Tilastokeskus
Statistikcentralen
Statistics Finland

Tutkimuksia
Forskningsrapporter
Research Reports
257

Paradigms in Statistical Inference for Finite Populations

Up to the 1950s

Vesa Kuusela

Tilastokeskus
Statistikcentralen
Statistics Finland

Tutkimuksia
Forskningsrapporter
Research Reports
257

Paradigms in Statistical Inference for Finite Populations

Up to the 1950s

Vesa Kuusela

Päätoimittaja – Chefredaktör – Principal editor:
Timo Alanko

Tutkimuksia -sarjan julkaisut eivät edusta Tilastokeskuksen kannanottoja.
Julkaisujen sisällöistä vastaavat kirjoittajat.

Taitto – Ombrytning – Layout:
Lehtonen Hilkka

ISSN 0355–2071
= Research Reports
ISBN 978–952–244–316–8 (pdf)
ISBN 978–952–244–315–1 (print)

Edita Prima Oy

Helsinki – Helsingfors 2011

Abstract

Sample surveys are an inherent part of a democratic society. Every day, important decisions are made basing on information that was obtained from a survey. Modern sample surveys started to spread after statistician at the U.S. Bureau of the Census had developed a sampling design for the Current Population Survey. In the beginning of 1950s, the theory was documented in textbooks on survey sampling. This thesis is about the development of the statistical inference for sample surveys.

For the first time the idea of statistical inference was enunciated by a French scientist, Pierre Simon Laplace. In 1781, he published a plan for a partial investigation in which he determined the sample size needed to reach the desired accuracy in estimation. The plan was based on Laplace's Principle of Inverse Probability and on his derivation of the Central Limit Theorem. They were published in a memoir in 1774 which is one of the revolutionary papers in the history of statistical inference, and its origin. Laplace's inference model was based on Bernoulli trials and binominal probabilities. He assumed that populations were changing constantly. It was depicted by assuming *a priori* distributions for parameters. Laplace's inference model dominated statistical thinking for a century.

Selection of the sample in Laplace's investigations was purposive. In 1894 in the International Statistical Institute meeting, Norwegian Anders Kiaer presented the idea of the Representative Method to draw samples. Its idea was that the sample would be a miniature of the population.

Arhtur Bowley realized the potentials of the sampling method and in the first quarter of the 20th century he carried out several surveys in the UK. As a professor of statistics, he developed the theory of statistical inference for finite populations. Bowley's theory leaned on Edgeworth's modification of the Laplace's model. Bowley's theory included also formulas for balanced stratification.

R.A. Fisher contributions in the 1920's constitute a watershed in the statistical science He revolutionized the theory of statistics and initiated estimation and inference theory. In addition, he introduced a new statistical inference model which is still the prevailing paradigm. The central idea is based on repeatedly drawing samples from the same population, and on the assumption that population parameters are constants. Fisher's theory did not include *a priori* probabilities.

Jerzy Neyman adopted Fisher's inference model and applied it to finite populations with the difference that Neyman's inference model does not include any assumptions of the distributions of the study variables. Applying Fisher's fiducial argument he developed the theory for confidence intervals. In addition, Neyman created optimal allocation for stratification.

Neyman's last contribution to survey sampling presented a theory for double sampling. This became the central idea for statisticians at the U.S. Census Bureau when they developed the complex survey design for the Current Population Survey. Important criterion was also to have a method that provided approximately equal interviewer workloads, aside of an acceptable accuracy in estimation.

Tiivistelmä

Otostutkimukset ovat demokraattisen yhteiskunnan luontainen osa. Joka päivä tehdään tärkeitä päätöksiä, jotka perustuvat otostutkimuksista saatuun informaatioon. Nykyaikaiset otantatutkimukset alkoivat levitä sen jälkeen, kun Yhdysvaltojen tilastoviraston (Bureau of the Census) tilastotieteilijät olivat 1940-luvun puolivälissä kehittäneet otanta-asetelman Current Population Survey -tutkimuksen tarpeisiin. Menetelmä dokumentoitiin 1950-luvun alussa otantateoriaa käsittelevissä oppikirjoissa. Tämä tutkimus käsittelee sitä, miten otantatutkimusten tilastollinen päättely kehittyi.

Ranskalainen tiedemies Pierre Simon Laplace muotoili ensimmäisenä tilastollisen päättelyn idean. Vuonna 1781 hän julkaisi osittaistutkimuksen suunnitelman, jossa hän määritteli, kuinka suuri otos tarvittaisiin vaaditun estimointitarkkuuden saavuttamiseksi. Suunnitelma perustui Laplacen käänteisen todennäköisyyden periaatteeseen sekä hänen johtamaansa keskeiseen raja-arvolauseeseen. Nämä oli julkaistu vuonna 1774 muistiossa, joka on yksi tilastollisen päättelyn vallankumouksellisista esityksistä ja sen lähtökohta. Laplacen päättelymalli perustui Bernoullin kokeisiin ja binomi-todennäköisyyksiin. Hän oletti, että perusjoukko muuttui koko ajan. Tämän hän otti huomioon olettamalla, että perusjoukon parametreilla oli *a priori* todennäköisyysjakauma. Laplacen päättelymalli hallitsi tilastollista ajattelua yli sadan vuoden ajan.

Laplacen tutkimuksessa otos poimittiin harkinnanvaraisesti. Vuonna 1985 norjalainen Anders Kiaer esitteli Kansainvälisen tilastoinstituutin kokouksessa niin sanotun edustavan menetelmän otosten poimimiseen. Pyrkimyksenä oli, että otoksesta tuli perusjoukon pienoismalli.

Arthur Bowley oivalsi otantamenetelmän mahdollisuudet ja 1900-luvun ensimmäisen neljänneksen aikana hän teki useita otostutkimuksia Englannissa. Tilastotieteen professorina hän kehitti kiinteän perusjoukon tilastollisen päättelyn teorian. Hän teoriansa perustui Edgeworthin modifikaatioon Laplacen mallista. Bowelyn teoriaan sisältyivät myös kaavat tasapainoiselle osituksellE.

R.A Fisherin kirjoitukset 1920-luvulla olivat tilastotieteen vedenjakaja. Hän mullisti tilastotieteen teorian ja pani alulle estimointi- ja päättelyteorian. Lisäksi hän esitteli uuden tilastollisen päättelyn mallin, joka on vallitseva edelleen. Keskeinen ajatus perustuu siihen, että samasta perusjoukosta poimitaan otoksia toistuvasti ja että perusjoukon parametrit ovat vakioita. Fisherin teoria ei sisältänyt *a priori* -todennäköisyyksiä.

Jerzy Neyman omaksui Fisherin päättelymallin ja sovelsi sitä kiinteän perusjoukon päättelyssä sillä erotuksella, että Neymanin malliin ei sisälly oletuksia tutkimusmuuttujien jakaumista. Soveltamalla Fisherin niin kutsuttua fidusiaalista väitettä Neyman kehitti luottamusvälien teorian. Lisäksi Neyman kehitti optimaalisen allokaation ositteiden laatimiseksi.

Neymanin viimeisessä kirjoituksessa otantateoriasta aiheena oli kaksivaiheinen otanta. Tämä tuotti keskeisen oivalluksen Yhdysvaltojen tilastoviraston tutkijoille heidän kehitellessään monimutkaista otanta-asetelmaansa Current Population Survey -tutkimukselle. Tärkeä kriteeri oli tuottaa menetelmä, joka työllisti haastattelijoita tasaisesti, sen lisäksi, että estimoinnin tarkkuus oli hyväksyttävällä tasolla.

Preface

The main topic of this thesis is the origin and the historical development of statistical inference for finite populations. It is inherently linked with the history of survey sampling. The stimulus for this study emerged many years ago, while I was preparing a study of Anders Kiaer's influence on the birth of survey sampling. That inspired me to gain wider knowledge on the history of survey sampling and statistical inference for finite populations.

During the preceding sixty years, some ten well-known articles about the history of survey research have been published. Several books have been written about the history of statistics and probability and even more articles can be found in journals. In addition, few extensive textbooks have been written about the general history of statistics. Currently, many original texts – even very old texts – can be found on the Internet. The topic of this thesis seems to be well-covered and the unavoidable question is whether it is possible to discover anything new.

A slightly astonishing observation has been that either the published texts deal with the history up to the beginning of the 20th century, touching only superficially on the later development, or they begin from the first quarter of the 20th century and nearly ignore the earlier history. Another slightly astonishing observation was that very few of these texts dealt with statistical inference for finite populations. In the process of searching for facts about the growth of statistical thinking and the development leading to survey sampling, it slowly became apparent that the written history did not adequately cover the subject, included obvious misinterpretations, and was biased in some parts.

Survey methodology involves theoretical problems of survey sampling and epistemological problems of inference, but it also involves significant practical problems of survey undertaking. The practical questions have had an important role in the development of sampling methods and hence in inference within a finite population framework. The theoretical development cannot be analyzed apart from the practices of survey undertaking, but that has been noted only in very few of the texts

This thesis follows the development of ideas in a chronological order; this seemed a natural approach. There are two parallel streams: development of sampling techniques and development in statistical inference. Two streams have been followed because developments have advanced at a different pace and also because the development has been an interplay between survey practice and sampling theory. However, prominence is more on the inferential aspects. Development in sampling techniques is described mainly to make it easier to understand the development of statistical inference.

It was a slightly unexpected finding that a mathematical theory of statistical inference existed already at the end of the 18th century. A common conception implicitly given in contemporary statistical literature is that statistical inference started from R. A. Fisher's ideas in the second quarter of the 20th century. This finding led to an investigation of whether the history of statistical inference shows paradigms and a paradigm shift in the sense Thomas Kuhn (1962) described them.

To reach the aim of this study, the original texts are analyzed whenever possible. The written histories have been mainly used as a guideline. The focus has been on the contributions of those persons who have been most influential in the development of statistical inference methods. This delimitation has left many significant mathematicians and scientists without due attention.

Obviously, there are several angles from which to approach the history of survey sampling. One is to examine sampling in the context of the history of ideas: who formulated them, and how and why they were formulated, promoted, defended, and discarded or supplanted. Another perspective is to look at sampling theory as a branch of mathematics and then to fit this development into the general pattern of how mathematics – especially probability theory – evolves. A third approach is through the technical and practical developments, which enable applications of different methods. The third approach is relevant because much of the development of survey sampling has been motivated by practical problems of sample selection, data capture, data processing and data analysis, and not as much by abstract ideas.

The approach in this thesis has been to look at the development from the perspective of survey practice. Several detailed and extensive accounts about the history of survey sampling have been published, but the assessment of the development is often done from a theoretical point of view, often focusing mainly on the emergence of randomization. The practical approach means that development is analyzed more in respect to the implications that practical data collection and data processing tasks bring about. The development of methods in survey research can also be seen as an interplay between what is possible in practice and what is mathematically tractable.

Limitations

Survey research has been used for a variety of purposes, but this thesis is focused only on the enumerative use of sample surveys. The analytic applications are skipped almost entirely. Superpopulation approaches in the modern sense are often connected to analytical problems, and therefore they are touched on only on a few occasions. However, the concept of a superpopulation in a different sense than the modern one has been an implicit element of statistical inference before the current paradigm.

The scope of the current thesis is the early history of survey sampling up to 1950s. The classical theory of survey sampling was more or less completed in 1952 when Horvitz and Thompson (1952) published a paper on a general theory for constructing unbiased estimates. Most of the classical books about statistical sampling theory were also published roughly at the same time (Cochran 1953; Deming 1950, Hansen, Hurwitz and Madow 1953). In a manner of speaking, Horvitz and Thompson completed the classical theory of sampling techniques, and the random sampling approach was almost unanimously accepted.

There has been a lot of development in classical theory since then, but the paper by Horvitz and Thompson established the foundations for later development. The most notable development has taken place in the derivation of the model-assisted estimators. That is, estimators that utilise auxiliary information from the population via modelling. After the mid-1950s, a discussion started on the basics of statistical inference, and challenges to the random sampling approach appeared, but all of this will be excluded from this thesis.

A danger in analyzing the history lies in the fact that the practical problems of one hundred or more years ago were quite different from those of the current surveys. Undeveloped infrastructure, as compared to modern societies, had many implications on the statistical research. There is the risk of projecting the current world and ideas to the historical development, and that may lead to wrong conclusions.

There is also a risk to project current thoughts and ideas to the historical development. If there have been different paradigms, they have also been based on different world views. To analyze previous paradigms in terms of current knowledge, ideas, and ideals would probably lead to faulty conclusions.

The thesis focuses only on the major contributions to the subject as seen from the viewpoint of statistical science. Therefore, many slightly less important developments have been left out. However, it is the authors wish that this gives a sufficiently accurate description of the history of the prevailing statistical inference for survey sampling.

Acknowledgements

Professor Risto Lehtonen of the University of Helsinki, as my mentor, has supported me all the way in writing this thesis. Professor emerita Ene-Margit Tiit and Doctor Emmanuel Didier pre-examined the manuscript. Their pertinent comments helped essentially to amend its content and lessen errors. I am much obliged to both of them. In the course of the preparation of the text, I had many long discussions with Professor Carl-Erik Särndal on various topics in the context of statistical inference. His profound knowledge of the historical aspects and the fact that he personally knew some of the key persons kept me on the right track while searching for the culmination points. I am grateful for the time he spent with me. Professor Leif Norberg pointed out several inconsistencies and errors in the text and made several constructive recommendations to improve it. I thank him for reading the text with such strictness. Professor Stephen M. Stigler read an early version of the manuscript and gave some important advice and encouraging comments about it. Also Professor Jelke Bethlehem read the manuscript and commented on it. His comments helped appreciably to touch up the text. At an early stage, also Professor Alain Desrosiéres read an unfinished manuscript for this thesis. His comments directed me to look into sources and to follow happenings which proved to be vital for this study.

Director General of Statistics Finland twice granted me a leave of two months from my daily duties to prepare this thesis. These study leaves were indispensable to be able to dig into the vast material on the topic and to analyse it. In addition, the library of Statistics Finland turned out to be an ample source of sta-

tistical literature, not least because its origin goes back to the time when Finland was part of the Russian Empire, and the staff of the library proved to be very helpful. Ms. Mia Kokkila organised checking of the language and partly checked it herself and Ms. Hilkka Lehtonen took care of the painful task of editing the final version. I am grateful to both of them for their efforts.

Preparation of this thesis took several years. It was a lonesome process. To a great extent, the work was done aside of my daily duties and that time was away from many other activities of life. My wife, Joanna, probably mostly had to suffer from this. I thank her for her endurance.

In Espoo on the summer solstice of 2011

Vesa Kuusela

Contents

1 Introduction

Statistical inference here means methods that enable drawing probabilistic conclusions about a set of units, usually called a population, after observing only a part of it. These methods constitute a branch in the statistical science called sampling theory. They are inherently mathematical and are based on probability calculus. In addition, statistical inference involves significant philosophical, or epistemological, questions.

Statistical inference can also be defined as a formalized theory of inductive inference. That is, a set of methods that enable rational generalisations from observations to a wider domain than the one that has been observed. In this, the word "rational" denotes a probabilistic expression for inductive generalisations. In formal theory, a population is a central concept meaning the domain whose characteristics are inferred. Population can have different definitions as to intents and purposes. The subject of this thesis is a finite population. A finite population consists of distinct units that could be listed at least in theory.

An essential distinction between statistical inference for finite populations and for infinite, or hypothetical, populations is that a sample investigation on finite population could be, at least in theory, replaced by a complete enumeration or census. In research concerning an infinite population, a complete enumeration is not possible because the population cannot be defined in such a manner that it would be possible know all the units which constitute the population.

Statistical inference for a finite population is an inherent part of a more general method called survey research. Their relationship can be expressed by saying that statistical inference is a (mathematical) formulation for drawing conclusions in survey research. In general, the purpose of a survey is to describe the state (of the nature) of a population by estimating the values of some parameters, characterizing the properties of the population, from a sample. In a finite population, a parameter is a constant, but in separate samples, its estimates may obtain different values. The distribution of estimates obtained from all possible samples is called the sampling distribution. Inference methods also provide measures of the accuracy of the estimates obtained by sampling. In complete enumeration, there would be no need for statistical inference because there is no sampling error.

In sampling techniques, there are two central problems: (1) how to draw a sample from a population so that the sample can be expected to represent the population; and (2) how to calculate estimates from the sample. The latter problem is intrinsically related to the first.

The central concept in statistical inference for finite populations is the so-called **confidence interval**. It is a random interval having a stated probability of containing the unknown value of the population parameter that has been estimated. Särndal, Svensson and Wretman (1992) define a confidence interval related to a random sample, *s*, as a random interval $CI(s) = [t_L(s), t_U(s)]$, where $t_L(s)$ and $t_U(s)$ are the lower and upper endpoints. The endpoints are two statistics, $t_L(s) < t_U(s)$, which can be calculated for every sample obtained by a specified sampling design $p(s)$. The random element in the interval estimation is the randomly selected sample, s. A confidence level, $1-\alpha$, for a parameter, say population total t, is given as the probability

$$P[t \in CI(s)] = 1 - \alpha \tag{1.1}$$

where α is the probability that the selected sample s does not include t. The confidence level is interpreted to tell that $100*(1-\alpha)$ percent of the confidence intervals of all possible samples contain the parameter of interest. Jerzy Neyman introduced this type of confidence interval at the beginning of the 1930s (Neyman 1934).

If $\hat{t}$ is the point estimator for the unknown population total t, a confidence interval for t at level $1-\alpha$ is usually computed as

$$\hat{t} \pm z_{1-\alpha/2} s_{\hat{t}} \tag{1.2}$$

where $z_{1-\alpha/2}$ is the constant exceeded with probability $\alpha / 2$ by the $N(0,1)$[1] random variable, and $s_{\hat{t}}$ is the standard error of the estimate. Frequently in practical survey work, $\alpha = 0.05$ has been chosen and accordingly $z_{1-\alpha/2}= 1.96$. However, also $\alpha = 0.1$ and $\alpha = 0.01$ can be found in survey reports.

The 95% confidence interval for the population total t will be

$$\left[\hat{t} - 1.96 s_{\hat{t}}, \hat{t} + 1.96 s_{\hat{t}}\right]$$

This interval will contain the unknown total t for an approximate proportion of $1-\alpha$ of repeated samples s drawn with the same design, if the sampling distribution of $\hat{t}$ is approximately a normal distribution with mean t and standard deviation $s_{\hat{t}}$. Standard deviation $s_{\hat{t}}$ is usually called standard error and it varies between sampling designs. This condition is essentially equivalent to saying that the **Central Limit Theorem** applies for the random variable $\hat{t}$ (Särndal, et. al. 1992).

Current sampling techniques include a large number of different sampling designs, and the calculation of estimates and confidence intervals differs considerably due to the applied design. The gist of this study is the historical development of the prevailing inference methods in sample surveys.

1.1 Examples of sample surveys

This thesis deals primarily with sample surveys, which are undertaken by government agencies. The first sample survey in which the prevailing standards were applied was undertaken in the U.S. in the early 1940s (see Hansen and Madow 1976). It was called the Current Population Survey (CPS) and a similar survey,

[1] $N(0,1)$ designates the distribution function of the standardised normal distribution.

based on the same principles as the CPS, was soon started in several other countries under the name Labour Force Survey (LFS). Currently, nearly all National Statistical Institutes in the world conduct the Labour Force Survey.

Currently, the sample for the CPS is a multi-stage stratified sample of approximately 56,000 housing units from 792 sample areas, covering the entire U.S. It is composed of housing units drawn from lists of addresses obtained from the previous census. In the first stage of sampling, the country is divided into primary sampling units (PSUs). The PSUs are then grouped into strata that are sociologically and economically as homogeneous as possible. One PSU is sampled per stratum with the selection probability proportional to the size of the population in the stratum.

In the second stage of sampling, a sample of housing units within the sample PSUs is drawn. Ultimate Sampling Units (USUs) are clusters of housing units. The bulk of the USUs sampled in the second stage consist of sets of addresses, which are systematically drawn from sorted lists of addresses of housing units. Housing units from blocks with similar demographic composition and geographic proximity are grouped together. If addresses are not recognizable on the ground, USUs are identified using area-sampling techniques. Occasionally, a third stage of sampling is necessary when the actual USU size is extremely large.

Each month, interviewers collect data from the sample of housing units. Members of housing unit are interviewed for four consecutive months, then dropped out of the sample for the next 8 months, and then brought back for the following 4 months. In all, a selected housing unit is interviewed eight times.

During the interview week, field interviewers and telephone interviewers attempt to contact and interview a responsible person living in each sample unit. A personal visit interview is required for all households that are in the sample for the first time. This is because the sample is a sample of addresses and it is not possible to know in advance who the occupants of the household are or whether the household is occupied or eligible for an interview. The major results of the survey are released no later than two weeks after the completion of the interviews.

Based on the CPS, it was estimated in 2006 that 7,668,000 families lived in poverty in the U.S., and the upper and lower endpoints of the 90% confidence interval were 7,484,000 and 7,852,000, respectively.

Data collection is the major source of survey costs, and its organisation determines the time required for data collection. Continuous surveys, like the CPS, require a permanent interviewer corps to visit households or to call them from a telephone interview centre. The U.S. Bureau of the Census has more than 2,000 field interviewers solely for the CPS and nearly 300 interviewers in three telephone interview centres.

In addition, the data processing following the data collection is a very labour- and time-consuming phase of the survey process. In the early 1950s, when the first computer became available for the CPS, its computational power still had to be taken into account when designing sampling (see Bellhouse 2000 and Cochran 1942). Before the computer era, data processing and tabulations were done using punched card calculators, or Hollerith machines. In the 19th century, everything was done manually. Because of the data processing, a population census at that time was an enormous undertaking (see Bellhouse 2000 and Grier[2]).

The first partial investigation involving the characteristics of a modern survey was undertaken in France in 1802. Pierre Simon Laplace[3] carried out the partial investigation to estimate the size of the population in the country. His design was based on the fact that during the last quarter of the 18th century in France, all births were registered in parishes. Laplace took a sample of the departments, counted the total population in them on one day, and then, using a ratio estimator, estimated the population in the whole country (with the help of the information on registered births in the whole country). He concluded: "supposing that the number of annual births in France is one million, which is nearly correct, we find ... the population of France to be 28 352 845 persons." Before the survey was carried out, Laplace calculated what sample size was needed to attain the required accuracy in estimation. Finally, Laplace calculated that the "standard error", given the data, was 107,550 persons, and he concluded that it makes "the odds about 300 000 to 1 against an error of more than half a million". Laplace's survey and his method are described in Chapter 4.

Figure 1.1:
Processing of census data at the national statistical institute of France in the beginning of the 20th century

1.2 Aims of the thesis

The main topic of this thesis is the origin and the historical development of statistical inference for finite populations. This is inherently linked with the history of survey sampling. Statistical sampling theory was manifested in the beginning

[2] David Grier's article, "The Origins of Statistical Computing", is published on the website of ASA and has no other reference information (see http://www.amstat.org/about/statisticians/index.cfm?fuseaction=papers)

[3] **Pierre Simon Laplace** (1749–1827) was a French astronomer and mathematician. He was born in Normandy, reportedly in a modest family. At a young age, he sent a letter of introduction to the famous French mathematician and philosopher Jean Rond D'Alembert. The paper on the principles of mechanics excited D'Alembert's interest, and on his recommendation, a place in the École Militaire in Paris was offered to Laplace at the age of 19. He was later appointed as the professor of mathematics there. Despite being an ingenious and voluminous writer, Laplace was also a politician. In 1799, Laplace became the Minister of the Interior, but only for six weeks – Napoleon thought he was incompetent. Nevertheless, he became a member of the Senate.

of 1950s in two famous textbooks (Cochran 1953, Hansen, Hurwitz and Madow 1953). This period was preceded by a much longer and diversified period of a search for methods that could be generally accepted. One aim is to identify the most important turning points and developments that led to current theory.

Both statistical inference for finite populations and survey sampling can be seen as parts of a more general discipline called survey research methodology. Survey research methodology also involves the practical problems of survey undertaking, in addition to theoretical problems. The practical questions have had an important role in the development of sampling techniques, and therefore in statistical inference within a finite population framework, that it cannot be analysed detached from the practices of survey undertaking.

As noted earlier, survey sampling comprises two distinct but inherently linked parts: obtaining a representative sample from a population, and methods to draw conclusions from the sample about the population. The first part is practical, which also involves two distinct but equally important parts: (1) drawing a sample from a frame representing the population, and (2) data collection from sampled units. The practical problems of data collection are the most significant single factor in developing different sampling designs (see also Hansen and Hurwitz 1943). The latter part is called statistical inference.

Another aim is to find out whether survey sampling and statistical inference have involved paradigms and paradigm shifts in the sense Thomas Kuhn defined them (see Chapter 1.6).

1.3 Role of population in statistical inference

Population is a central concept in statistical inference because it spans the framework of the inference. In this respect, real populations and hypothetical populations have conceptually essential differences.

Real populations are composed of distinct real units. Real populations can still be divided into two categories: finite populations and infinite populations. The basic difference between these two is that a finite population is composed of a limited number of members. The number of units in a finite population is known, or they could be counted, and they could be labelled. The members of an infinite population cannot be counted or labelled. An infinite population is an ambiguous concept and in most cases, a hypothetical population better describes its nature.

Occasionally, finite population and fixed population have slightly different definitions. A finite population consists of a finite number of units, but their exact number is not known, and therefore they cannot be uniquely identified. For example, the fishes in a pond or the whales in a sea compose a finite population in this sense. A fixed population is composed of a known number of distinct units that have been, or could be, uniquely labelled. For probability sampling, it is required that there exists an operational representation of the units, for example a list of unit labels, called a **frame** or a **sampling frame**. Every unit of a fixed population is accessible with the help of the information in the frame.

In a hypothetical population, there are an infinite number of units. A hypothetical population is not defined through its members but by a rule or definition that confines them. R.A. Fisher, who introduced the concept of population, defined a hypothetical population to be "the conceptual resultant of the conditions we are studying" (see e.g. Fisher 1922). The definition means that population is defined through features that every existing or potential member of the population possesses. In a hypothetical population, it is not possible to know, or to list, its members.

Consequently, a full enumeration of a hypothetical population is not possible, and neither is it possible to estimate the total sum of any characteristic – actually, it does not even exist. Inference within a hypothetical population framework usually aims at revealing an abstract cause mechanism. Inference in a fixed population framework aims at estimating population parameters.

Formally, the distinction between inferences for fixed and hypothetical populations is in the assumed stochastic structure: in a fixed population, both the measured values of sample units' characteristics and the population parameters are constants. No probability distribution is attached to observations. The stochastic element in inference is induced by random selection of a sample. In sampling from a hypothetical population, observations are assumed to have a known probability distribution *f(x)*, and the probability to obtain a given sample $x_1, x_2, \ldots x_n$ of size *n* is given by the product $\prod_1^n f(x_i)$. A hypothetical population is a theoretical quantity that is helpful in designing the mathematical setup of statistical inference. A finite population is a real entity whose parameters are to be estimated.

Superpopulations

In analytic sample surveys, interest usually is focused on parameters of a "superpopulation". They are associated with a stochastic mechanism that is assumed to have generated the observed values. R. A. Fisher coined the concept of superpopulation in the 1920s, but in current statistical texts, superpopulation has a slightly different connotation than what Fisher meant. The superpopulation approach is often thought to constitute a bridge between analytic and enumerative surveys.

Deming (1953) considered a "superpopulation" to be a hypothetical infinite population from which the finite population is itself a sample. An investigator samples the finite population and draws inferences from the sampled values. Unlike in classical sampling theory, where the targets of inference are parameters of a finite population, a stochastic model for the finite-population values is used to evaluate and suggest sample designs and estimators. However, for addressing scientific questions (as opposed to, e.g., administrative questions), the parameters associated with the stochastic model are typically of more interest than the finite-population parameters.

Deming (ibid.) refers to inference for superpopulation parameters as an "analytic" use of survey data. A simple example of superpopulation inference is when comparing two domain means, where it is of interest to ask whether the superpopulation means are equal, but seldom of interest to ask whether the finite population means are equal. (Actually, that question is futile, since the means in two real populations would be equal only very rarely.)

In modern superpopulation inference, it is assumed that a process or a model has generated the observable population U of size N. The model M is thought to describe the relationship between the observable variables y and $x_1, x_2, .., x_p$. The model states, for example, that for each unit of the observable population holds:

$$\begin{Bmatrix} y_k = \beta_1 x_{1k} + \beta_2 x_{2k} + \ldots + \beta_p x_{pk} + \varepsilon_k; k = 1, \ldots, N \\ E_M(\varepsilon_k) = 0; V_M(\varepsilon_k) = \sigma^2 \end{Bmatrix} \tag{1.3}$$

In addition, it is assumed that the random errors ε_k are independent and normally distributed. In this mode of inference, the interest is not in the finite population U at the present time, but rather in the process or the causal system relating y and $x_1, x_2, .., x_p$.

1.4 Epistemological features of statistical inference

1.4.1 Inductive inference

A central question in scientific inference is how is it possible to draw conclusions of something that we are not capable of observing directly or completely. In order to obtain knowledge and understanding about the surrounding world it is necessary to have both methods to acquire data and methods to reveal particulars and relations between the observed facts to establish generalisations and theories. A central part of scientific activity, or the pursuit of knowledge in general, is the logic by which investigators end up with conclusions from observations, experiments, and initial premises.

The two main methods of scientific inference are called deduction and induction. In some respect, they can be regarded as opposites: deduction goes from general to specific, and induction goes from specific to general. Induction is an argument or theory starting with empirical observations and leading to a conclusion, while deduction goes in the opposite direction, from theory to observation.

Deduction is an old method to draw conclusions from given premises, postulated already by Aristotle. The power of deductive inference comes from the fact that from true premises, correctly deduced conclusions are necessarily true. A classic example of the competence of deduction is Euclidian geometry, where the whole system is deduced from a few axioms. The growth of mathematical theories in general, including the probability theory, is an example of the capability of deductive reasoning.

In scientific experimentation, the so-called hypothetico-deductive method is frequently applied. Schematically, the method works as follows: From a general hypothesis and particular statements of initial conditions, a particular predictive statement is deduced. The statements of initial conditions, at least for the time, are accepted as true; the hypothesis is the statement whose truth is at issue.

By observation or experiment, we determine whether the predictive statement turned out to be true. If the predictive consequence is false, the hypothesis is disconfirmed. If the observation reveals that the predictive statement is true, we say that the hypothesis is confirmed. The design of a scientific experiment aims at creating such an experimental setup that the deductive procedure could be applied to draw conclusions.

In empirical research, deductive inference is not sufficient. Francis Bacon[4] recognized that the scientific method embodies a logic essentially different from that of Aristotle. Bacon commended the method of careful observation and experimentation. He put forward that scientific knowledge must somehow be built on inductive generalisation from experience.

A simple example of inductive inference is the following: if we draw balls from an urn and we only have white balls, we tend to infer that all balls in the urn are white. Every new observation of a white ball strengthens our conviction on the rule (that all balls in the urn are white), but we can never be absolutely sure. On the other hand, a single observation of a black ball ruins the rule. Induction is said to be ampliative and undemonstrative. That is, it expands the observations to a wider domain than what was originally observed, but inductive inference cannot demonstrate that a rule is true.

More than a century after Bacon's works, David Hume[5] published a book in which he criticised the principle of inductive inference. His critique began with a simple question: How do we acquire knowledge about the unobserved? (Hume 1739 and 1748) Hume's basic problem can be described as follows: Given that all the balls that were drawn from an urn have been white so far, and given that the conclusion has been entertained that the unobserved balls are also white, do the observed facts make up sound evidence for that conclusion? Basically, the problem of induction is a problem of explaining the concept of evidence.

Hume's answer was sceptical. It is out of the scope of this study to deal comprehensively with this question, but several authors, for example, Salmon (1967), have analysed it thoroughly. Hacking (1975) treated Hume's philosophy in the context of probability theory. In addition, Chatterjee (2003) has analysed profoundly Hume's philosophy in relation to statistical inference.

Hume's critique essentially rested on his attack on **the principle of the uniformity of nature**. It is obvious that inductive inferences cannot be expected to yield correct predictions if nature is not uniform. For example, if we do not know whether the future will be like the past, it is not possible know which facts will hold. Likewise, if it is not believed that a population under study is uniform or stable in all of its parts, it is not feasible to generalize the results obtained from a sample.

[4] **Francis Bacon** (1561–1626) was an English politician and philosopher. He put forth the view that only through reason are people able to understand and have control over the laws of nature. His famous adage, 'Knowledge is power', reflects this conception. Francis Bacon's influence on empirical research has been so strong that he has been called "the Father of Modern Science".

[5] **David Hume** (1711–1776) was a **Scottish** philosopher and historian who has been regarded as the founder of the sceptical, or agnostic, school of philosophy. He had a profound influence on European intellectual life.

Hume's problem has been approached from many points of view. An example is the so-called induction by enumeration. Suppose that a coin has been thrown a large number of times. Given that *m/n* of observed throws has been heads, we infer that the "long run" relative frequency of heads is *m/n*. It is obvious that induction by enumeration is closely related to the long-run frequency interpretation of probability.

Another, slightly different, example was given by Laplace at the end of the 18th century. He posed the question: how certain can we be that the sun will rise tomorrow, given that we know that it has risen every day for the past 5,000 years (1,825,000 days). One can be pretty sure that it will rise, but we cannot be absolutely sure. In response to this question, Laplace proposed **the Law of Succession**. In its simplest form, it means the following: If we have had x successes in n trials and ask what is the probability of success in the next trial, we add one to the numerator and two to the denominator ($(x + 1)/(n + 2)$) (see Chapter 4, Formula 4.7). Applying this procedure, one could be 99.999945% sure that the sun will rise tomorrow.

Induction by enumeration and hypothetico-deductive method are inherently different approaches. Induction by enumeration actually consists in simple inductive generalisations from instances, and the hypothetico-deductive method is in contrast to it. The hypothetico-deductive method aims at confirming or disconfirming hypotheses derived from previous knowledge, while induction by enumeration aims at deriving scientific hypotheses.

An answer to Hume's critique is that inductive conclusions are probabilistic, not absolutely certain. An inductive inference with true premises only establishes its conclusions as probable. At the time when Hume published his critique, mathematicians dealt only with the problems of direct probability. The critique gradually initiated development of the methods for the calculation of inverse probability to address the problems of induction.

Inverse probability and statistical inference can be seen as a formal approach to apply induction in empirical research. Inverse probability in statistical science involves two problems: the problems of direct probability are mathematical and hence involve deductive inference; in inverse probability, known probability distributions are applied to make inferences about the unobserved part of nature and it is inherently inductive. Statistical inference in the modern sense can be seen as an outgrowth of inverse probability.

The American mathematician, C.S. Peirce, defined induction to be "reasoning from sample taken at random to the whole lot sampled" (see Stigler 1978, p. 247).

The famous Theorem of Thomas Bayes (Bayes 1763) is often regarded as the first method to calculate inverse probability (see Chapter 3). However, Laplace gave the first precise formulation of inverse probability in a careful scientific context in a mémoire in 1774. Laplace's contributions on inverse probability are analysed in Chapter 4.

1.4.2 The inference model

It is obvious that probability and probability models play a central part in inductive inference. A probability model can be seen as an abstract description of mass events in the real world by which one is able to predict the frequency of future events and to analyze observations from such events, but probability models cannot be applied directly in inductive inference.

In direct probability, it is generally conceded that knowing the value of a stochastic probability factor, say *s*, the probability for an arbitrary or 'random' occurrence of a chance event can be determined, like in coin-flipping, dice-rolling, or selecting balls from an urn.

In inverse probability, the question is reversed: Given the outcome of an experiment or observations, what can be concluded about the underlying causes of outcomes and their stochastic characteristics? Obtaining an answer to the question requires the use of direct probability in one way or another.

Thought experiment

Abstract probability models cannot be applied directly in real world phenomena because the situations to be analysed are much too diverse and usually too complex. The inference model requires an intermediate model, a thought model, which links an abstract probability model to the real-world phenomenon. Characteristic of a thought experiment is that it involves such a setup that can be (or could be) tested experimentally if necessary.

One of the oldest thought experiments is the so-called **urn problem** or **urn trial**. The urn problems have been a part of probability theory since at least the publication of the *Ars conjectandi* by Jakob Bernoulli in the beginning of the 18th century (see later). Bernoulli considered the problem of determining from a number of pebbles drawn from an urn the proportions of different colours. The urn trial is often called a Bernoulli trial or Bernoulli experiment.

In an urn trial, an urn is thought to contain *n* balls (or tickets), *x* white and *y* black. One ball is drawn randomly from the urn and its colour is observed. It is then placed back in the urn[6], and the selection process is repeated. Occasionally, the two possible outcomes are called "success" and "failure". For example, a white ball may be called "success" and a black ball "failure". The urn trial induces a Binomial Distribution.

Another example of an inference model is the one which Thomas Bayes' applied in formulating his theorem: the model was based on the positions of balls on a (billiard) table after they were rolled on it (see Chapter 3).

R.A. Fisher introduced a new inference model in the 1920s. Its central idea is to repeatedly draw samples from the same known probability distribution (see Chapter 9). Fisher's thought model is still the predominating one in statistical inference. In the 1930s, Jerzy Neyman adapted it in a modified form to finite

[6] In the setup with the replacement of balls, the subsequent drawings are independent. In another setup, the urn is assumed to contain an infinite number of balls and then the drawings can also be regarded as independent.

population inference (see Chapter 10). Neyman's idea of drawing samples repeatedly from the finite population is the core of modern sampling theory.

Thinking models are also applied in a wider scope. A common thinking model up to the 20th century originated from the planetary system, which was also incorporated into social research. A parameter describing a state of population was paralleled with a planet and its position. Measurements gave varying results so that observations had a distribution around the true value. In addition, the planet was moving all the time, and therefore its position could not be considered to be constant. The resulting uncertainty was described by *a priori* probability. In social research, this idea led to thinking that a society should be approached as a constantly changing universe, a superpopulation, and every observable population was a realization of some phase of the superpopulation. The world view behind these thought models was mechanistic, comprising of distinct units, and often a Greater Cause was assumed to act behind the events. This originated from Newton's philosophy, and it dominated thinking until the beginning of the 21st century.

1.5 Research on the history of survey methods

Currently, survey research is applied in a variety of different areas, such as scientific research, public administration, agricultural research, marketing and opinion research, etc. The first applications of sample surveys, in the modern sense, concerned human populations. The most significant impetus was to have a method to be used alongside a population census to explore population characteristics. An aspiration was to have a method that was faster to carry out and less costly than a total enumeration, as well as a method that was easier to apply for varying needs, and to focus on more specialized questions than what was possible in a census.

1.5.1 Research on the history of sampling techniques

During the past 60 years, several papers have been written about the history of survey sampling. For example, Stephan (1948), Yates (1946), Seng (1951), Chang (1976), Kruskal and Mosteller (1980), Sukhatme (1966), O'Muircheartaigh and Wong (1981), Hansen, Dalenius, and Tepping (1985), and Bellhouse (1988) have written comprehensive accounts on development in the 20th century. In most articles, survey research in the modern sense is considered as starting from Anders Kiaer's presentation at the ISI meeting in 1894. Kiaer's Representative Method did not involve sampling methods in the same sense they are presented in the classical textbooks, but in the written history, it is a common opinion that Kiaer's contributions were the starting point for the development of current sampling techniques.

However, Kendall (1960) argued that the first example of partial investigation, or survey, was the one that John Graunt carried out in 1662 to estimate the size of the population in London. Graunt's estimation was intuitive and did not involve any reference to probability. Nevertheless, Kendal (ibid.) regarded Graunt's investigation as the starting point of statistical science. The early history of statistics and Graunt's survey are described more closely in Chapter 2.

The second early example of a partial investigation is Pierre Simon Laplace's estimation of the size of the population in France in 1802. Laplace's method involved a sound theoretical setup that was based on probability. Laplace published the outline of the theory already in 1783, 20 years before the actual survey (Laplace 1783). Laplace's major contribution in the mémoire published in 1774 was his Principle (of Inverse Probability), which addresses the same question as statistical inference: to draw probabilistic conclusions about a population from a sample of observations (Laplace 1774).

Using the Principle, Laplace also calculated what sample size was needed to obtain the required accuracy of estimation. After the data collection, he calculated a probabilistic interval estimate of the size of the population. Laplace's interval estimate is close to the modern confidence interval, although it was based on a different probabilistic setup. Therefore, it is often called a credibility interval. Laplace's survey and the methods he applied are presented in Chapter 4.

Laplace's Principle is close to Thomas Bayes' method with equal prior probabilities. Bayes' Essay was published a few years earlier than Laplace's Principle, and there has been some discussion about whether Laplace was aware of Bayes' Essay. The current understanding is that he was not (Laplace was 14 years old when Bayes Essay was published in England). Bayes' Essay did not have greater influence on the development of probability theory and statistical science in the 19th century. Bayes' Essay and other contributions during the same era are presented in Chapter 3.

Several textbooks deal with the history of official statistics in the 19th century and beginning of the 20th century. Westergaard (1932), Porter (1986), Hacking (1990), and Desrosières (1998) give comprehensive accounts of the rise of statistical thinking. All three authors analyse and describe how statistics became a central part of administration in western countries and how the statistical professions started and assumed their current roles. Westergaard (ibid.) describes the beginning of statistics and the institutional changes that fostered the status of statistics. Hacking describes how the "avalanche of printed numbers" began, and how it eventually became possible to think of statistical patterns as natural parts of societies. Chapter 6 is devoted to the description of the emergence of statistical thinking and the consolidation of the ideas of Laplace (and Gauss) and the development that paved the way for the representative method.

Obviously, no surveys or partial investigations within human populations were carried out in the course of the 19th century because the prevailing conception was that such populations were so heterogeneous that only a full enumeration could be truly representative.

Indirectly, the Belgian scientist Adolphe Quetelet was central in justifying partial investigations: he was the central pacemaker in the tradition to carry out standardized censuses on regular basis, thus providing basic information about populations; he was also involved in starting statistical institutions in which new statistical methods could be presented and discussed; he was the first to establish the regularity of social phenomena; and lastly, he showed that the greatest part of social, biological, and economical phenomena followed the Normal Distribution (see Chapter 6).

The original rationale for this study was the significance of Anders Kiaer in bringing forth the Representative Method. He presented this method for the first time at the International Statistical Institute (ISI) meeting in Bern (Kiaer 1895). Kiaer's aim was to introduce a new data collection method for social studies that was less expensive to carry out than a total enumeration and more flexible. The literature of the history of survey sampling emphatically regards Kiaer's first presentation of the method in 1895 as the starting point for survey research and sampling techniques. Obviously, that is true in the sense that Kiaer raised the topic in the agenda of the ISI and defended the method in subsequent meetings. Because of Kiaer's persistence, the ISI had to take a stand on the method and eventually accept it as a method that national statistical offices could apply. However, partial investigations evidently were carried out already before Kiaer's survey. Kiaer's work and contributions are analysed in Chapter 7.

Arthur Bowley from London University College realized the usefulness of Kiaer's method in shedding light on the living conditions of the working class in England. During the first quarter of the 19th century, he carried out several living condition surveys in England. In addition, he derived a mathematical apparatus to calculate the accuracy of estimates in the form of a credibility interval, which was close to a confidence interval. It was published in the mèmoire to the ISI in 1926.

A practical problem was that random sampling could not be applied, because the only known method, simple random sampling, was not feasible due to practical constraints. In the 1930s, Jerzy Neyman wrote three papers in which he established the basic theory of statistical inference for finite populations. In the third paper, published in 1938, he directly addressed the practical problem of taking a survey of a large human population. Only that paper gave tools to design complex sample surveys with reasonable costs and sufficient accuracy. After that paper, a period of rapid developments in sampling methods took place in the United States. The most important contributions came from the U.S. Bureau of the Census and the practical impetus game from the need to design a sampling method for the newly established Current Population Survey (CPS). Hansen and Hurvitz, applying the principles Neyman had presented, developed a method by which the data collection of a large social survey could be undertaken with acceptable costs and manageable fieldwork. After that, the development was very rapid, and by the first half of the 1950s, the classical sampling theory was established. The final formulation of modern sampling techniques is described in Chapter 12.

1.5.2 History of statistical inference

The history of probability and the development of its theory is a well-covered topic. For example, the books written by Stephan Stigler (1986) and Anders Hald (1998 and 2007) give very detailed accounts up to the beginning of the 20th century. In addition, Todhunter's (1886) textbook[7] on this topic is worth men-

[7] Obviously, Gouraud (1848) published the first book on the history of probability, but Stigler (1978) says that it is outdated. Todhunter's book was the first comprehensive account of the history of probability.

tioning. Statistical inference is not treated much in these books, and statistical inference for finite populations is not touched on at all. Dale (1999) analyses the history of inverse probability, and hence the history of statistical inference, from a general point of view, but not specifically in connection with finite populations. Characteristically, all these authors focus their attention on developments in the 19th century or earlier, and they only briefly comment on developments after the beginning of the 20th century.

Recently, a number of textbooks have been published on the contemporary history of statistical science and probability theory. Textbooks by Kruger, et al. (1987 and 1989), Gigerenzen et al. (1989), and Salsburg (2001) give a comprehensive account of the development of statistical methods during the past century. A common feature in all these books is that they do not mention statistical inference for finite populations and survey sampling, or they mention them only superficially. These authors mainly deal with the development since the 1920s and only briefly mention earlier development. Ian Hacking has written several textbooks (e.g., Hacking, 1965, 1975, and 1990) about the philosophy of statistical science and scientific inference, but he only treats statistical inference within a hypothetical population framework.

Besides textbooks, there is an abundance of articles giving historical accounts of the development of probability theory. All the texts that touch on the history of statistical inference before 1930 deal only with statistical inference within infinite hypothetical populations.

Stigler (1986), Hald (1998, 2007), and Dale (1999) all recognize the importance of Laplace's Principle of Inverse Probability in the history of statistical inference. Laplace's mémoire published in 1774 was the first attempt to attack analytically the problem of induction. Later, Laplace wrote two well-known textbooks on probability (Laplace 1812 and 1814), which were frequently referred to by mathematicians in the first half of the 19th century. His most famous followers were Siméon-Denis Poisson and Adolphe Quetelet, who both strongly fostered Laplacian science. Later, Quetelet wrote a very popular book on probability (Quetelet 1849), which was based on Laplace's ideas. This book, "*Quetelet's letters*", was the basis for subsequent developments for Francis Galton, among others.

After Laplace, the problems of partial investigation were not noticeably treated. More than a century after Laplace's contributions, Arthur Bowley derived formulas for both random sampling and purposive selection (Bowley 1926). He applied Laplacian methodology in deriving the formulas for random sampling. Bowley also introduced formulas for proportional stratification and pointed out the circumstances when it was gainful. Obviously, it was the first English text on sampling theory. Bowley's impact on survey sampling is the topic of Chapter 8.

The decade of 1920–1930 can be regarded as a watershed in the development of statistical theory. All English statisticians before that, including Karl Pearson, Gosset, Edgeworth, and Bowley, were working from the Laplace theory (or paradigm). In the 1920s, R.A. Fisher sharply attacked that theory, especially the method of inverse probability, and presented his estimation theory. In doing that, Fisher completely renewed statistical theory. Later, Fisher presented his method of statistical inference that he called fiducial inference. It seems that Fisher developed the theory alone and outside academia while working at the

Rothamsted experimental station. Fisher did not contribute directly to the finite population inference, but his indirect influence was vital. Fisher's contributions and his influence on survey sampling is analysed in Chapter 9.

A common conception is that Jerzy Neyman is the architect of the theory of statistical inference for finite populations. In the 1930s, he wrote three papers on statistical inference for finite populations and sampling theory in which established the foundations of modern statistical sampling theory. Neyman applied Fisher's estimation theory and inference method (repeated samples from the same population) to sampling from finite populations. The impetus for Neyman came from Bowley's mèmoire to the ISI (Neyman 1934), and his main target was the purposive selection that Bowley had presented. Only later, Neyman presented a mathematically sound theory for inference within a finite population framework (Neyman 1937). Neyman's works are described and analysed in Chapter 10.

Neyman discarded Fisher's inference model, inductive reasoning, and instead proposed inductive behaviour. It has proved to be essential for the development survey method because it gives a quick and objective interpretation to epistemological probability. Survey statisticians soon unequivocally accepted this and discussion about the nature of inductive inference disappeared from the discussion on statistical inference and from the sampling literature.

Neyman's third paper on survey sampling, published in 1938, directly addressed a practical problem in undertaking a survey in a large human population. Only that paper gave tools to design complex sample surveys. After Neyman's contribution, a period of rapid development of sampling methods started.

Hansen and Hurwitz started to develop a new sampling design for the CPS, which was based completely on probability sampling. The result of this work is best documented in the paper titled "*On the Theory of Sampling from Finite Populations*" by Hansen and Hurwitz (1943). They developed the basic theory of stratified two-stage sampling with one primary sampling unit (PSU) within each stratum drawn with a probability proportional to size measure (PPS sampling) and then sub-sampled at a rate that ensured self-weighting within strata. In Hansen's and Hurwitz' (ibid.) method, the possibility to have varying probabilities in selecting sampling units was explicitly articulated for the first time. Unequal inclusion probabilities were already implicitly present in Neyman's optimal allocation designs, although he did not pay attention to that.

An important breakthrough in the classical theory of survey sampling theory was the paper of Horvitz and Thompson in 1952 on a general theory for constructing unbiased estimates (Horvitz and Thompson 1952). Hansen and Hurwitz (1943) obtained results on sampling with probability proportional to size and with replacement. Horvitz and Thompson extended this idea to sampling without replacement. After the Horvitz and Thompson paper, development was very rapid, and by the first half of the 1950s, the classical sampling theory was established. The final formulation of modern sampling techniques is analysed in Chapter 12.

A peculiar detail is that in the 1930s and 1940s, Cochran did not contribute on sampling from finite (human) populations, although he has been referred to frequently. In modern terminology, Cochran's approach in his early papers was model-based. Only later did Cochran develop methods for finite and fixed populations. In his famous textbook (Cochran 1953), he deals with sampling from finite (human) populations, and his approach is design-based.

1.6 Paradigm shifts in the development of scientific disciplines

According to popular conception, science is supposed to be a steady and cumulative acquisition of knowledge where new findings and results of experiments are added to previous knowledge to form more accurate or extensive theories, and typical scientists are considered to be objective and independent thinkers.

In his famous book, *The Structure of Scientific Revolutions*, Thomas Kuhn (Kuhn 1962) brought this view under suspicion. He argued that scientific research and thought are defined by paradigms, or conceptual world views, which consist of formal theories, classic experiments, and trusted methods. Scientists typically accept a prevailing paradigm and try to extend its scope by refining theories, explaining puzzling data, and establishing more precise measures of standards.

The thesis of Kuhn was that scientific disciplines, once they have emerged from the pre-paradigmatic stage, undergo periods of so-called normal science, which allow them to obtain rapidly a high degree of precision and progress. During the period of normal science, acquisition of knowledge is a more or less steady and cumulative process. Normal science is dependent on the adoption of a universally accepted paradigm that defines research problems for the scientist, tells him or her what to expect, and provides the methods that he or she will use in solving them.

For Kuhn 'normal science' meant research based on past achievements that a scientific community acknowledges for a time as supplying the foundation for its further practice. Today such achievements are told and described both in elementary and advanced textbooks. These textbooks explain the body of accepted theory, illustrate many of its applications, and compare these applications with exemplary observations and experiments.

Kuhn says that he used the term 'paradigm' in two different senses. On the one hand, it stands for entire constellation of beliefs, values, techniques, and so on shared by the members of a scientific community. On the other, it denotes one sort of elements in that constellation, the concrete solutions which, employed as models or examples, can replace explicit rules as a basis for the solution of the remaining problems of normal science.

The first sense Kuhn called sociological and says that the definition is intrinsically circular. A paradigm is what the members of a scientific community share, and, conversely, a scientific community consists of men and women who share the paradigm. He claims, however, that all circularities are not vicious and defends his argument at length.

In the course of research, scientists stumble upon anomalies that the paradigm is unable to explain. If the paradigm repeatedly fails to explain the anomaly, a crisis ensues and alternative theories develop. Eventually a competing theory proves relatively successful in explaining the anomaly and it replaces the old paradigm. Kuhn called this replacement a scientific revolution or paradigm shift. At first, the scientific community resists the replacement, but with time, the success of the new paradigm gains enough support to win out. The scientists within the discipline thus see the world in a different way than it "was" under the old paradigm. Once the old paradigm is replaced and the revolution has ended, normal science re-emerges only to await the discovery of new anomalies.

Kuhn also claims that the ultimate resolution of the conflict between competing paradigms is not wholly the result of reasoning and comparative analysis; it is also affected by external factors. In the process of debating the merits of respective paradigms, theorists tend to "talk through" each other. Kuhn argued that a scientific revolution is a non-cumulative developmental episode in which an older paradigm is replaced in whole or in part by an incompatible new one. However, the new paradigm cannot build on the preceding one. Rather, it can only supplant it, for "the normal-scientific tradition that emerges from a scientific revolution is not only incompatible but actually incommensurable with that which has gone before." (Kuhn 1962)

In addition, Kuhn argued that typical scientists tend to be conservative individuals in the sense that they agree to what they have been taught and apply their knowledge to solving the problems that their theories point out. The training of new scientists partly aims at teaching the paradigm so that they would continue to foster the established tradition. The study of paradigms is what mainly prepares the student for membership in the scientific community. Scientists whose research is based on shared paradigms apply the same rules and standards for scientific practice. This dedication is a prerequisite for normal science, i.e., for the genesis and continuation of a particular research tradition. From the very beginning, new scientists are indoctrinated to the prevailing paradigm. Only young scientists who are not yet so deeply indoctrinated into accepted theories – like Newton, Lavoisier, or Einstein – can manage to sweep an old paradigm away.

One should also bear in mind that paradigms can also exist on a smaller scale. Probably in any science, one can identify "sub paradigms" and "sub paradigm shifts" within a general paradigm. A new paradigm can also surface when two or more related paradigms merge.

Kuhn was mainly thinking about natural sciences and astronomy (he received his Ph. D. in physics), and he illustrated his theory of the evolution of science with examples from the physical sciences. As examples of major paradigm shifts, Kuhn mentions the overthrow of Ptolemaic cosmology by Copernican heliocentrism, and the displacement of Newtonian mechanics by quantum physics and general relativity.

Kuhn's book has revolutionized the history and philosophy of science, and his concept of paradigm shifts has been extended to such disciplines as political science, economics, and sociology. Can Kuhn's theory also be applied to formal or methodological sciences like statistical science? Many formal or methodological sciences are closer to the traditional conception of science with gradual progress and cumulative acquisition of knowledge because of their deeply deductive natures.

In a methodological discipline, an anomaly is rarely something that the paradigm cannot explain. Rather it is the inability to solve or explain certain problems, or the paradigm is not able to respond to the needs. The anomalies that methods face may be in their capabilities to provide answers to questions arising in other (real) sciences.

In mathematics, including probability theory, it is difficult to imagine a revolution like the one that took place in physics. Statistical inference is inherently based on probability theory, but it differs from many other methodological disciplines due to its intimate connection to inductive reasoning. It is not straightforward how to attach to an estimate a measure of probability, which indicates how certain it is that the estimate reflects the true state of nature. Real world observations cannot be linked with probability distributions using merely deductive reasoning.

1.6.1 Research on paradigms in survey sampling

The history of statistics or statistical science has not been analysed much in respect to paradigms, but it is not a totally nonexistent topic. A central theme in many writings on the history of survey sampling has been the use of randomization in sampling. In some papers, for example, Brewer (1999) and Bellhouse (1988), this theme was prominent. A typical feature in nearly all papers, not only in the two mentioned, is that randomisation is assessed from a current perception of survey sampling techniques rather than through an attempt to figure out the reasoning in each epoch.

Brewer (1999) divided the history of survey sampling into three parts. He argues that in the first part, from the end of the 19th century up to around 1945, survey designers could select between randomisation sampling and purposive sampling "...on an arbitrary basis, apparently without serious fear of criticism". During the next 25 years, Brewer claims, the random selection of samples went virtually unchallenged. Then during the 1970s, the choice re-emerged in the form of balanced sampling. Brewer (ibid.) calls the first period 'pre-paradigmatic' in the sense Kuhn defined it. The next period, according to Brewer, was dominated by the randomisation paradigm.

In Kuhn's theory, the paradigms are closely related to 'normal science'. During periods of normal science, the primary task of scientists is to bring accepted theory and facts into closer agreement or in methodological sciences, to develop the accepted methods to better address practical needs. During these periods, the scientific community works from a single paradigm or from a closely related set. Kuhn's argument that a scientific community is defined by its adherence to a single paradigm implies that at the pre-paradigmatic (or multi-paradigmatic) phase of a discipline, scientists do not form a truly scientific community[8].

Is Brewer's characterization justified? Obviously it is, if the only criterion considered relevant is whether selection probabilities of sample units are used in statistical inference (implying randomisation in the selection of a sample). However, sample surveys were already an important method to collect data during the period Brewer called pre-paradigmatic. Did they lack scientific basis?

Probability as a central factor in statistical inference is a distinct issue from that of random selection of units. Random selection as a method providing rep-

[8] Kuhn also suggested that questions about whether a discipline is or is not a science can be answered only when members of a community who doubt their status achieve a consensus about their past and present accomplishments.

resentative samples has a considerably longer history than its role in estimation (see Bowley 1906). Brewer's evaluation of the historical development of survey sampling seems slightly contemptuous because it is quite obvious that the virtues of random selection were already known before Hansen and Hurwitz designed the sampling scheme for the Current Population Survey (Hansen and Hurwitz 1943).

In the beginning of the 20th century, there were serious statisticians and scientists who applied partial investigations, and it seems an underestimation of their professional ethics to say that they did not use scientific methods, even though their methods were different from the methods that are currently applied. Brewer talks about randomization induced by selection probabilities of sampling units, but there are also other lines of thought about how probability is assumed to enter the inference setup. Brewer's claim that the sampling method could be selected freely between random and purposive sampling on an arbitrary basis without fear of criticism appears somewhat doubtful.

Bellhouse (1988) argues that the initial paradigm in survey sampling is that of the desire to collect a representative sample as presented by Anders Kiaer in the 1890s. Bellhouse also says that there are earlier examples of partial investigations but that they illustrate the randomness in research as is typical for the pre-paradigmatic times. Bellhouse also analyses the development of survey sampling in relation to the adoption of randomisation in statistical inference. Another paradigm that Bellhouse identifies is the one starting from Neyman's paper in 1934. In Bellhouse's mind, the reasons are twofold: the first is that by that paper, randomization was pointed to as the recommended solution in sample selection (and the problems of purposive selection were shown indisputably); the second reason was that it provides a theory of point and interval estimation under randomisation. An apparent question is: did not there exist formalized statistical inference for finite populations before Neyman?

Kish (2002) argues that sampling is a branch of and a tool for statistics, and that field of statistics was founded as a new paradigm in 1810 by Quetelet. That happened, according to Kish, when the predictable, meaningful and useful regularities in the behaviour of population aggregates of less predictable individuals were named "statistics". Kish (ibid.) maintains that it was a great discovery at that time. Kish's arguments do not seem to be well-grounded, however. Quetelet never carried out a survey or partial investigation because he believed that populations are so heterogeneous that only a full enumeration could be representative (see Chapter 6).

Although paradigms of randomisation have been a topic in some studies, the development of statistical inference through the history has not been analysed in respect to paradigms. The documented history suggests that there occurred paradigm shifts in the method called statistical inference, first because of the contributions of Ronald Fisher, and then by Jerzy Neyman. Neyman's work resulted in a new paradigm in statistical inference for fixed populations as Bellhouse (Bellhouse 1988) argued. However, there existed a method for statistical inference before Fisher's and Neyman's contributions. The paradigm that Neyman initiated replaced an older paradigm that was based on Laplace's Principle of Inverse Probability.

Another paradigm shift took place when Anders Kiaer introduced the Representative Method to be used instead of censuses to explore population characteristics. It was a revolution in the Kuhnian sense, but it was not truly a paradigm shift because the representative method and census continued to exist side by side. This was also Kiaer's intention.

The view of randomization that the papers of Brewer and Bellhouse reflect seems ill-founded. The merits of randomization in sample selection were acknowledged already at an early stage. Before the modern age, randomization was not applied in sampling from human populations because of practical reasons. Drawing a (simple) random sample was not possible or it was too difficult, and data collection from a truly random sample had been too labour- and time-consuming. Another aspect of randomization is its role in statistical inference. The argumentation about randomisation inference has been slightly confused: what is meant by randomisation inference and when did it start?

2 Origins of statistical science

Some years ago, Maurice Kendall wrote an article asking, "Where shall the history of statistics begin?" (Kendall 1960). He opens with a paradox: "A history must start somewhere, but history has no beginning." By this, Kendall means that history is a continuous flow of ideas and activities, which could be followed endlessly. However, a starting point has to be selected in order to be able to assess the paths of development.

What does "statistics" means? In modern usage, "statistics" as a word and as a concept has two different meanings: Statistics may mean a systematic collection of facts and their organization, usually in tables. This is occasionally called **official statistics** because it is typically an activity of administration. The other meaning is related to analyzing the data and drawing conclusions from it. The latter will be called here **statistical science**[9]. Statistical science also includes applications of probability in statistics.

2.1 Early examples of official statistics

Activity that can be described as statistical in the wide sense has been pursued for a very long time. It is possible that all organized societies have practiced some sort of statistical activity. The first known censuses of agriculture were already undertaken in Babylonian times (3000 B.C.). That means that the first statistics were compiled relatively soon after the art of writing was invented. According to Rao (2006), in India a treatise called Arthasastra by Kautilya, probably written during 321–296 B.C., had a detailed description of the system of data collection relating to agricultural, population, and economic censuses in villages and towns during that period. Much later, the tradition of collecting data in detail continued in India during the period around 1590 A.D. Ancient China also counted its people to determine the revenues and the military strength of the different provinces. There are also accounts of statistical overviews compiled by the Egyptian rulers long before Christ. Rome regularly took a census of people and of property. This was used to establish the political status of citizens and to assess their military and tax obligations to the state. There is a very famous example of counting the people of Israel, leading to the birth of Jesus in Bethlehem.

Recently, Missiakoulis (Missiakoulis 2020) has discovered evidence that Cecrops, the legendary first king of Athens, may have take a census of his subjects in the 16th century B.C.: Each person was commanded to cast a single stone on a pile, and by counting the stones, it was established that they were twenty thousand inhabitants. However, Missiakoulis (ibid.) is a little hesitant about the census because Cecrops is a mythical figure, who may or may not have existed in

[9] In later chapters, the word statistics means statistical science unless otherwise indicated explicitly.

person, and the evidence of Cecrops' census is all from classical literary sources and not from archaeological sources.

In the Middle Ages, attempts to conduct a census were rare. The most famous one was the census of England taken by the order of William the Conqueror, King of England. The compilation of this Doomsday Book started in the year 1086 A.D. The book records a wealth of information about each manor and each village in the country.

Another interesting example of the history of official statistics can be found in the Inca Empire that existed between 1000 A.D. and 1500 A.D. in South America. Incas did not have a written language, but each Inca tribe had its own statistician, called the Quipucamayoc. He kept records of the number of people, the number of houses, the number of llamas, the number of marriages, and the number of young men who could be recruited for the army. All these facts were recorded on a quipus, a system of knots in coloured ropes.

At regular intervals, couriers brought the quipus to Cusco, the capital of the kingdom, where all regional statistics were compiled into national statistics. The system of Quipucamayocs and quipus worked well, but the system vanished with the fall of the Inca Empire.

An early census also took place in Canada in 1666. Jean Talon, the intendant of New France (later Quebec), ordered an official census of the colony to measure the increase in population since the founding of Quebec in 1608. The enumeration, which recorded a total of 3,215 persons, included the name, age, sex, marital status, and occupation of every person. The first censuses in Europe were undertaken in the Nordic countries. The first census of modern times was carried out in Iceland in 1703; the first census in Sweden-Finland took place in 1746; and in Denmark-Norway, the first census was done in 1769.

The first statistical account of Scotland has been considered one of the cornerstones of modern statistics. It was undertaken at the end of the eighteenth century under the direction of Sir John Sinclair. It was published in twenty-one volumes between 1791 and 1799. Sinclair's plan involved a pre-planned set of 160 questions sent to all parishes in Scotland: 40 questions covered the geography and topography of the parish, its climate, natural resources, and natural history; 60 questions addressed population and related matters; and the remaining questions concerned the parishes' "agricultural and industrial production" and miscellaneous matters. In 1799, Sinclair was able to lay before the General Assembly "a unique survey of the state of the whole country, locality by locality".

Obviously for a very long time, records of population and related matters have been collected intermittently in a variety of places. Usually, the chief purpose of statistical activity has been the promotion of bureaucratic efficiency. Without detailed records, centralized administration is almost inconceivable.

These examples can be described as bookkeeping of the population in a country without any attempts to reason about the data. They can hardly be called examples of statistical science. Kendall (ibid.) argues that "the true ancestor of modern statistics is not seventeenth-century statistics, but Political Arithmetic". By modern statistics, Kendal meant statistical science.

2.2 Political Arithmetic

William Petty[10] coined the term "political arithmetic" in the middle of the17th century for a discipline of empirical collection of population records and the preparation of accurate life tables. In his view, political arithmetic was an application of "Baconian principles to the art of government" (see Porter 1986). He is best known for his economic history and statistic writings, but he also attempted to do some simple statistical analysis. Petty's work in political arithmetic, along with the work of John Graunt[11], laid the foundation for modern census techniques.

The third important person behind political arithmetic was Edmond Halley[12]. In 1693, Halley published an article on life annuities, which presented an analysis of age-at-death taken from archives in Breslau, which was known for keeping careful and exact records. Halley's work had a strong influence on the development of actuarial science. The construction of the life-table for Breslau has been regarded as a major event in the history of demography.

Political arithmetic had a strong influence on early statistical thinking, and it dominated the thinking right up to the beginning of the nineteenth century, when political arithmetic gradually gave way to the new social science of statistics (social calculus). This coincided with a profound change in the social infrastructure which was caused by the age of industrialisation and as a consequence of rapid urbanisation. According to Porter (ibid.), by that time, statistical writers had become increasingly convinced that society was more than a passive recipient of legislative initiatives. Rather, society was considered dynamic, sometimes intractable, and also possessing some autonomy, and therefore had to be understood before the aims of the state could be put into effect.

2.3 First sample survey

The first documented attempt to make statements about a population by using information only about a part of it was made by John Graunt. In a famous tract (Graunt, 1st edition 1662), he described a method to estimate the population of London, based on a sample. His motivations were somewhat obscure, but obviously the main motive was not to conduct a scientific investigation. In the beginning of his tract he said:

> "I Have been several times in company with men of great experience in this City, and have heard them talk seldom under Millions of People to be in London, all which I was apt enough to believe, untill, on a certain day, one of eminent Reputation was upon occasion asserting, that there was in the year 1661 two Millions of People more than Anno 1625, before the great Plague; I must confess, that, untill this provocation, I had been frighted with that misunderstood Example of

[10] **William Petty** (1620–1683) was an English economist, scientist, and philosopher

[11] **John Graunt** (1620–1674) was an English merchant and a collaborator of William Petty

[12] **Edmond Halley** (1656–1742) was an English astronomer, geophysicist, mathematician, meteorologist, and physicist.

> David, from attempting any computation of the People of this populace place; but hereupon I both examined the lawfulness of making such enquiries, and, being satisfied thereof, went about the work itself in this manner: " (Graunt 1662)

Graunt's survey was based on the fact that from the beginning of the 17th century in England, parishes were obliged to keep records on births, christenings, marriages, and burials. He surveyed families in a sample of parishes where the registers were well kept. He found out that on the average there were 3 burials per year in 11 families. Assuming this ratio to be more or less constant for all parishes, and knowing the total number of burials per year in London to be about 13,000, he concluded that the total number of families was approximately 48,000. Obviously, this estimation is the first documented example of using a ratio estimator. Putting the average family size at 8, he estimated the population of London to be 384,000. The estimated number of inhabitants in London was one-third of what was commonly believed.

Although Graunt was aware of the fact that averages like the number of burials per families and family sizes varied in space and time, he did not make any provisions for it. In Graunt's time, there were no known methods to take into account such a variation in estimation. On the other hand, he tried to verify his estimate by calculating it with another method.

> "And lastly I took the Map of London set out in the year 1658 by Richard Newcourt, drawn by a scale of Yards. Now I guessed that in 100 yards square there might be about 54 Families, supposing every house to be 20 foot in the front: for on two sides of the said square there will be 100 yards of housing in each, and in the two other sides 80 each; in all 360 yards: that is 54 Families in each square, of which there are 220 within the Walls, making in all 11880 Families within the Walls. But forasmuch as there dy within the Walls about 3200 per Annum, and in the whole about 13000; it follows, that the housing within the Walls is 1/4. part of the whole, and consequently, that there are 47520 Families in, and about London, which agrees well enough with all my former computations: the worst whereof doth sufficiently demonstrate, that there are no Millions of People in London, which nevertheless most men do believe, as they do, that there be three Women for one Man, whereas there are fourteen Men for thirteen Women, as elsewhere hath been said." (Graunt 1662)

Apart from using a ratio estimator, Graunt made two significant inventions, which later appeared to be important in survey sampling. First he observed, and rested on, the fact that some social and demographic indicators and ratios remained stable in time and space. Stability of social phenomena is an essential assumption, without which social surveys would not be justified. For example, he observed that nearly the same proportion of boys and girls were born – but slightly more boys. This proportion remained constant in all parishes in London and the surrounding countryside and over time. These facts were not known before Graunt established those using church records. Graunt's estimates had not been plausible without the awareness of the stability of the ratios.

Graunt's second invention was to use averages to estimate total values. Essential to his method was the observation that the proportion of burials in a year remained around 3 to 11 families, and that the average family size was 8 persons. These averages were first expanded to estimate the number of families and then the number of people.

2.4 The origin of averages in estimation

The use of averages as the basis of estimation of the total amount is central in sample surveys. The way in which Graunt used average values in estimation was ingenious, but the very use of average values has a longer history. Actually, the idea of combining observations to reduce variation is very old. According to Plackett (1958), the problem of estimating parameters from observations appears to have been presented itself already to the Babylonian astronomers a few centuries B.C. Between 500 and 300 B.C. they developed a systematic mathematical theory to account for the motions of the sun, moon, and planets. Also the Greeks had mathematical techniques to combine observations in astronomy. According to Plackett (ibid.), the technique of repeating and combining observations made on the same quantity appears to have been introduced into the scientific method by Tycho Brahe[13] at the end of the 16th century.

[13] **Tycho Brahe** (1546–1601), born **Tyge Ottesen Brahe**, was a Danish nobleman known for his accurate and comprehensive astronomical and planetary observations. He adopted the Latinized name «Tycho» at around age fifteen. In his lifetime, Tycho was well known as an astronomer and alchemist.

3 Inverse probability and Bayes' Theorem

3.1 Beginnings of probability theory

The origins of probability theory can be found in the period from 1650 to 1700 in the mathematical analyses of "games of chance" and in the systematic study of mortality data. A possibility to have a better understanding of future events – or maybe to predict them – has always enchanted people, but gambling has a special role because of the immediate benefits. Therefore, analysis of the games of chance was the source of the greatest impetus for the early contributions to the probability calculus. The second half of the 17th century has come to be known as the age of scientific revolution. The thoughts of famous scientists such as Newton, Leibniz, and Halley indirectly also paved the way for the development of probability theory (see Hald 1990). Especially Newton's contributions have been regarded as being important because his research eventually brought about a new world view and his philosophy of science dominated the intellectual world for more than two centuries.

At the end of the 17th century, the central problem in probability theory was the calculation of probabilities in different (game) events. The main interest was in the problems of direct probability. That consisted of descriptions of the distributions of outcomes of experiments, which were composed of equally likely simple events. Hald (1990) claims that during the decade from 1708 to 1718, there was a great leap forward (in probability theory) because of the great number of significant contributions published in that period. Probability theory expanded greatly from its original questions.

JACOBI BERNOULLI,
Profess. Basil. & utriusque Societ. Reg. Scientiar.
Gall. & Pruss. Sodal.
MATHEMATICI CELEBERRIMI,
ARS CONJECTANDI,
OPUS POSTHUMUM.
Accedit
TRACTATUS
DE SERIEBUS INFINITIS,
Et Epistola Gallicè scripta
DE LUDO PILÆ
RETICULARIS.

BASILEÆ,
Impensis THURNISIORUM, Fratrum.
cIↃ IↃCC XIII.

Figure 3.1:
The cover page of Ars Conjectandi.

Jakob Bernoulli's[14] famous book *Ars Conjectandi* was published in

[14] Jakob Bernoulli (1654–1705), is also known by names Jacob, Jacques and James Bernoulli, was a Swiss mathematician and scientist and was one of the many prominent mathematicians in the Bernoulli family. Jakob Bernoulli studied theology and entered the ministry. But he also studied mathematics and astronomy. He traveled throughout Europe from 1676 to 1682, learning about the latest discoveries in mathematics and the sciences. He became familiar with calculus through a correspondence with Gottfried Leibniz, then collaborated with his brother Johann on various applications. In 1690, Jakob Bernoulli became the first person to develop the technique for solving separable differential equations. Upon returning to Basel in 1682, he founded a school for mathematics and the sciences. He was appointed professor of mathematics at the University of Basel in 1687, remaining in this position for the rest of his life

1713[15], and it turned out to be very influential all over Europe. In this book, Bernoulli analyzed probability and stochastic phenomena (as they would be called today) from a wider perspective and more systematically than his precursors had done (see Stigler 1986 and Hald 1990). Bernoulli created many central concepts that have remained in probability theory. Maybe the best known is the Bernoulli trial. He also coined the terms *a priori* and *a posteriori* to distinguish two ways of deriving probabilities. He presented several central inventions in *Ars Conjectandi*, but the most influential was the Law of Large Numbers[16]. Bernoulli's book is claimed to be the first systematic treatment on probability theory and to be groundbreaking on the topic (see Stigler 1986). It was soon translated into English, and it had a strong influence on English mathematicians such as Abraham De Moivre[17] and Thomas Simpson, and obviously it eventually led to Thomas Bayes' Essay (see Schneider 2006).

Five years after *Ars Conjectandi* was published, Abraham de Moivre published a book entitled *The Doctrine of Chances*[18]. De Moivre drew from the results that Bernoulli had attained, but he also elaborated many of those problems that Bernoulli was not able to solve. Eventually, de Moivre also managed to solve some of the mathematical problems Bernoulli could not (see Stigler, ibid.). Published in 1738, the second edition of de Moivre's book was considerably more elaborate than the first edition: In it he already foreshadowed the idea of normal distribution as an approximation of binomial distribution, but he did not firmly establish it. That was done by Laplace a half century later, and therefore this approximation has occasionally been called the de Moivre-Laplace Theorem[19]. The works of de Moivre had a profound influence in the 18th century, and many of the publications on probability theory have been said to take up their motivation from his writings. The title of de Moivre's book came to be synonymous with probability theory (see Stigler, ibid.). That was also the source for the title used in Bayes' essay (*"An Essay Toward Solving a Problem in the Doctrine of Chances"*).

Although mathematicians during the first half of the 18th century mainly dealt with problems of direct probability induced by the games of chance, there were

[15] In fact, Jakob Bernoulli (1654–1705) wrote the book already in 1705, but it was published eight years after his death by his nephew Nicholas.

[16] The Law of Large Numbers says that in repeated, independent trials with the same probability p of success in each trial, the chance that the percentage of successes differs from the probability p by more than a fixed positive amount, $\varepsilon > 0$, converges to zero as the number of trials n goes to infinity, for every positive ε. It follows from the law that the empirical probability of success in a series of Bernoulli trials will converge to the theoretical probability.

Originally the law was called Bernoulli's Theorem. In 1835, Poisson elaborated it further and coined the name «*La loi des grands nombres*» («Thc law of large numbers»). After Bernoulli and Poisson, especially Russian mathematicians contributed to refinement of the law, including Chebyshev, Markov, Kolmogorov, and Khinchin..

[17] **Abraham De Moivre** (1667–1754) was born in France but he fled to England because of religious reasons at the age of 21 after being in prison for two years. He stayed in England the rest of his life and wrote only in English, except one paper in Latin.

[18] The whole title of the book was "The Doctrine of Chances: or a Method of Calculating the Probability of Events in Play"

[19] Karl Pearson held the association of the Normal curve to Gauss one of the four capital flaws in the history of statistics. He said (see Pearson 1978) "There is a fundamental curve in statistical theory which goes by the name Gauss. Laplace discovered it ten years at least before Gauss, and its real discoverer was De Moivre – some half-century before Laplace."

also more "serious" areas of application. They were motivated by problems from annuities, insurance, or sciences like meteorology and especially astronomy. In annuities and in the insurance business, direct probability played an important role because probability theory provided a tool to determine the risks and payments. In empirical research, or for example in astronomy, direct probability usually is not sufficient. Often the problem is the reverse: based on observations at hand, what can be concluded about the cause system that has brought about these observations? In their books, both Bernoulli and de Moivre had trains of thoughts concerning inverse probability in the context of games of chance (see Dale 1999), but they never explicitly treated the topic. Only after Hume had published his critique on the inductive method in 1748 did mathematicians become captivated with questions about inductive inference and inverse probability.

R.A. Fisher claimed that Bayes' Essay was the first attempt to systematically treat inductive reasoning (see Fisher 1935). On the other hand, Stigler (1986) claims that the first attempts to deal with the problem took place in England by two persons: Thomas Simpson and Thomas Bayes. Bayes' formula is one of the best known formulas in probability theory, and it is the first formula to explicitly address the inverse problem. Simpson published his papers before Bayes did, but they addressed a slightly different inference problem than Bayes did in his Essay.

As the topic of the thesis is the path how statistical inference, or inverse probability, for finite populations has developed into a mature discipline, a natural starting point is the work of both Thomas Simpson and Thomas Bayes.

3.2 Simpson's analysis of error

Next to de Moivre, Thomas Simpson[20] was the most important writer on probability theory and actuarial mathematics in Britain in the first half of the 18th century (Hald 1998). He wrote two textbooks in 1740s which were partly based on the early works of de Moivre. Stigler claims that Simpson's motivation was to popularize de Moivre's works (Stigler 1986). However, Simpson also made an original contribution on statistical error theory, which has proved to be significant for the later development of statistical inference.

The Simpson's work on error distribution was written in the form of a letter, which was read to the Royal Society in 1755. It carried the title: "*On the Advantage of Taking the Mean of a Number of Observations in practical Astronomy*". *(Simpson 1755)*

Simpson's treatment of this problem was fairly limited. He showed that it is better to take a mean than a single observation, provided that the mean is based on 6 measurements. The idea to combine observations in a mean to reduce the influence of variation was not new. The novelty in Simpson's idea was that it was applied to inaccurate observations (in astronomy). He started his letter, "... in order to diminish the error arising from the imperfection of instruments, and of the organs

[20] **Thomas Simpson** (1710 – 1761) was and English mathematician and inventor. Besides his books on probability, he is known by the Simpson's rule to approximate definite integrals.

of sense, by taking a Mean of several observations…". In this short letter, Simpson focused on two important points: the conceptual and technical developments.

In the conceptual development, Simpson did not focus on observations, i.e., the actual position of the astronomical object, but on the errors made in the observations, i.e., the difference between the recorded observations and the actual position of the observed object. According to Stigler (ibid.) this was a critical opening in the mid-eighteenth century mathematics that lacked theories of inference. It was the first step to a more successful analysis of uncertainty.

Simpson assumed a specific distribution for the errors. Because of the known error distribution, he was able to focus his attention on the mean error instead of the mean observation. This way, he could avoid the problem arising from the stochastic structure for the unknown position. The induced inferential problem was similar to what R. A. Fisher later called the fiducial argument[21].

Simpson supposed that each of the n independent observations was susceptible to the errors (with discrete values)

$$-v, -v+1, \dots, -2, -1, 0, 1, 2, \dots, v-1, v$$

and the probability distribution of the errors proportional to either

$$r^{-v}, \dots r^{-2}, r^{-1}, r^{0}, \ r^{1}, r^{2}, \dots r^{v} \text{ or}$$

$$r^{-v}, 2r^{-v+1}, 3r^{-v+2}, \dots, (v+1)r^{0}, \ \dots, 3r^{v-2}, 2r^{v-1}, r^{v}, \text{ where } r > 0.$$

The error distributions were derived from those of de Moivre, but the use to which Simpson put them was new. He was actually interested in the case where r=1 (i.e., symmetric error distribution). Stigler (Stigler, ibid.) claims that Simpson did not recognize the advantage of writing it in a more general way. According to Stigler, this way he had been able to anticipate the use of generating functions in statistics, which was the technique to be used later in the works of Lagrange and Laplace.

Later Lagrange embraced the idea of Simpson and presented a detail discussion of discrete error distributions on the lines essentially the same as those followed by Simpson (see Plackett 1958). Lagrange also purported to show that the mode of the distribution of sample means is the same as the populations mean. Lagrange's mathematical developments and results were appreciated by Laplace who subsequently made the technique a basic part of his attack on the problem of combining observations and the analysis of error (see Plackett, ibid.).

[21] Fisher's argument runs as follows: If e represents the error, O the observation and P the point observed, then O = P + e can be written as well P = O – e. Fisher argued: "Its is not important which end is considered as fixed." In fiducial argument the symmetrical difference, or the error, e = O – P is treated as randomly distributed (see Chapter 9).

3.3 Bayes' inverse probability

In its modern form[22], for the discrete case, Bayes' Theorem is:

$$P(a_i \mid E) = \frac{P(E \mid a_i)P(a_i)}{\sum_i P(E \mid a_i)P(a_i)} \tag{3.1}$$

In the current literature, the formula is said to provide an answer to the question: what is the probability of an "event" a_i if we know that "event" E has happened. In the common mathematical interpretation of the theorem, a_i corresponds to outcomes at some intermediate stage of a compound experiment, or events, and E to some final outcome that is readily observed.

The probability $P(a_i)$ is often called the *a priori* probability of a_i. A special situation where all events a_i are equally likely is found by setting all $P(a_i)$ equal to each other. Then $P(a_i)$ factors out of the denominator and cancels the term in the numerator, and the formula gives the same result without any a priori probabilities $P(a_i)$:

$$P(a_i \mid E) = \frac{P(E \mid a_i)}{\sum_i P(E \mid a_i)} \tag{3.2}$$

In some writings, the Bayes' formula has been given an epistemological interpretation: given that event E has occurred, the probability that it was due to the hypothetical cause a_i is equal to the probability that a_i should produce the event times the probability that a_i should occur in the first place, all divided by a scaling factor that is equal to the sum of such terms over all i's. This interpretation involves the idea of the inverse probability which is thought to address the question: Given that an event that may have been the result of any of two or more causes has occurred, what is the probability that the event was the result of a particular cause? Bayes himself did not explicitly mention this epistemological interpretation, although that may have been his ultimate motivation (see Hald 1998).

In the continuous case of Bayes' rule, there is a continuous range of possible causes with a continuous range of probabilities, θ, ranging from 0 to 1. The probability that one particular cause should produce p successes in $n = p + q$ trials is

$$\frac{(p+q)!}{p!q!}\theta^p(1-\theta)^q$$

This is divided by the sum of all possible causes, which in the continuous case becomes an integral. The continuous form of the Bayes' rule is usually written as

$$P(S) = \frac{\theta^p(1-\theta)^q}{\int_0^1 \theta^p(1-\theta)^q d\theta} \tag{3.3}$$

[22] The modern forms of the Bayes' Theorem have been given only after his death. In the Essay, they cannot be found. According to Fienberg (2006) the modern formula was introduces in the beginning of 19th century.

The combinatorial coefficients in the numerator and denominator are the same and have been cancelled out.

3.3.1 The Bayes' Essay

It is well known that the Presbyterian minister Thomas Bayes (1701–1761) invented the Bayes' Theorem. It was published as a paper often called the Essay of Bayes (Bayes 1763). Its complete title was *"An Essay Toward Solving a Problem in the Doctrine of Chances"*[23]. Stigler (1986) claims that he has found some evidence indicating that Bayes got interested in this topic because of Simpson's earlier paper (see also Bellhouse 2004).

Bayes never published the Essay by himself, though. Richard Price, another Presbyterian minister and a friend of Bayes, found it from his left property and sent it to the Royal Society two years after Bayes' death. Although Price is not as well known in the statistical literature as Bayes, he was not a layman in what comes to probability. Already the fact that he realized the value of the Essay proves it. In addition, he wrote an introduction and an appendix in the Essay, commented and explained Bayes results, and obviously added a few details to the text (see Hald 1998 or Dale 1999). There have even been some speculations about Price's role and about how much he really changed the text (see Fienberg 2006), and even about the true origin of the Essay (see Stigler 1983). The Price's Appendix was called "*Containing an Application of the foregoing Rules to some particular Cases*". In that, he discusses a number of examples which illustrate the use of the results of the Essay. Dale (ibid.) gives a detailed analysis of Price's Appendix. Also Hald (2007) makes interesting comments about it.

Bayes started the Essay by defining the problem that he was addressing and went on to prove nine propositions which eventually solved the problem:

> "**Problem**: Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named."

In modern notation, the problem could be written as follows: Let X be the number of times the event happens in n trials and θ the probability it happens in a single trial. Bayes sought to find the probability $P(b \leq \theta \leq f \mid X)$ for any given b and f.

The formulation of the problem aims explicitly at finding an interval estimate for an unknown parameter based on observations. Price observed the importance of this problem and he wrote a long paragraph devoted to a discussion of this matter. He notes that the discussion of the problem was needed to determine "in what degree repeated experiments confirm a conclusion". Later he mentions that the problem "is necessary to be considered by any one who

[23] Bayes' Essay was first published in 1763 but it has been published later in Biometrika in 1958 accompanied with biographical notes by G. A. Barnard (Bayes 1958). The Essay can be found also from the Internet.

would give a clear account of the strength of *analogical* or *inductive reasoning.*" In effect, Price was saying that solving the problem was essential to the solving of the problem of induction (see also Hacking 1975).

It is not known whether the problem that Bayes put forward was new, but at least it had not been solved before. One reason why Bernoulli or de Moivre did not treat the problem may have been the fact that at the beginning of the 18th century the problem of induction had not been brought up as a central problem of philosophy (see Hacking, ibid.).

At the time when Bayes prepared his Essay, the concept of probability was vague, and mathematicians gave the concept different meanings or they did not give it a definite meaning at all. Wanting to make his point clear, Bayes first defined what he meant by probability. He defined the "ground laws" in the following manner:

> "**Definition 1.** Several events are inconsistent, when if one of them happens, none of the rest can.
> 2. Two events are contrary when one, or other of them must; and both together cannot happen.
> 3. An event has said to fail, when it cannot happen; or, which comes to the same thing, when its contrary has happened.
> 4. An event is said to be determined when it has either happened or failed.
> 5. The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon it's happening.
> 6. By chance I mean the same as probability.
> 7. Events are independent when the happening of any one of them does neither increase nor abate the probability of the rest."

In the introduction to the Essay, Price wrote: "He [Bayes] has also made an apology for the peculiar definition he has given to the word chance or probability. His design herein was to cut off all dispute about the meaning of the word, ... of the proper sense of the word probability, he has given that which all will allow to be its proper measure in every case where the word is used."

Bayes' Essay was composed of what he called propositions and some of their corollaries. Only in Proposition 9 does he address the problem and derive a solution to it using the previous propositions. Propositions 1 through 8 define rules of probability calculus and define the intermediate results needed to establish Proposition 9.

> **Proposition 1**: "When several events are inconsistent the probability of the happening of one or other of them is the sum of the probabilities of each of them."
> **Proposition 2**: "If a person has an expectation depending on the happening of an event, the probability of the event is to the probability of its failure as his loss if it fails to his gain if it happens."
> **Proposition 3**: "The probability that two subsequent events will both happen is a ratio compound of the probability of the 1st, and the probability of the 2d on supposition the 1st happens."
> **Corollary**: "Hence if of the two subsequent events the probability of the 1st be a/N and the probability of both together be P/N, then the probability of the 2d on supposition the 1st happens is P/a."

In modern notation, this means that if E_1 and E_2 are any two events, ordered in time so that E_1 happens earlier than E_2, then

$$P(E_2 \mid E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} \tag{3.4}$$

This is the first explicit formulation of conditional probability. It was significant because conditional probability is a central concept in inverse probability.

> **Proposition 4**: "If there be two subsequent events to be determined every day, each day the probability of the 2d is *b/N* and the probability of both *P/N*, and I am to receive *N* if both the events happen the 1st day on which the 2d does; I say, according to these conditions, the probability of my obtaining *N* is *P/b*."
> **Proposition 5**: "If there be two subsequent events, the probability of the 2d b/N and the probability of both together P/N, and it being 1st discovered the 2d event has happened, from hence I guess that the 1st event has also happened, the probability I am in the right is P/b."

In modern notation, this proposition says that if E_1 and E_2 are two events, ordered in time $(E_1 < E_2)$, then

$$P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} \tag{3.5}$$

This proposition addresses the problem of induction, or inverse problem[24], as it gives a probability for an unknown event based on an observed event. Proposition 5 can be considered as the first formulation of the Bayes' Theorem. Hald (1998) regarded this proposition as Bayes' most important result — not because of Bayes' proof, but because of his interpretation and because its application to statistical inference. It is important to note that Bayes regarded propositions 3 and 5 fundamentally different. Shafer (1982) has given a careful analysis of Bayes' reasoning between these two propositions.

> **Proposition 6**: "The probability that several independent events shall all happen is the ratio compounded of the probabilities of each."

Bayes gave two corollaries to this proposition dealing with failures of events: "Failure of an event may always be considered as the happening of its contrary."

> **Proposition 7**: "If the probability of an event be a, that of its failure be b in each single trial, the probability of its happening p times, and failing q times in $p + q$ trials is $\underline{Ea^pb^q}$ if E be the coefficient of the term in which occurs when the binomial $\overline{a + b}|^{p+q}$ is expanded."

In essence, Bayes derived the binominal distribution in this proposition with the help of proposition 6. In modern notation, when $n = p + q$ and probability of success is θ, it can be written

[24] Bayes spoke about *converse* problem, not inverse problem.

$$P(p,q \mid \theta) = \binom{n}{p} \theta^{p} (1-\theta)^{q} \tag{3.6}$$

To solve the initial problem, Bayes created a though-experiment which was based on the positions of balls on a flat square table[25] marked ABCD. To begin, he gives a postulate:

> **Postulate 1**: I suppose the square table or plane ABCD to be so made and leveled, that if either of the balls *o* or W be thrown upon it, there shall be the same probability that it rests upon any one equal part of the plane as another, and that it must necessarily rest somewhere upon it.

First, one ball, *W*, is rolled onto the table, ABCD, with unit width, and the position of the ball, line *os*, is observed. *W* corresponds to the unknown parameter *T* (see Figure 3.2). Next, another ball, O, is rolled *n* times onto the same table, *p* of which end up in a position to the left of the original ball and *q* to the right ($p + q = n$). The number of times the ball O is to the right of the line *os* is marked *X*.

Next Bayes proves a lemma:

> **Lemma 1**: "The probability that the point *o* will fall between any two points in the line AB is the ratio of the distance between the two points to the whole line AB."

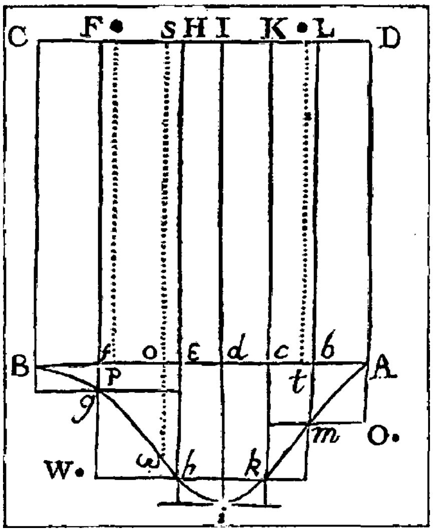


Figure 3.2:
Bayes' table: the square ABCD and two balls W and O. The ball W is rolled once and stops on line os. The ball O is rolled n times and the number of times it is in osDA is counted. (Bayes 1764 or Stigler 1986).

Bayes assumed that θ has uniform distribution on the unit interval and once θ is determined *X* has a binomial conditional distribution with θ representing success (or failure) on a single trial. A consequence of this thought experiment is that the distribution of θ can be considered as continuous. In a Bernoulli trial, θ had had a discrete distribution. Stigler (1986) noted that the continuous nature of θ provided a new symmetric character for the problem which opened a new way of solving it.

Geometrically, the position where the ball *W* comes to rest determines a line *os* parallel to sides BC and AD; θ is the ratio of Ao to AB. *X* is then the number of times O comes to rest in the rectangle osDA. Bayes called this event *M*.

Bayes wrote the next propositions using geometrical expressions related to the table in Figure 3.2. In modern notation, Proposition 8 can be written as follows:

[25] Some writers have later said that Bayes spoke or meant a billiard table (see e.g. Bellhouse 2004).

> **Proposition** 8: Before the ball *W* is thrown, the probability that the point O should fall between *f* and *b*, and withal that the event *M* should happen *p* times and fail *q* times in $n = p + q$ trials is

$$P(b < \theta < f \cap M_n = p) = \int_b^f \binom{n}{p} \theta^p (1-\theta)^q d\theta \qquad (3.7)$$

Corollary:

$$P(M_n = p) = \int_0^1 \binom{n}{p} \theta^p (1-\theta)^q d\theta = \frac{1}{n+1}$$

In the Essay, Bayes gave a ratio of surface areas in the figure instead of the formula. Finally, Proposition 9 addresses the inverse problem given in the beginning of the Essay.

> **Proposition** 9: "If before any thing is discovered concerning the place of the point O, it should appear that the event *M* had happened *p* times and failed *q* times in $n = p + q$ trials, and from hence I guess that the point *o* lies between two points such as *b* and *f*, and consequently that the probability of the event *M* in a single trail was somewhere between *b* and *f*, the probability I am in the right is [in modern notation]

$$P(b < \theta < f \mid M_n = p) = \frac{\int_b^f \theta^p (1-\theta)^q d\theta}{\int_0^1 \theta^p (1-\theta)^q d\theta} \qquad (3.8)$$

Also this expression cannot be found in the Essay. The proof was based completely on Newtonian geometric reasoning using both the corollary of Proposition 8, Proposition 8, and Proposition 5.

The problem (of inverse probability) was to determine $P(b < \theta < f \mid X)$ for any given *b*, *f*. Using modern notation, the problem which Bayes treated can be stated in the following manner: Given the number *p* of balls to the left, what is the probability of θ lying in the interval $(b, f) \supset [0, 1]$?

This can be rewritten as follows: Let *X* be a random variable, indicating the number of times an event happens in *N* trials, and θ is the probability that the event happens in a single trial. In the modern literature, the distribution for θ is referred as the "prior" distribution, as it represents the uncertainty about θ prior to making any observation *X*. The resulting conditional distribution for $(\theta \mid X = p)$ is called "posterior" distribution. Bayes considered only a uniform prior distribution for θ because there was no reason to believe any value more probable than others. Bayes said many times in the Essay that the principle is to be used only in cases where we have no grounds for choosing between the alternatives. Later, Laplace has called this principle the **principle of insufficient reason**, and later it has been ironically said to indicate equal distribution of ignorance (see Fisher 1930).

3.3.2 Discussion

Evidently, Bayes' Essay was the first explicit formal treatment of inverse probability. The Essay is difficult to read today even though it is quite obvious what to look for. Mostly the difficulty is due to the applied geometric mode of reasoning. Stigler (1982) says that "Bayes' essay is one of the more difficult works to read in the history of statistics". According to Stigler (1986), Bayes adopted the non-analytical geometric mode from Newton. Also Bellhouse (2004) argued that "Bayes was a strong Newtonian in his scientific outlook". However, the Newtonian method was uncommon among mathematicians of the time. For example, neither Bernoulli, de Moivre, nor Simpson embraced it.

As noted earlier, Bayes himself did not publish the Essay. His friend Richard Price, who found the manuscript among Bayes' papers, communicated it to the Royal Society in 1763. There have been speculations about the reasons why Bayes wrote about the topic but did not publish it (see Stigler 1986, Bellhouse 2004, or Fienberg 2006). One possibility is that he wanted to take up some of Simpson's ideas and find a more plausible explanation, but he was not satisfied with the result or he was to develop it further. Another theory is that Bayes was not satisfied with what he had achieved and hoped to be able to solve some more mathematical problems during his lifetime. There have also been speculation on what the real contribution of Bayes was and how much Price changed the text (see Fienberg 2006).

Bayes' Essay did not raise much interest in the beginning. According to Stigler (1982, 1986, and 1999), it remained nearly unnoticed for ten years. Hald (2007) contemplated why it it did not evoke any response from British mathematicians and natural scientists. It was only rarely referred to in the early 19th century writings on probability. Todhunter (1865) only briefly mentions Bayes, and in the early textbooks on statistics, such as Yule (1911) and Bowley (1901, 1910) Bayes was not mentioned at all. Also Westergaard (1932) dos not mention Bayes in his account of the history of statistics. A greater interest in it was shown only in the second quarter of the 20th century, and its current significance has been greatly borne only after World War II (see Fienberg, 2006). Today, Bayes' formula is one of the best known formulas in probability calculus, and it is the basis of a branch of statistical science. However, the origin of Bayes formula is rarely recognized and its original meaning has not been retained in the current applications.

A significant insight of Bayes was that the basic problems of induction or inductive inference incorporate inverse probability, or converse probability as he called it. He also understood that inverse inference couldn't be a straightforward or mechanical application of direct probability but required a different approach.

4 Laplace and estimating the population of France

4.1 Introduction

In the second half of the 18th century, France was rich in prominent mathematicians dealing with problems of probability theory or topics around probability calculus: such as Jean le Rond D'Alembert (1717–1783), Joseph Louis Lagrange (1736–1813), Marie Caritat de Condorcet (1743–1794), Antoine Lavoisier (1743–1794), Andrien-Marie Legrendre (1752–1833), and Pierre Simon Laplace (1749–1827)[26]. Their works were frequently published in the memoirs of the French Academy of Science. The French school, consolidated by the ideas of the Enlightenment, developed many central ideas and methods in mathematics and probability theory on which modern probability theory and statistical science rest. Karl Pearson gave a series of lectures on the history of statistics in the 17th and 18th centuries in France. Egon Pearson later edited and published them as a textbook (Pearson 1978). In these lectures, Karl Pearson gave a thorough account of the lives and works of all these French mathematicians and their collaboration.

Figure 4.1:
Portrait of Pierre Simon Laplace.

The French mathematicians also had connections with Daniel Bernoulli (1700–1782), Leonhard Euler (1707–1783)[27], and Heinrich Lambert (1728–1777), who mainly worked in Switzerland. Euler's contributions in mathematical analysis and series expansions paved the ground for Condorcet's and Laplace's developments in probability theory. Laplace's mathematical analysis leaned heavily on Euler[28]. It is evident that the French mathematicians were also aware of some of the works of the British mathematicians. De Moivre's and Simpson's works were especially well-

[26] Many of the French mathematicians were in a way or another involved in the French Revolution. The life of Lavoisier ended on a guillotine and Condorcet died in custody for unknown reason during the French revolution. (see Pearson 1978)

[27] **Leonhard Paul Euler** was a Swiss mathematician and physicist who spent a great part of his life in St. Petersburg and Berlin. Aside from Gauss, he is considered to be the preeminent mathematician of the 18th century, and one of the greatest of all time. He is also one of the most prolific mathematicians ever: he wrote some 25 monographs and about other 850 publications. He made important discoveries in fields as diverse as infinitesimal calculus and graph theory. He also introduced much of the modern mathematical terminology and notation, particularly for mathematical analysis.

[28] Laplace gave three advices to his students how to learn mathematics: "Read Euler! Read Euler! Read Euler!"

Both Dale (1999) and Hald (1998) give comprehensive accounts of the mathematical contributions of the French 18th century mathematicians. Both authors also analyse and comment on the mathematical side of inverse probability and describe how the ideas were developed in the course of time. The French mathematicians' contributions were so influential that it is justified to say that the foundations of inverse probability (and hence statistical inference) were laid in France at the end of the 18th century. In particular, Laplace's and Condorcet's contributions were central[29]. Pearson (1978) and Hacking (1975) claim that they developed together central parts of the theory of inverse probability. However, Condorcet's writing style was very difficult and his papers never gained a wider audience. Dale (1999) gives a comprehensive account and analysis about Condorcet's papers, though. Laplace's contributions, on the other hand, gained a wide audience and they were path-breaking, not only for the development probability theory and for statistical methods in the 19th century, but also for statistical science in general (see Hald 1998 and Stigler 1986).

Unlike Bayes, Laplace developed an analytic approach, obviously inspired by Euler's mathematical innovations. Laplace's thought model was urn trial and mathematical analysis typically involved (Euler's) series expansions, solving large factorials with Stirling's formula[30], and omitting terms that were negligible in large samples. Laplace's mathematical developments and the contributions on probability theory, especially on inverse probability, are comprehensively and minutely explored and analysed by Stigler (1986), Dale (1999), and Hald (1998, 2007).

Within probability theory, Laplace dealt with a great variety of different topics, as well as gave examples on how probability could be applied in science and in social situations. For example, he developed a "test" for evaluation of reliability of a witness in court. In this context, it is not possible, even superficially, to touch all topics that Laplace dealt with in probability theory. Only those topics that are directly or indirectly related to survey research and inverse probability will be touched on, and in that the mathematics is explored only to an extent that illustrates the lines of thought that explain how Laplace ended up with his statistical methods. A more thorough analysis can be found from the extensive literature covering Laplace's contributions.

According to Pearson (1978), Laplace wrote some eighteen memoirs dealing with the theory of the probability. Pearson (ibid.) also maintained that the most significant were written in the years 1772–1783. Two of them, "*Mémoire sur la probabilités des causes par les èvènemens*"[31] (PCE) in 1774 (at the age of 25) and "*Mémoire sur les probalités*"[32] (MOP, written in 1778 but published in 1781) were so influential that they can be regarded as the first significant contributions

[29] Many other famous mathematicians lived during the same period, such as Gauss, whose works have remained in the history of statistical science but they did not contribute directly to statistical inference (see Hald 1998 and Stigler 1986).

[30] Strirling's formula for ln(n!):

$$\ln(n!) = \frac{1}{2}\ln(2\pi) + (n + \frac{1}{2})\ln(n) - n + \frac{1}{12}n^{-1} - \frac{1}{360}n^{-3} + \frac{1}{1260}n^{-5}\ldots$$

[31] "Memoir on the Probability of the Causes of Events" (PCE)

[32] "Memoir on Probabilities" (MOP)

on inverse probability. When referring to Laplace's contributions, it is customary to refer to Laplace's later publications, "*Essai Philosophique sur les Probabilités*" and "*Théorie Analytique des Probabilités*". They were both first published in 1812 (Laplace 1812a, 1812b), and several editions were published in the first half of the 19th century. Although these two (and few others) are the best-known contributions of Laplace, they were partly composed of the results already published in the earlier memoirs.

In the first of the two early memoirs (PCE), Laplace had already presented his idea of inverse probability. Compared to some of Laplace's other works, this memoir reads fairly clearly and effortlessly and, as Stigler (1986) points out, even after more than two centuries, it seems almost like a contemporary work. Stigler (ibid.) also claims that the influence of this piece of work was immense. It was from this memoir that the ideas, now called "Bayesian", first spread through the mathematical world. Hald (1998) also considers Laplace's 1774 memoir as one of the revolutionary papers in the history of statistical inference. The second of these two memoirs (MOP) deals with more topics than the first one and is more elaborate. In it, Laplace completed his application of probability calculus to the analyses of errors in observations, which was left unfinished in the previous memoir. He also introduced the principle of probabilistic inferences in a more explicit manner. For example, he introduced the principle of statistical hypothesis testing, which actually existed latently already in the earlier text (see Hald, ibid.).

4.2 Laplace's inverse probability[33]

In the beginning of MOP, Laplace identified three different types of probabilities:

> "In the analysis of chance, we intend to know the probability of composite events, following any law, of simple events of which the possibilities are given; these are able to be determined in these three ways: 1° *a priori*, when, by the like nature of the events, we see that they are possible in a given ratio; it is in the same way, in the game of *heads* and *tails*, if the piece that we cast into the air is homogeneous and if its two faces are entirely similar, we judge *heads* and *tails* equally possible; 2° *a posteriori*, by repeating a great number of times the experience which can bring about the event of which there is question, and by examining how many times it has happened; 3° finally, by the consideration of the grounds which can resolve for us to say on the existence of this event; if, for example, the respective skills of two players A and B are unknown, as we have no reason to suppose A more strong that B, we conclude from it that the probability of A to win a game is ½. The first of these ways gives the absolute probability of the events; the second makes it known very nearly as we will just see in the following, and the third gives only their possibility relative to the state of our knowledge.
>
> Each event being determined by virtue of the general laws of this universe, it is probable only relatively to us, and, for this reason, the distinction of its absolute possibility and of its relative possibility can seem imaginary; but we must observe that, among the circumstances which compete in the production of the events, there are some variables at each instant, such as the movement that the

[33] Laplace did not use the term *inverse probability*. It was first used by the English scientist, August de Morgan in the 1830s (see Dale 1999, p. 4).

> hand imprints on the dice, and it is the reunion of these circumstances which we name: it is of others which are constant, *chance* such as the ability of the players, the inclination of the dice to fall on one of their faces rather than on the others, etc.; these form the *absolute possibility* of the events, and their knowledge more or less extensive forms their *relative possibility*; alone, they do not suffice to produce them; it is more necessary that they be joined to the variable circumstances of which I speak; they serve thus only to augment the probability of the events, without determining necessarily their existence." (Laplace 1778)

Following the contemporary classification of probabilities (see, e.g., Weatherford 1982), Laplace's first definitions correspond to the definition of classical probability, and the second to the definition of frequentist probability. They are both regarded as definitions of objective probability. The third definition of Laplace's corresponds to the modern meaning of subjective probability. Unlike in the modern statistical texts, for Laplace, *a priori* probability meant objective probability. In modern writings, *a priori* probability is obtained usually by subjective judgements. In this chapter, there is a danger of mixing up the modern definitions of probability with those of Laplace's. However, it will be indicated in which meaning of the concept of probability is used. In all direct citations from Laplace's text, the concept naturally denotes the same as he had intended.

4.2.1 The Principle of Inverse Probability

Laplace introduced the Principle of Inverse Probability in his *Memoir on the Probability of the Causes of Events* to the Royal Academy of Sciences (Laplace 1774)[34]. Laplace's Principle and Bayes' method are similar, and some writers, e.g., Pearson (1920) and Fisher (1930), claim that Laplace had copied the Principle from Bayes. However, according to Stigler (1978), Bayes' Essay was ignored until after 1780 and played no important role in the scientific debate until the 20th century. Todhunter (1865) is the first thorough account on the history of probability. There he says:

> "This memoir [Laplace 1774] is remarkable in the history of the subject, as being the first which distinctly enunciated the principle of estimating the probabilities of the causes by which an observed event may be produced." (Todhunter 1865)

Todhunter considers this memoir as the first contribution on inverse probability, and not that of Bayes'.

In the introduction to the memoir, Laplace says:

> "I propose to determine the probability of the causes of events, a question which has not received due consideration before, but which deserves even more to be studied, for it is principally from this point of view that the science of chances can be useful in civil life." (Laplace 1774)

An obvious conclusion from this citation is that Laplace was not aware of Bayes' Essay. Stigler (1978) claims that there are many reasons why it is reason-

[34] Stephan Stigler has translated the memoir and it was published in *Statistical Science* (Stigler 1986b).

ably certain that Laplace was unaware of Bayes' earlier work. Hald (1998) held that Bayes' and Laplace's theories are conceptually and mathematically so different that they cannot be related.

First Laplace defines the difference between direct and indirect (or inverse) probability. An example of direct probability is an urn that is known to contain only white and black tickets in a given ratio, and one seeks the probability that a ticket drawn by chance will be white. Laplace says that in this case, the event is uncertain but the cause on which the probability of occurrence depends is known. He continues by defining the *Problem* (of inverse probability) for which he later gives solutions in different formats:

> "An urn is supposed to contain a given number of white and black tickets in an unknown ratio; if one draws a ticket and finds it white, determine the probability that the ratio of white and black tickets is that of p to q. The event is known and the cause is unknown." (Laplace 1774)

In order to solve the Problem, he defined the *Principle* (of inverse probability)[35]:

> "If an event can be produced by a number n of different causes, the probabilities of these causes given the event are to each other as the probabilities of the event given the causes, and the probability of the existence of each of these is equal to the probability of the event given that cause, divided by the sum of all the probabilities of the event given each of these causes." (Laplace, ibid.)

In modern mathematical form[36], Laplace's Principle can be written as follows: If the "event" is denoted by E and $a_1, a_2, \ldots, a_n$ the n potential causes, then

$$\frac{P(a_i \mid E)}{P(a_j \mid E)} = \frac{P(E \mid a_i)}{P(E \mid a_j)}$$

and

$$P(a_i \mid E) = \frac{P(E \mid a_i)}{\sum_j P(E \mid a_j)} \qquad (4.1)$$

It is easy to notice that Laplace's Principle has the same form as the Bayes' formula with equal *a priori* probabilities (see formula 3.2). Hald (1998) argued that Laplace did not know of Bayes' Essay, and did not realize that his Principle could be derived as a conditional probability under the assumption that the causes are uniformly distributed. Only in the second edition of *Théorie Analytique des Probabilités* (published in 1814), did Laplace prove the general version of his Principle, which is the same as Bayes' theorem (see Hald, ibid.).

In effect, the formula (4.1) says that the probability of a_i being the cause of the observed event E is proportional to the direct probability of the event being

[35] Later, the *Principle* refers to this definition.

[36] Laplace never used the discrete form. All his mathematical treatments were composed of continuous functions. The discrete forms of the formula appeared later.

the cause, that is $P(a_i \mid E) \propto P(E \mid a_i)$. Laplace applied this idea in solving his problems where the observations, say $x_1, \dots , x_n$, are obtained for a given value of a continuously varying parameter θ, i.e.

$$P(\theta \mid x_1,\dots,x_n) \propto P(x_1,\dots,x_n \mid \theta) \tag{4.2}$$

Hald (ibid.) concluded that the intuitive background of the Principle may have been the same reasoning that led Lambert (1760) to the Maximum Likelihood principle: If the probability of the observed event for a given cause is large relative to the other probabilities, then it is more likely that the event has been produced by this cause than by the other causes.

To illustrate an application of the Principle, Laplace gave an example: There are two urns, *A* and *B*. The first contains *p* white tickets and *q* black tickets, and the second contains *p'* white tickets and *q'* black tickets. One draws *f* white and *h* black tickets from either of these urns, but not knowing which of the urns. It is required to determine what is the probability that the urn where the tickets were drawn was *A* or *B*?

Laplace gave the following solution using the Principle: Assuming that the urn was *A*, the probability of getting *f* white and *h* black tickets from it is

$$K = P(f,h \mid A) = \left. \frac{(f+h)!(p+q-f-h)!}{f!h!(p-f)!(q-h)!} \right/ \frac{(p+q)!}{p!q!} \tag{4.3}$$

The probability that the urn was *B*, i.e., $K' = P(f, h|B)$, can be obtained in a similar way replacing *p* and *q* by *p'* and *q'*, respectively. Applying the Principle, Laplace concluded that the probability that the urn was *A* is $P(A) = K / (K + K')$ and the probability that it was *B* is $P(B) = K' / (K + K')$. (See Laplace 1774)

Hald (1998) argues that this example, in modern terminology, may be regarded as an example of testing a simple hypothesis against a simple alternative if "causes" are replaced by "hypotheses".

The same idea by which Laplace solves the previous problem, i.e., the Principle, he applied in many different instances, also in the planning and the realization of the survey to estimate the population of France.

4.2.2 The Principle of Insufficient Reason

A priori probability was a significant element in both Bayes' and Laplace's inverse probability. There is little evidence that its use had been notably challenged before R.A. Fisher in the first quarter of the 20th century. Some critical writings concerning *a priori* probabilities were published during the 19th century, but they did not seem to have greater influence. One reason for the use of *a priori* probability obviously was the implicit idea of a superpopulation[37], which was involved in all inferential developments. Obviously, this was a consequence of the Newtonian worldview, which assumed that observations of the real world were realizations of an unknown cause system, which governed the world in the background. The parameters of the superpopulation ("states of the nature") were considered as unknown and therefore random variables, which had a probability distribution. This distribution was usually unknown and therefore, according to the *Principle of Indifference,* the rectangular prior distribution was necessary.

The principle of indifference: if on the background information B the hypotheses $(H_1, H_2, \dots H_N)$ are mutually exclusive and exhaustive and B does not favor any one of them over any other, then

$$P(H_i \mid B) = \frac{1}{N}, 1 \leq i \leq N$$

The principle of indifference or the *principle of insufficient reason* is often dedicated to Laplace. Stigler (1986) claims that the application of the principle of indifference was not a metaphysical assumption concerning the unknown structure of the world (equally likely causes). Rather, it was an implicit assumption that for ease of analysis, the problem had been specified in such a manner that this principle of insufficient reason was reasonable. Laplace writes:

> "When we have nothing given *a priori* on the possibility of an event, it is necessary to assume all the possibilities, from zero to unity, equally probable; thus, observation can alone instruct us on the ratio of the births of boys and of girls, we must, to consider the thing only in itself and excepting of the events, to assume the law of possibility of the births of a boy or of a girl constant from zero to unity, and to start from this hypothesis into the different problems that we can propose on this object." (Laplace 1778)

Fienberg (2006) argues that Laplace's introduction of the notion of "indifference" as an argument in specifying a prior distribution was the first in a long line of efforts "to discover the statistical holy grail": prior distributions reflecting ignorance.

In the 1774 memoir, Laplace writes:

> "...we assume that the coin which was tossed in the air had no tendency to favor either heads or tails. Now, this supposition is only mathematically admissible because physically there must be an inequality. But as the [players] are ignorant ... of which side has greater tendency, we can believe that this uncertainty neither increases nor decreases the advantage. We shall see, however, that nothing is less founded than this supposition, that it follows that the science of chances must be used with care, and must be modified when we pass from mathematical case to the physical." (Laplace 1774)

[37] The concept of superpopulation was not used nor even recognized by Laplace or any of his contemporaries or followers. Only R.A. Fisher introduced it (Fisher 1922).

4.2.3 Applications of the Principle in statistical inference[38]

For analyzing the causes of events, Laplace continues by applying his Principle to solve three different problems.

The *first problem* was defined in following way:

> "If an urn contains an infinity of white and black tickets in an unknown ratio, and we draw $p + q$ tickets from it, of which p are white and q are black, then we require the probability that when we draw a new ticket from the urn, it will be white."

Laplace assumed that the unknown ratio, x, of white tickets to all tickets in the urn has a continuous range of probabilities, $0 \leq x \leq 1$, all values equally possible. The probability of drawing p white tickets and q black tickets from the urn in a single drawing is $x^p(1-x)^q$. It should be noted that Laplace assumed the urn to be infinite, and therefore subsequent drawings were independent. Laplace concluded by applying the Principle[39] that the probability that x is the true ratio is:

$$P(x \mid p,q) = \frac{x^p(1-x)^q dx}{\int_0^1 x^p(1-x)^q dx} \tag{4.4}$$

The right-hand side of the formula is the form that Laplace used. Actually, it gives the probability that x falls in the range $[x, x + dx]$.

Assuming that x is the true ratio of white tickets to all tickets, the probability of drawing a white ticket from the urn is x. The probability of drawing a white ticket from the urn with true ratio x is obtained by multiplying (4.4) "by the probability of the supposition" (probability of drawing a white ticket from an urn with true ratio x × probability that x is the true ratio):

$$x \times P(x \mid p,q) = \frac{x^{p+1}(1-x)^q dx}{\int_0^1 x^p(1-x)^q dx} \tag{4.5}$$

And the total probability of drawing a white ticket from the urn, E, Laplace shows to be

$$P(E) = \frac{\int_0^1 x^{p+1}(1-x)^q dx}{\int_0^1 x^p(1-x)^q dx} \tag{4.6}$$

The expression for probability of E, i.e., the ticket in a new drawing being white, reduces after repeated integration by parts to

$$P(E) = \frac{p+1}{p+q+2} \tag{4.7}$$

[38] In this chapter, Laplace's original style of writing about mathematical topics is used, to demonstrate his reasoning and how he derived formulas. In general, his style was complicated and difficult to read for a modern reader.

[39] Only continuous distributions were analyzed in all 18th century writings on probability.

This result was later called Laplace's *Rule of Succession.* Using his Principle and

> "Euler's series", Laplace continues to show that if $m + n$ new tickets are drawn from the urn, the probability of getting m white tickets and n black tickets is
>
> $$P(E) = \frac{p^m q^n}{(p+q)^{m+n}} \tag{4.8}$$

Due to approximations, Laplace expected this to hold "without fearing any appreciable error" when p and q are very large, and m and n very small in comparison to p and q. However, he points out immediately that this approximation is inadequate for larger values of n and m. If $m = p$ and $n = q$, the probability should be approximated by

$$P(E) = \sqrt{1/2}\,\frac{p^m q^n}{(p+q)^{m+n}} \tag{4.9}$$

Laplace concluded that the solution to this problem provided a direct method to determine the probability of future events after those that have already occurred. This principle later came to be called the *Principle of Learning from Experience.* Laplace maintains that it is a broad subject and therefore gives only a "rather singular proof" of the following theorem:

> One can suppose that the numbers p and q are so large that it becomes as close to certainty as one wishes that the ratio of the number of white tickets to the total number of tickets contained in the urn is included between the two limits $p/(p+q) - w$ and $p/(p+q) + w$, one can suppose w to be less than any given quantity.

Using the results of the preceding examples, Laplace concludes that the probability of ratio x being between the given limits is

$$P\big(p/(p+q) - w \le x \le p/(p+q) + w\big) = \frac{\int x^p (1-x)^q dx}{\int_0^1 x^p (1-x)^q dx} \tag{4.10}$$

if the integral in the numerator is taken from $p/(p+q) - w$ to $p/(p+q) + w$. Marking the ratio $x = p/(p+q) + z$, Laplace shows, using approximations, that for infinitely large p and q, and "w infinitely less than $(p+q)^{-1/3}$ and infinitely greater than $(p+q)^{-1/2}$", the probability (4.10) is approximately

$$\frac{(p+q)^{3/2}}{\sqrt{2\pi pq}} \int_0^w 2e^{-(p+q)^3 z^2/2pq} dz \tag{4.11}$$

Then using "M. Euler's integral calculus", Laplace shows that this integral is approximately 1 and concludes that

> "...neglecting infinitely small quantities, we can consider it certain that the ratio of the number of white tickets to the total number of tickets is between the limits $p/(p+q) - w$ and $p/(p+q) + w$, where w is equal to $1/\sqrt[n]{(p+q)}$ and n is greater than 2 and less than 3, a fortiori when n is greater than 3 and therefore w can be supposed smaller than any given quantity."

The *second problem* that Laplace treated in PCE dealt with the division on wins in a game of chance, which had to broken before an orderly end. It is not strictly related to the current topic of this thesis and is therefore passed over.

The *third* and obviously most famous *problem* of inverse probability in PCE that Laplace described was "to determine the mean that one should take among three given observation of the same phenomenon". The motivation to take up this problem originated from a writing of Daniel Bernoulli, who in a footnote stated that this was an important problem, which he had not solved. Lagrange had also touched on the problem, but had not solved it. Laplace's treatment of this topic was important because he was able to show how the new Principle of Inverse Probability could be applied to nontrivial practical problems.

Laplace phrased the problem as follows (see Figure 4.2):

> Given three observations *a, b, c* of a phenomenon along a time axis *AB*. The time interval between *a* and *b* is *p* seconds and between *b* and *c*, *q* seconds. We wish to find the point *V* on the line where we should fix the mean that we should take between the three observations. It is supposed to represent the "true time" of the phenomenon.

He assumed that any observation differing from *V* by a factor x would lead to a probability which could be represented by a curve $y = \varphi(x)$. He stated three conditions for the error curve $\varphi(x)$, which should help to determine its true form:

1. $\varphi(x)$ must be symmetrical about V, since errors occur in both directions equally likely;
2. $\varphi(x)$ must decrease asymptotically to ordinate as $|V - x|$ gets greater, because "the probability that the observations differs from truth by an infinite distance is evidently zero";
3. $\int\varphi(x)dx = 1$ since it is certain that the observation will fall on a point under the curve.

Laplace concluded that the probability that three observations deviate from point *V* by distances *Va*, *Vb*, and *Vc* is $\varphi(x)\cdot\varphi(p-x)\cdot\varphi(p+q-x)$. If it is assumed that the true instance is *V'* and that $V'a = x'$ then the probability would be $\varphi(x')\cdot\varphi(p-x')\cdot\varphi(p+q-x')$. Applying the "fundamental Principle", Laplace concluded that the probabilities that the true instance is at the point *V* or *V'* are related to each other as

$$\frac{\varphi(x)\cdot\varphi(p-x)\cdot\varphi(p+q-x)}{\varphi(x')\cdot\varphi(p-x')\cdot\varphi(p+q-x')} \tag{4.12}$$

Next, Laplace noted that in seeking the mean to be chosen, there are two things that may occur: it is equally probable that the true instant of the phenomenon falls before or after it. Laplace called this the mean of probability. The second is the instant that minimizes the sum of the "errors to be feared"[40] multiplied by their probabilities. Laplace called this the mean of error, or the astrological mean.

Laplace continues by a similar approach by which Simpson had derived his error distribution (see Chapter 3): Given the true instant V, the posterior probability of the three observations a, b, and c can be expressed as:

$$\varphi(x|p,q) = \cdot\varphi(x)\varphi(p-x)\varphi(p+q-x)$$

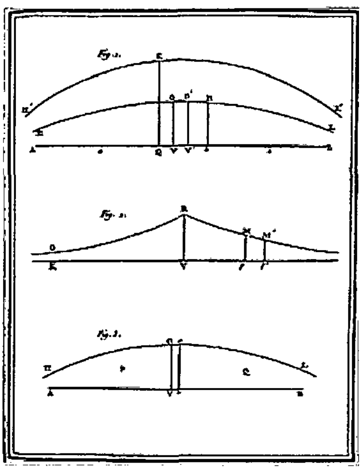


Figure 4.2:
Diagrams in the Laplace's memoire of 1774. The figure 2 shows the double exponential density.

The point, V, has disappeared and it is involved only in the first error term, x, which was an unobserved and random quantity. The error term, x, became the object of Laplace's investigation, the "*cause*" to be found from the observable "events" p and q. The distribution of "events" p and q, given the cause, x, $\varphi(p,q \mid x)$ was proportional to $\varphi(x,p,q)$. Therefore, by his principle, the distribution of x was

$$\varphi(x|p,q) \propto \varphi(x,p,q) \qquad (4.13)$$

Laplace had a small error in his formula, and therefore he ended up with an error distribution that had a wrong "shape", the double exponential distribution in Figure 4.2. Stigler (1986) analysed in detail Laplace's ideas and his further discoveries. In his later writings, Laplace, after a debate with Gauss, ended up with the correct form for the distribution. Later, this "law of error" has come to be known as Normal Distribution.

4.2.4 Plan to estimate the population of France

Laplace presented the idea for estimating the population of France in a memoir already in 1783[41], nearly 20 years before the survey actually was undertaken. In 20 years, Laplace had gained respect as a scientist and attained a high position in the French administration. Hald (1998) claims that the execution of the survey should also be seen in relation to the fact that the French government in 1800 had established a Central Statistical Office and prescribed a general enumera-

[40] Laplace's "error to be feared" ("*errour á craindre*") is conceptually close to the modern notion of standard error.

[41] "On the births, the marriages and the deaths at Paris, from 1771 to 1784; & at the whole extent of France, during the years 1781 & 1782" or "Sur les Naissances, les Mariages et les Morts" (Laplace 1783)

tion of the population in 1801[42]. It took more than two years before the census returns were received and processed, and the resulting figure of 27,349,003 inhabitants was considered unreliable. The next census, which took place in 1806, gave 29,107,425 inhabitants.

Laplace's plan was based on the fact that during the last quarter of the 18th century in France, all births were registered in parishes and published. Laplace writes in the beginning of the memoir:

> "... The Academy is determined ... to insert each year into its Mémoires, the list of births, of marriages & of deaths in the whole extent of France. A respectable magistrate by his light & and his zeal for the public good, & who since longtime occupies himself with success on research on the population, has well wished to produce to himself all the information which it was able to desire on this matter; it is to him that we are indebted of the following lists." (Laplace 1783)

According to Bru (1988), the "respectable magistrate" who had placed some of his tables at the disposal of Laplace, was a French demographer, Intendant Francoise de la Michodiére[43]. Bru (ibid.) gives a detailed account on Michodiére's contributions and demographic analyses in France.

Laplace's plan was to take a sample of the departments (small administrative districts), count the total population in the sampled departments for a single day, and then estimate the population of the whole country, using that information combined with information on registered births in France. Based on earlier demographic studies (e.g., Graunt's and Michodiére's), Laplace assumed that the ratio of population to births during a year was relatively stable. Another essential assumption was that the proportion of women of childbearing age in the population remained stable.

4.2.5 Determining the required sample size

Laplace selected 30 departments distributed over the area of France applying two criteria. First, all types of climates were represented. In this way, the effects of climate on the birth rate were compensated. Second, Laplace selected departments that had communes with mayors he thought were capable of providing accurate information. In modern terms, Laplace applied a two-stage cluster design. The method is close to modern cluster sampling except that the departments were selected with a purpose.

It is not known how Laplace ended up with exactly 30 departments, but the size of the sample was based on calculations on how large a sample was needed to obtain the required precision. In the memoir of 1783, Laplace showed how he estimated the needed size of the sample. His reasoning in the memoir seems surprisingly modern:

[42] Perrot and Woolf (1984) give an extensive and detailed account of the statistical activity in France 1789–1815.

[43] **Francoise de la Michodiére** (1720–1797) was a very productive and influential person in France. He became 'conseiller d'Etat' already at the age of 19. He was the first to use empirical data to argue for a positive association between wheat prices and excess mortality, which was later called the Michodiére's law.

> "The ratio of the population to the births ... can never be rigorously exact: by supposing in it even a rigorous precision, there would remain still on the population of France, the incertitude which is born of the action of the variable causes. The population of France, drawn from the annual births, is therefore only a probable result, & consequently susceptible to errors. It is to the analysis of chances to determine the probability of these errors, & to what point we must carry the denumeration, in order that it be very probable that they are contained within narrow limits. These researches depend on a new & yet little known theory, that of the probability of future events takes from observed events; they lead to some formulas of which the numerical calculation is impractical, because of the great numbers which we consider: but having given in this Volume & in the preceding, the principles necessary to resolve this kind of questions, & a general method to have in highly convergent series, the functions of great numbers; I have made application of it to the theory of the population deduced from births. The denumerations already made in France, & compared to the births, give very nearly 26 for the ratio of the population to the annual births; now if we take a mean among the births of the years 1781 & 1782, we have 973054½ for the number of annual births in the whole extent of the Realm, containing in it Corsica; by multiplying therefore this number by 26, the population of the whole of France, will be 25299417 inhabitants. Now I find by my analysis, that in order to have a probability of a thousand to one, of not being deceived by a half-million in this evaluation of the population of France, it would be necessary that the denumeration which has served to determine the factor of 26 had been of 771469 inhabitants. If we would take 26½ the ratio of the population to the births, the number of the inhabitants of France will be 25785944; & in order to have the same probability of not being deceived by a half-million on this result, the factor 26½ must be determined after a denumeration of 817219 inhabitants. It follows thence that if we wish to have for this object the precision which its importance requires, it is necessary to carry this denumeration to a million or twelve hundred thousand inhabitants." (Laplace 1783)

At the end of this citation, Laplace discusses how large a sample is needed to attain the required accuracy with a given probability. The end of the citation reads almost like contemporary text. It is noteworthy that Laplace realized that an estimate of the accuracy of the estimate was needed and that the accuracy depended on the size of the sample. He does not explain which analysis led to this conclusion or whether it was based on intuition. Twenty years later, Laplace published the Central Limit Theorem, which might have alluded to this.

Laplace continues explaining how he applied his Principle for inverse probability to calculate the size of the sample:

> "We consider an urn which contains an infinity of white & black tickets in an unknown ratio, & and we suppose that in a first drawing we have extracted p white tickets & and q black tickets; we suppose next that in a second drawing we have extracted q' black tickets, but we are ignorant of the number of white tickets brought forth in this drawing; the mean which naturally presents itself in order to know this number in an approximate manner, is to suppose it with q' in the ratio of p to q, that gives pq'/q for this number. We determine presently the probability that the true unknown number will be contained in the limits
>
> $$\frac{pq'}{q}(1-\omega),\frac{pq'}{q}(1+\omega)$$
>
> or that which returns to the same, that the error of the result $\frac{pq'}{q}$ will not surpass $\frac{pq'\omega}{q}$." (Laplace 1783)

The ratio estimate of population is the number of white tickets, that is, $p'=pq'/q$. Karl Pearson (1928) pointed out that his model was unsatisfactory in several respects (in view of modern understanding): the births are not regarded as part of the population, and the sample is not considered as part of the finite population. Nevertheless, Laplace creates some novel and useful ideas in statistical science, even though they might not be exactly correct.

In order to find out how large a sample would be needed, it was necessary to derive the (sampling) distribution of p'. This Laplace solved using his Principle for inverse probability in a manner that does not open readily (see Dale 1999, p. 218). If the unknown ratio of white tickets to the total number of ticket is x, and the unknown number of white tickets in the second drawing is p', the probability is

$$P(p',q'|x)=\frac{(p'+q')!}{p'!q'!}x^{p'}(1-x)^{q'} \tag{4.14}$$

Laplace assumed that p' may obtain all values from p' = 0 to p' = ∞ and "these values are more or less probable, according as they render the second drawing more or less probable."

Since $P(q' \mid x)=\sum_{p'=0}^{\infty}\frac{(p'+q')!}{p'!q'!}x^{p'}(1-x)^{q'}=\frac{1}{1-x}$ it follows that the probability is:

$$P(p' \mid q',x)=\frac{P(p',q' \mid x)}{P(q' \mid x)}=\frac{\frac{(p'+q')!}{p'!q'!}x^{p'}(1-x)^{q'}}{\sum_{p'=0}^{\infty}\frac{(p'+q')!}{p'!q'!}x^{p'}(1-x)^{q'}}$$
$$=\frac{(p'+q')!}{p'!q'!}x^{p'}(1-x)^{q'+1}, p'=0,1,2,... \tag{4.15}$$

Laplace assumed that all values of the ratio are equally probable in the range from $x = 0$ to $x = 1$, i.e., that the *a priori* distribution of x is uniform.

The formula (4.4) gives the probability of x, given p and q. By multiplying the probabilities of p' and x, Laplace obtained the "entire probability" of p' as

$$P(p' \mid p,q,q')=\frac{(p'+q')!}{p'!q'!}\frac{\int_{x=0}^{1}x^{p+p'}(1-x)^{q+q'+1}dx}{\int_{x=0}^{1}x^{p}(1-x)^{q}dx} \tag{4.16}$$

Both Hald (1998) and Stigler (1986) maintain that at the time of this memoir, Laplace had not found the asymptotic expansion solutions, which were so typical in his later works. Therefore, he had to attack the problem in a different manner.

To find probability $P(0 \le p' \le s)$, one needs to sum the terms in the numerator, depending on p'. For this purpose, Laplace (using results from another memoir), assuming that q' and s are "very large numbers", shows that:

$$\sum_{p'=0}^{s} \frac{(p'+q')!}{p'!} x^{p'} = \frac{1}{(1-x)^{q'+1}} \frac{\int_{x'=x}^{1} x'^{s}(1-x')^{q'}dx'}{\int_{x'=0}^{1} x'^{s}(1-x')^{q'}dx'}$$

Hence, the probability is

$$P(0 \leq p' \leq s \mid p,q,q') = \frac{\int_{x=0}^{1}\int_{x'=x}^{1} x^{p}(i-x)^{q}dx \cdot x'^{s}(1-x')^{q'}dx'}{\int_{x=0}^{1}\int_{x'=0}^{1} x^{p}(i-x)^{q}dx \cdot x'^{s}(1-x')^{q'}dx'} \qquad (4.17)$$

Based again on earlier results, Laplace concludes that if s is less than and hardly different from pq'/q, then (4.17) can be approximated by an integral, which in modern probability theory is called the normal probability:

$$P(0 \leq p' \leq s \mid p,q,q') = \frac{1}{\sqrt{\pi}}\int_{T}^{\infty} e^{-t^2} \qquad (4.18)$$

where $T^2 = \dfrac{\left(\dfrac{p}{p+q} - \dfrac{s}{s+q'}\right)^2 \cdot (p+q)^3(s+q')^3}{2sq'(p+q)^3 + 2pq(s+q')^3}$

The right hand side of formula 4.18 is approximately the normal distribution with zero mean and unit variance, i.e. $N(0,1)$. The approximation holds 'near the most probable values' of p'. If s is greater than pq'/q but close to it, the probability becomes

$$P(0 \leq p' \leq s \mid p,q,q') = 1 - \frac{1}{\sqrt{\pi}}\int_{T}^{\infty} e^{-t^2} \qquad (4.19)$$

Hence, it follows that

$$P(s \leq \frac{pq'}{p} \leq s') = 1 - \frac{1}{\sqrt{\pi}}\int_{T}^{\infty} e^{-t^2} - \frac{1}{\sqrt{\pi}}\int_{T'}^{\infty} e^{-t^2}$$

where T' is defined as T replacing s with s'. If one sets

$$s = \frac{pq'}{q}(1-\omega), s' = \frac{pq'}{q}(1+\omega)$$

and disregards terms of order ω^3 and two values of T^2 and T'^2 and takes $V^2 = \dfrac{pqq'\omega^2}{2(p+q)(q+q')}$ then

$$P(\frac{pq'}{q}(1-\omega) \leq \frac{pq'}{p} \leq \frac{pq'}{q}(1+\omega)) = 1 - \frac{2}{\sqrt{\pi}}\int_V^\infty e^{-t^2} \qquad (4.20)$$

Laplace continues considering how p (the "size of sample") should be determined to obtain a large probability that the error in p' (the predicted population size) is small. For this purpose, he denotes that proportion of white tickets to black by $i = p/q$ and the accepted error by $a = \frac{pq'}{q}\omega$ and hence $\omega = \frac{a}{iq'}$. The previous expression of V^2 yields $p = \frac{2i^2(i+1)q'^2V^2}{a^2 - 2i(i+1)q'V^2}$. The value of a depends on the limits between which the error of the estimate is supposed to be.

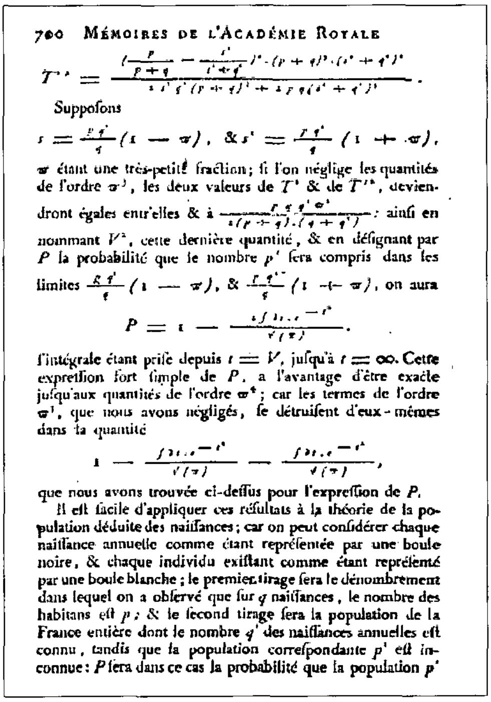
Laplace gives an example by first letting $a = 500000$. The value of q' was the number of annual births in France, which was known to be $q' = 973054.5$ (the decimal expression resulted from a three-year average). The value of V depends on the probability, P, that the population would be enclosed within the limits

$$P\left(\frac{pq'}{q} - a \leq p' \leq \frac{pq'}{q} + a\right)$$

Laplace set the probability to "a thousand to one", that is P =1000/1001;

$$\frac{2\int e^{-t^2}dt}{\sqrt{\pi}} = \frac{1}{1001}, or$$

$$\int e^{-t^2}dt = \frac{\sqrt{\pi}}{2002}$$



700 MÉMOIRES DE L'ACADÉMIE ROYALE

$$T'^2 = \frac{(\frac{p}{p+q} - \frac{s'}{s'+q'})^2 \cdot (p+q)^3 \cdot (s'+q')^3}{2s'q'(p+q)^3 + 2pq(s'+q')^3}.$$

Suppoſons

$$s = \frac{pq'}{q}(1 - \varpi), \ \& s' = \frac{pq'}{q}(1 + \varpi),$$

ϖ étant une très-petite fraction; ſi l'on néglige les quantités de l'ordre ϖ^3, les deux valeurs de T^2 & de T'^2, deviendront égales entr'elles & à $\frac{pqq'\varpi^2}{2(p+q)\cdot(q+q')}$: ainſi en nommant V^2, cette dernière quantité, & en déſignant par P la probabilité que le nombre p' ſera compris dans les limites $\frac{pq'}{q}(1 - \varpi)$, & $\frac{pq'}{q}(1 + \varpi)$, on aura

$$P = 1 - \frac{2\int dt \cdot c^{-t^2}}{\sqrt{(\pi)}}.$$

l'intégrale étant priſe depuis $t = V$, juſqu'à $t = \infty$. Cette expreſſion fort ſimple de P, a l'avantage d'être exacte juſqu'aux quantités de l'ordre ϖ^4; car les termes de l'ordre ϖ^3, que nous avons négligés, ſe détruiſent d'eux-mêmes dans la quantité

$$1 - \frac{\int dt \cdot c^{-t^2}}{\sqrt{(\pi)}} - \frac{\int dt \cdot c^{-t^2}}{\sqrt{(\pi)}},$$

que nous avons trouvée ci-deſſus pour l'expreſſion de P.

Il eſt facile d'appliquer ces réſultats à la théorie de la population déduite des naiſſances; car on peut conſidérer chaque naiſſance annuelle comme étant repréſentée par une boule noire, & chaque individu exiſtant comme étant repréſenté par une boule blanche; le premier tirage ſera le dénombrement dans lequel on a obſervé que ſur q naiſſances, le nombre des habitans eſt p; & le ſecond tirage ſera la population de la France entière dont le nombre q' des naiſſances annuelles eſt connu, tandis que la population correſpondante p' eſt inconnue: P ſera dans ce cas la probabilité que la population p'

Figure 4.3:
Copy of a page in Laplace's 1783 memoir.

where the integral is taken from $t = V$ to $t = \infty$. Laplace concludes, "it is clear that this equation determines V", which is $V^2 = 5.415$. Hence the number i could be obtained from an enumeration, but the purpose of the memoir was to make a plan for the enumeration. Therefore, p and q were unknown. Based on the earlier enumerations, it was known "very nearly" that $i = 26$. Laplace does not publish or discuss la Michodière data leading to the value of i. He just takes it into use. However, Laplace also carries out similar calculations for $i = 25.5$ and $i = 26.5$, which give values $p = 727520$ and $p = 817219$. Obviously, there was some uncertainty in his mind concerning the value of i.

His conclusion was that in order to have a "probability of one thousand against one, of not being deceived by more than one half-million in the evaluation of the population of France", it is necessary that the "sample size", p, is 727,210 inhabitants (in the case where it is expected to give the first factor), 771,469 inhabitants (in the case of the second factor), and 817,219 inhabitants (if it is expected to lead to the third factor).

Finally, Laplace concluded that "if we wish to have for this object, the probability that its importance requires, it is necessary to carry to a million or twelve hundred thousand inhabitants, the denumeration p which must determine the factor i." Obviously, he had doubts concerning the assumed value for i and therefore ended up to suggest larger sample size than what his calculations showed.

4.3 Estimating the population size

Earlier works in actuarial mathematics by Halley, de Moivre, and Simpson had shown that the size of a population could be calculated from yearly number of births and the life table, assuming a stationary population (see Hald 1990). The problem was that populations are rarely stationary and life tables at that time were not representative. This may have been one reason why Laplace proposed to use a ratio estimator. Ratio estimation was not invented by Laplace. Graunt had estimated the population of London in 1662 by a similar method (see Chapter 2). However, Laplace extended the survey to the whole of France while Graunt investigated only London and in a more intuitive manner[44].

When the estimation took place, Laplace applied a slightly more elaborate method in estimation than he had proposed earlier in the derivation of the required sample size. Since the first paper, he had managed to solve some mathematical problems, and he had also developed new mathematical tools for probability calculus (see Hald 1998 and Stigler 1986). The inference model was still based on a Bernoulli trial, i.e., a box with white and black tickets.

A black ticket still represents a birth and each white ticket represents an individual living in the country. The first drawing represents the enumeration in which it is observed y births and the number of inhabitants is x[45]. The second drawing represents the population of the whole of France and the total number of births, X, is known while the corresponding population, Y, is unknown.

From the known total number of registered births, x, during the preceding year in the selected departments and that in the whole country, X, the ratio estimate of the population of France, Y, could be calculated as:

$$\hat{Y}_R = X\frac{y}{x}$$

[44] Graunt also planned to estimate the population of England but he never put his plan into action (see Hald 1998).

[45] In this chapter, Laplace's notations are replaced by a more modern style found, e.g., in Cochran's book on sampling (Cochran 1953).

The combined population of the sampled Departments as of September 22, 1802, was 2,037,615. As for births, Laplace totalled the sample births for the three-year period from September 22, 1799, to September 22, 1802, obtaining a value of 215,599, so that his sample x is 215599/3. By taking the average of the number of births, he hoped to eliminate random variation (or 'fluctuation', as it was called). Finally, in the numerical estimate of the sample ratio, y/x, was 28.352845 and then

> "supposing that the number of annual births in France is one million, which is nearly correct, we find, on multiplying by the preceding ratio (y/x), the population of France to be 28 352 845 persons."

Laplace assumed that an infinite urn consisted of white and black tickets representing a population of French citizens on a specified day. In modern terms, he regards the number of known births in the country, X, as a random variable from a sample of unknown size Y, the population of France.

White tickets represented registered births in the preceding year. The ratio p (proportion of white tickets) is unknown. He regarded the ratio x/y from the sample as a binomial estimate of p. Cochran (ibid.) says that the choice of the model presupposes that the birth rate p varies from department to department.

Laplace assumed that the unknown ratio of births to population, p, follows uniform prior dp ($0<p<1$). Cochran (1978) argued that obviously an essentially tighter prior had been justified, for example, $p < 0.2$. Hald (1998) pointed out that this had complicated significantly Laplace's theoretical analysis. Given the binomial sample data from the communes (x successes out of y trials), the posterior distribution of p is then

$$\frac{p^x(1-p)^{y-x}}{\int_0^1 p^x(1-p)^{y-x}dp} \tag{4.21}$$

which is an application of Laplace's Principle (giving the relative likelihoods of various values of unknown ratio p).

In the second drawing, only X (the total number of births) is supposed to be known. The problem was to find the distribution Y for given values of X, x, and y. Cochran claims that Laplace assumed (implicitly) that he had a second independent binomial sample,

For a given value of X, the probability is

$$p(X \mid Y,p) = \binom{Y}{X} p^X(1-p)^{Y-X} \tag{4.22}$$

and combining it with the posterior distribution of p from the first sample gives

$$p(X \mid Y,x,y) = \frac{\binom{Y}{X}\int_{p=0}^1 p^{X+x}(1-p)^{(Y-X)+(y-x)}dp}{\int_{p=0}^1 p^x(1-p)^{(y-x)}dp} \tag{4.23}$$

According to the Principle of Inverse Probability, $p(Y \mid X,x,y) \propto p(X \mid Y,x,y)$

In order to analyse the error in the ratio estimate, Laplace marked $Y = \hat{Y}_R + z$ and focuses on the distribution of z. The probability of X successes in $(Xy/x) + z$ trials is

$$\frac{\left(\dfrac{Xy}{x} + z\right)!}{X!\left[\dfrac{X(y-x)}{x} + z\right]!} p^x(1-p)^{[X(y-x)/x]+z} \qquad (4.24)$$

To find an approximate form of the distribution function $f(z)$ when x, X, and y are all large, Laplace first multiplies (4.21) by (4.24), implicitly assuming that the two samples are independent. Drawing tickets two times from an infinite urn can be regarded as two independent trials, but that does not hold with a finite population. The error is insignificant, however (see Cochran, ibid.).

Then by applying Stirling's approximation to $n!$ and expanding to a Taylor series up to terms z^2, he ends up with the approximation of the frequency distribution of z. (For details, see Cochran 1978, or Hald 1998.)

$$f(z) \cong \exp\left\{-\frac{1}{2}\frac{x^3}{X(X+x)y(y-x)}\left[z^2 + z - \frac{2X(y-x)}{x^2}z\right]\right\} \qquad (4.25)$$

Laplace thus showed that in large samples, the distribution of the error z in the ratio estimate $\hat{Y}_R$ was approximately normal, with a bias whose leading term is $X(y\text{-}x)/x^2$ if x/X is negligible, and a variance

$$V(z) = X(X+x)y(y-x) \,/\, x^3 \qquad (4.26)$$

Laplace calculated that the "standard error", given the data, was 107.550 persons, and he concluded that it makes "the odds about 300,000 to 1 against an error of more than half a million".

This conclusion is basically the same as the modern expression concerning the accuracy of a sample estimate, bearing in mind the different conceptions and definitions of probability. In another context, Laplace also defined an "error to dread", which in modern terms can be described as a probabilistic interval estimate, which is a kind of confidence interval.

Thatcher (1964) compared so-called binomial prediction, based on Laplace's theory, and the theory of confidence limits. The interpretations of the inference models are different, but the comparison is interesting anyway. He found that the confidence limits lie outside the Laplacian limits, but the difference between them is no larger than the effect of one extra observation.

4.4 Laplace's and Brewer's ratio estimators

Ratio estimator has been included in all textbooks on survey sampling. However, Cochran (1978) claimed that an infinite superpopulation to study the properties of estimators had not been applied in the same manner as Laplace did since Brewer published his paper in 1963 (Brewer 1963).

Brewer assumed that the population values (y_i, x_i) are a random sample from a superpopulation in which

$$y_i = \beta x_i + \varepsilon_i$$

where ε_i and x_i are independent and $x_i > 0$. In arrays in which x_i is fixed, ε_i has mean 0 and variance cx_i. The values x_i $(i = 1, 2, \ldots, N)$ are known. In other words, the population total (the population of France), $Y = \beta X + \sum_{1}^{N} \varepsilon_i$, is assumed to be a random variable (*X* is the number of known births in the country, see Cochran 1978).

Applying Brewer's model to Laplace's problem, results conditional on the known value of *X* are obtained by writing $\hat{Y}_R = X/p_y$ and $Y = X/p_Y$, where p_y and p_Y are estimates of *p* obtained from binomial samples of sizes *y* and *Y*, respectively. Then

$$\hat{Y}_R - Y = X\left(\frac{1}{p_y} - \frac{1}{p_Y}\right) \cong X(p_y - p_Y)/p^2 \qquad (4.27)$$

Averaging over repeated selections of both drawings (France and a sample of communes) gives (see Cochran 1978)

$$E(\hat{Y}_R - Y) = EV(\hat{Y}_R) \cong \frac{X^2pq}{p^4}\left(\frac{1}{y} - \frac{1}{Y}\right) = \frac{X^2q}{yp^3}\frac{(Y-y)}{Y} \qquad (4.28)$$

This is nearly the same as that of Laplace's estimate if the binominal selection was replaced by the hypergeometric. Cochran (ibid.) pointed out that the difference is caused by Laplace's (implicit) assumption that France itself and the sample of departments were independent binomial samples from an infinite superpopulation, or from an infinite urn, with an unknown ratio of births to population. In Brewer's model, the sample of communes is a subsample drawn from France.

4.5 Laplace's other contributions to probability theory

Laplace's contributions in probability theory were diversified, and only part of them can be explored here. He had a great number of ideas but he was not able to finalise all of them, and some of them were in a hidden form. For example, Laplace's Principle involves a similar reasoning as the idea of maximum likelihood. Another example is his work on a topic that is currently known as **sufficiency**. Stigler (1973) notes that Laplace did similar investigations as Fisher did a century later. Stigler states that it was surprising how close Laplace came to discovering sufficiency in 1818.

4.5.1 Central limit theorem

From the standpoint of statistical inference, probably the most important invention of Laplace was the Central Limit Theorem (CLT), where "Central" should be understood as a synonym for fundamental[46]. Stigler (1986) and many others consider this theorem as Laplace's major result in probability theory. The CLT asserts that, under certain general conditions, the sum of a large number of independent variables is approximately normally distributed. In modern form, the theorem is:

If $\bar{x}$ is the mean of a sample of size n from a distribution having finite variance σ^2 and mean μ, then

$$\lim_{n\to\infty} P\left[\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \leq y\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-\frac{1}{2}u^2}\,du = \Theta(y)$$

where $\Theta(y)$= N(0,1), the normal distribution function.

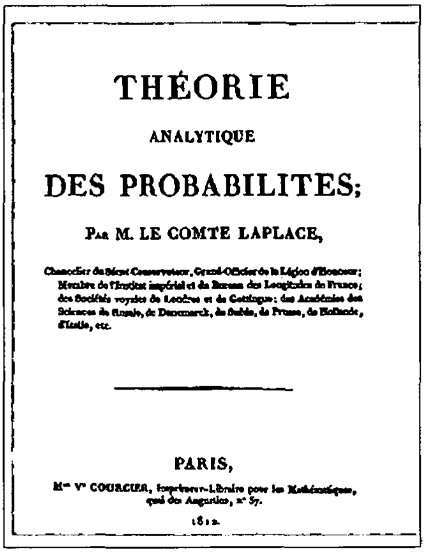
Before Laplace published the CLT, the distribution of the arithmetic mean had been studied for many error distributions, resulting in very complicated formulas, and therefore the need for approximations was obvious (see Hald 1998, especially Chapter 3). The only previous theorem, according to Hald (Ibid.), was due to de Moivre, who in 1730 proved that

$$\binom{2n}{n}\left(\frac{1}{2}\right)^{2n} \cong \frac{1}{\sqrt{\pi n}}$$

and in 1733, de Moivre had derived the normal approximation to the binomial distribution.

Laplace realized that a new mathematical technique was required. He obtained the necessary new method for approximating $P(s_n = s)$, $s_n = x_1 + x_2 + \ldots x_n$, ($x_i$ is the number of points at the i^{th} throw), by a combination of two of his main lines of mathematical methods: the theory of generating functions and the method of asymptotic expansions of an integral. The characteristic functions, or Fourier transforms, were an outgrowth of a technique Lagrange had employed a few years earlier (see Stigler 1986).

[46] The actual term "central limit theorem" (in German: "zentraler Grenzwertsatz") was first used by George Pólya in 1920 in the title of a paper. Pólya (1920), "Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem", Mathematische Zeitschrift 8: 171–181.



THÉORIE

ANALYTIQUE

DES PROBABILITES;

PAR M. LE COMTE LAPLACE,

Chancelier du Sénat Conservateur, Grand-Officier de la Légion d'Honneur; Membre de l'Institut impérial et du Bureau des Longitudes de France; des Sociétés royales de Londres et de Gottingue; des Académies des Sciences de Russie, de Danemarck, de Suède, de Prusse, de Hollande, d'Italie, etc.

PARIS,

Mme Ve COURCIER, Imprimeur-Libraire pour les Mathématiques, quai des Augustins, n° 57.

1812.

Figure 4.4:
Cover page of Théorie analytic des probabilites

The paper where the CLT was published was read to the Academy of France in 1810 (Laplace 1810). Laplace's work did not include the regularity conditions familiar in the modern form of CLT. They and the exceptional cases came later, and they were treated by several mathematicians e.g., Poisson, Cauchy, Liapounov, Lindeberg, etc. Hald (1998) wondered why Laplace, being so near the general solution to CLT, did not continue his analysis, and answers that Laplace took up this problem very late in his career. Hald (ibid.) suspects that an explanation may be the fact that from 1786 on, Laplace concentrated on one of his major works in astronomy *Mécanique Céleste,* which he completed in 1805.

The importance of CLT for all statistical inference is undeniable. Only after the CLT was invented could the development of partial investigations, or sample surveys, be justified. However, it took nearly a century before the applications started to appear.

Laplace also treated, and partly derived, the Law of Error (Normal Distribution) together with Gauss, although the origins of the distribution are older (see Stigler 1986 or Cramér 1970). In older textbooks on probability, the Normal Distribution is occasionally called the Gauss-Laplace distribution (see Cramér 1970). In Russian texts in the early 20th century, Normal Distribution was often called the Gauss-Laplace law.

4.5.2 Hypothesis testing

Laplace was also keen in applying his methods on new areas. For example, as an *application of tools he had developed using* binomial probability, he undertook an analysis on the sex ratio at birth. He had data from a twenty-six-year series in Paris. He found the total number of births to be y = 251527 for boys and z = 241945 for girls. If x represents the probability that a given birth is male, he calculated in a straightforward application the posterior probability:

$$P(x \leq ½ \,|\, y = 251527, z = 241945) = 1.1521 * 10^{-42}$$

He therefore regarded it as certain that the probability for a male birth is $p_x > ½$. This is clearly an example of a test of a hypothesis the null hypothesis being H_0: $p_x \leq ½$.

Below is a citation of another application that is also close to modern hypothesis testing.

> "We have observed that, in the interval of the 85 years elapsed from 1664 to 1757, there are born, in London, 737629 boys and 698958 girls, which gives around 19/18 for the ratio of births of boys to those of girls; this ratio being greater than the one of 105 to 101 which took place in Paris, and the number of births observed in London being very considerable, we would find for this city a greater probability that the births of boys are more possible than those of girls; but, when the probabilities differ likewise little from unity, they can be counted equal and confused with certitude.
>
> The preceding method gives a quite simple solution to an interesting problem, which it is perhaps very difficult to resolve by other methods: we have seen that the ratio of the births of boys to those of girls is sensibly greater in London than in Paris; this difference seems to indicate in London a greater facility for the birth of boys: the question is to determine how much this is probable.
>
> This value of P (the probability that the birth of a boy is less possible in London than in Paris = 1/410458) is a little too great; but, since taking in it only the first two terms of the series we would have a value too small, it is easy to conclude from it that the preceding can differ from the truth by the 1/142 part of its value, so that it is a strong approximation: there is therefore odds of more than four hundred thousand against one that the births of boys are more facile in London than in Paris. Thus we can regard as a very probable thing that it exists, in the first of these two cities, a cause more than in the second, which facilitates the births of boys, and which depends either on the climate or on the nourishment of the mothers." (Laplace 1783)

4.6 Laplace's influence on statistical science

Laplace has been regarded as a Newtonian scientist (see e.g. Stigler 1986). Newtonianism is the doctrine of following the principles and making use of the methods of the natural philosopher Isaac Newton. Newtonianism was an influential intellectual program during the 18th century Enlightenment. The followers of Newton tried to apply Newtonian principles to a wide variety of new fields. Other famous Newtonian scientists of that time were Leonhard Euler and David Hume.

Weatherford (1982) claims that the classical theory of probability reached its zenith in the work of Laplace and that Laplace solved more problems and developed more important mathematical tools, including statistical methods, than any of his predecessors. His contributions were so influential that they dominated statistical thinking nearly for a century. His mathematical treatment of the statistical problems also provided new tools for the development of the theory. Laplace's contributions can be regarded as the origin of statistical science. They opened a new era in the development of probability theory and its application to empirical sciences. In the context of this thesis, Laplace's derivation and application of the inverse probability are of central importance.

Hald (1998) concludes that Laplace is a pioneer in sample surveys, and that his theory is essentially correct for simple random sampling, although his model actually did not correspond to this mode of sampling. Laplace conducted the first scientifically ambitious partial investigation and thus created a method of statistical inference. Laplace's method can be partly seen as a response to Hume's critique of the inductive method because it aims at providing tools for inductive reasoning.

It is surprising how little attention Laplace's work has received in textbooks on statistical science. It seems that his contributions have nearly completely fallen into oblivion. For example, Hansen and Hurwitz wrote in their account of the historical basis for modern sampling theory:

> "The theory for independent random sampling of elements from population where the unit of sampling and the unit of analysis coincide was developed by Bernoulli more than 200 years ago. The theory that would measure the gains to be had from introducing stratification into sampling was indicated by Poisson a century later. Subsequently, Lexis systematized previous work and provided the theoretical basis for sampling clusters of elements. The adaptation of the work of Bernoulli and Poisson to sampling from finite populations was summarized by Bowley in 1926 approximately a century after the work of Poisson." (Hansen and Hurwitz 1943)

The authors mention Poisson, a disciple of Laplace who developed Laplace's ideas but not Laplace.

As late as in 1978, Cochran was also surprised by the discovery of the use of the ratio estimator by Laplace (Cochran 1978). It is a slightly peculiar nuance in the history of statistics that R.A. Fisher regarded Bayes' contribution as the first attempt to formalize inductive reasoning while not giving any credit to Laplace. However, he obviously knew Laplace's works well because he sharply attacked them several times[47].

There is little evidence available on how Laplace originally derived the general theory. He was only 25 when he wrote PCE. The greatest motivation for his work, at least in the beginning, may have been the problem of merging discrepant observations in astronomy (see Stigler 1986).

Laplace was primarily an astronomer. His interest in population statistics was apparently less motivated by social or political concerns than by the scientific aim of making evident that the social world can basically be approached by the same probabilistic methods as the physical world (see Fischer 2001). However, his work and his teaching had a far-reaching influence on the social sciences in the 19th century. For example, Quetellet's *Social Physics* was essentially based on Laplace's idea of social phenomena being analogous to natural ones. The basic ideas of Laplace influenced the 19th century mathematicians with the resulting expectancy that all random fluctuations in nature and in society could be treated correspondingly to a pattern of errors in observations. Fischer (2001) claims that this concept, together with Laplace's frequent approximations by normal distributions, paved the way for the latter "Quetelism".

[47] Fisher explained that the reason why he appreciated Bayes and did not appreciate Laplace was the fact that Bayes did not publish his results! (See Fisher 1936)

5 Laplace-Bayes paradigm for statistical inference

In developing his Principle of Inverse Probability, Laplace created a method that in modern terminology can be called statistical inference. In modern statistical science, inference is based on a different inference model than in Laplace's method, but the purpose is the same: to infer from sample to population.

Typical features of Laplace's method are:

1. The inference model that is based on Bernoulli trials introducing a binominal setup and binominal probabilities.
2. The universe (population in modern terms) is supposed to change constantly. Therefore, its parameters cannot be considered as constants but as having (a priori) probability distributions. In modern terms, the inference setup is close to the superpopulation approach in which the observable population is a sample of a superpopulation.
3. The inference setup is formed by using the inverse probability principle of Laplace.
4. The indifference principle (or the principle of insufficient reason) is an inherent part of inference, and it is usually considered to justify the use of the rectangular distribution for a priori probabilities.

The concept of *a priori* probability should not be confused with concepts like "credibility" or "degree of confirmation," or "strength of expectation," etc. as is often done in modern Bayesian theory. In Laplace's and Bayes' theory, *a priori* probability is an objective probability, but its value is not known and its value cannot be found experimentally.

In the writings on probability theory, Laplace's patterns of thought are prevalent throughout the 19th century (see e.g. Chang 1976). For example, Poisson's, Quetelet's and Lexis' contributions are based on these ideas, as well as the theory building in Russia (e.g., Tchuprov, Spława-Neyman, and Kovalevsky). There is indirect evidence that the texts of Laplace, Poison, and Quetelet were in common use both in European and Russian universities. Maybe the most assuring indication is what R.A. Fisher wrote in 1936:

> "...In the latter half of the nineteenth century the theory of inverse probability was rejected more decisively [than Boole] by Venn and by Chrystal, but so retentive is the tradition of mathematical teaching that I may myself say that I learned it at school as an integral part of the subject, and for some years saw no reason to question its validity." (Fisher 1936)

Another indirect evidence of the predominant role of Laplace's inference model is Karl Pearson's article from 1920, *The Fundamental Problem of Practical Statistics*. In this article, Pearson referred to "inverse probabilities" and concluded that in practical statistics, it takes the following form:

> "An 'event' has occurred p times out of $p + q = n$ trials, where we have no *a priori* knowledge of the frequency of the event in the total population of occurrences. What is the probability of its occurring r times in a further $r + s = m$ trials?" (Pearson 1920).[48]

Pearson continued that the problem had been considered in 1774 by Laplace "whose approximation by aid of Stirling's Theorem leads us directly to the normal curve". In this article, Pearson analyzed and modified "Laplace's investigation on a broader basis".

The inference model in Jerzy Neyman's early paper was based on drawing balls simultaneously from several urns (Spława-Neyman 1923, p. 467). His probabilistic setup was an elaborated version of the classical urn model (balls in urns have labels giving the values of the variable). This setup originated from Lexis and it was common in Russian writings before World War I. For example, Tchuprov applied a similar setup in his analysis of distributions (see Tchuprov 1918, 1923a, and 1923b). In addition, Kovalevsky applied the Bernoulli trial in deriving estimators for sample surveys. He started his paper on sampling by saying: "Suppose we have an urn containing white and black balls in an unknown ratio..." (Kovalevsky 1924).

Hald (1998) called this method the Bayes-Laplace model, and Stigler (1986) called it the Bayes-Laplace method, but it also fulfils the characteristics of a paradigm in the sense Kuhn described it. In the 19th century, Laplace's inference model had a central role; obviously Laplace's method was the dominant one in universities. At that time, it was a method that was taught in universities all over Europe as a self-evident and natural approach.

Up to the beginning of the 20th century, Laplace's and later Poisson's and Quetelet's textbooks were important study materials for new students. On the other hand, the writings of that time do not project rival methods (see Chapter 6). As a result, calling Laplace's method a paradigm seems warranted.

The method is mainly a creation of Laplace, and obviously partly Condorcet's, but Bayes' influence was not noticeable (see also Chang 1976). In fact, it is not very important whether Laplace was aware of Bayes' Essay or not. During the 19th century, neither the Bayes inference model nor Bayes' thought patterns were referred to in writings on probability theory. Hence, the method could only be called the Laplace paradigm, although there are similarities with Bayes' thinking model. The term Bayesian is strongly attached to modern statistical language, while Laplacean is not. Mentioning both gives a more illustrative expression of the nature of the paradigm. Therefore, calling the method the Laplace-Bayes paradigm is warranted.

[48] At the same time, Pearson was also extremely critical of Laplace. First he claims that Laplace was really only following Bayes; he continues in a footnote:

"... I do not think it is correct to say that Laplace was the first to treat the problem analytically. It all turns on the evaluation of the incomplete beta function. The methods of quadrature of Bayes and Price may be somewhat primitive, but I cannot see that they are much rougher than those used on this occasion by Laplace. There is no special merit in reducing any integral to terms in exponentials, unless these give an adequate approximation to the sought value. And Laplace does not really measure the closeness of his approximation nor indicate where it fails." (Pearson 1920)

6 The rise of statistical thinking in the 19th century

6.1 Development of probability theory after Laplace

The foundations of direct probability and inverse probability were laid at the end of 18th century and in the beginning of 19th century, and the French mathematicians, especially Laplace, made the most important contributions in both. Another central figure of that time was Carl Friedrich Gauss[49], who contributed nearly as extensively on probability theory as Laplace but from a slightly different perspective. Gauss and Laplace partly dealt with the same or similar problems and thus paved the way for each other. Hald (1998 and 2007) gives a thorough account of their contributions and "collaboration". Also Stigler (1986) devoted several pages to describe so-called Gauss-Laplace synthesis. Gauss has not been referred to here because he did not contribute much on inverse probability, which is the main topic of this thesis. Nevertheless, Gauss' impact on the development of probability theory and later on statistical science has been momentous, especially his contributions to the development of the least square method and the Normal Distribution and applications linked to it. These topics have been vital in shaping statistical science in its current form.

Since Laplace and Gauss, the development of probability theory subsided for a long time. In 1924, Rietz wrote:

> "The mathematical theory of statistics dates back to the first publication relating to Bernoulli's theorem in 1713. The line of development started by Bernoulli was carried forward by DeMoivre, Stirling, Maclaurin, and Euler culminating in the formulation of the Bernoulli theorem by Laplace in substantially the form in which it still holds a fundamental place in mathematical statistics.
> The *Théorie Analytique des Probabilités* of Laplace is undoubtedly the most significant publication at the basis of the development of mathematical statistics. Strangely enough, for a period of more than fifty years following the publication of the work of Laplace in 1812, little of importance was contributed to the subject. To be sure, the second law of error of Laplace was developed by Gauss and given its important place in the adjustment of observations, but there was on the whole relatively little progress...." (Rietz 1924)

Rietz, in fact, referred to the development in Western Europe where the period 1830–1890, starting after Laplace's and Gauss' most productive times, has been described as the one of clarification and consolidation of the works of Laplace

[49] **Johann Carl Friedrich Gauss** (1777–1855), a German mathematician and scientist, contributed to many fields, such as statistics, number theory, analysis, differential geometry, geodesy, electrostatics, astronomy, and optics. Gauss had an outstanding influence in many fields of mathematics and science and is often ranked as one of history's most influential mathematicians.

and Gauss. During that period, probability theory was extended to applications from the natural sciences to the social and biological sciences (see Hald 1998). The applications to social sciences led to the emergence of statistics (official statistics) and eventually to modern survey research. A significant line of development can be found in Germany, where the sovereign states founded statistical bureaus starting from the beginning of the 19th century, thus laying the foundations for economic and moral statistics.

Siméon-Denis Poisson (1781–1840) is a person who needs to be mentioned in this context. He was one of the most important apprentices of Laplace, and he is said to be the first to understand the fundamental importance of Laplace's probabilistic work. He was the one who introduced Laplacian theory on probabilistic problems to practical problems. Poisson was also a central person who systematized and extended Laplace's works and extended its application to vital statistics and law. Bru (2001) called him "the apostle of Laplacian science".

Poisson has been regarded as a mathematical genius. He published over 300 mathematical works covering a wide range, including applications from pure mathematics and probability theory to electricity, magnetism, and astronomy. In probability theory, Poisson is probably best known because of the distribution he invented, the Poisson Distribution, as a limiting distribution of binomial distribution. The Poisson distribution describes the probability that a random event will occur in a time or space interval under the conditions that the probability of the event occurring is very small, but the number of trials is very large so that the event actually occurs only a few times (Poisson 1829). According to Bru (ibid.), Poisson himself did not much appreciate his discovery. Its importance was observed only a century later in specific applications like queue theory. Poisson distribution is also in the kernel of von Bortkewicz' Law of Small Numbers (Bortkewitcz 1898), which eventually indicated the value of Poisson Distribution.

One of the Poisson's best known achievements in probability theory is currently known as the Weak Law of Large Numbers or the Poisson law of large numbers[50]. This provided a rationale for applying probability to social matters. It was deemed to explain how the statistical stability of social affairs was possible (see Hacking 1990). Poisson also extended Laplace's theory of errors to a situation in which the distribution of errors is not necessarily normal.

Poisson was a highly appreciated person, and he held several high academic positions and positions in the central administration. Bru (2001) claims that in every position he held, Poisson tried to demonstrate how Laplacian theory could be used to validate statistical data. He sought to popularize the statistical theories of Laplace also among "practical men". It has been credited to Poisson that Laplace's ideas spread so quickly and widely in France and other parts of Europe.

For nearly a century after Laplace, there were no noteworthy new developments in the theory of direct or inverse probability in Western Europe. However, important developments in probability theory emerged in Russia. Mathematicians like Bunyakovsky, Chebysev, Markov, Liabounov, Bernstein, and Tchuprov, to mention a few, made several important contributions in probability theory in the course of the 19th century. Their writings were often very theoretical, pub-

[50] Poisson coined the phrase 'law of large numbers' (see Hacking 1990).

lished in Russian, and their applications in empirical research were rare. Therefore, their influence in Western Europe was slight and slow.

Statistical thinking and statistics concerning human population started to emerge only at the end of Napoleon's era. In the beginning, the statistics were put together according to diverse principles, and therefore they were incoherent, irregularly collected, and difficult to have access to. One of the greatest obstacles for studying human population was the general belief that it was too heterogeneous, irregular, and unstable to be a subject of research. During the first half of the 19th century, research on human populations started to extend rapidly as a consequence of the increasing observations on invariances in social phenomena. An outcome of this was that statistics was perceived as a function with an autonomous role. As Armatte (2001) put it, the period 1820–1845 meant the hibernation of probabilities but a golden age of statistics.

Westergaard, in turn, called the period from 1830 to 1849 the Era of Enthusiasm. He wrote:

> "But enthusiasm is particularly evident in the two decades concerned. During this period statistics attracted public interest to an unusual degree. Official statistical institutions were founded or re-established in several countries, and numerous statistical societies sprang up and worked in co-operation with these institutions. Statistical journals were started. Highly talented authors, as, for instance, A. Quetelet, made the statistical results accessible in lucidly written books, in which the apparently dry observations were interpreted in an attractive manner. Of course this was not without drawbacks: there was a great temptation to dilettantism, and many statistical publications were superficial and full of hastily acquired results. No wonder that a reaction took place later on, with a deeply rooted suspicion against the statisticians. It proved necessary to pull down several badly constructed buildings and to replace them by more solid structures. But statistics profited from this passing popularity, inasmuch as large fields came under tillage, and the mass of statistical observations got an enormous increase." (Westergaard 1932)

The development took place in two different but related areas: in the development of statistical institutions and infrastructures; and in the discovery of social structures and their regularity – occasionally called social laws.

6.2 *Establishment of the infrastructures of statistics*

Westergaard (ibid.) called 1853–1888 the period of statistical congresses because during that period a series of international statistical congresses were organized. The importance and status of statistics was corroborated in successive meetings. Westergaard (ibid.) claims that the initiative leading to the establishment of the international statistical congress was principally due to Quetelet[51]. The first con-

[51] **Adolphe Quetelet** (1796–1874) was a Belgian scientist. He received his first doctorate in 1819 and after receiving this doctorate he taught mathematics for a while. In 1823, he went to Paris to study astronomy. Aside of astronomy he learned the theory of probability under Fourier and Laplace. In Belgium he first worked at an observatory but later he was appointed as the director of statistical bureau.

gress was held in Brussels in 1853 with Quetelet as main organizer and chairman. In the following 23 years, eight other congresses took place.

The activity around statistics was remarkable already before statistical Congresses, and throughout Europe, different bodies for statistics were established, such as statistical institutions and statistical societies. Especially the impact of the statistical societies was significant for the development of modern statistics and the professional principles of statistical work.

6.2.1 First statistical societies

Willcox (1934) listed fifteen statistical societies founded between 1834 and 1844. There existed some societies already before that period, such as the French Statistical society and those of Württemberg, Marseilles and Saxony. Westergaard (ibid.) commented that "Everybody seemed to have got statistics on the brain!" The societies were mainly founded as citizens' organizations outside statistical institutions. The most prominent motive seems to have been the interest in social questions (see Westergaard 1932 and Desrosiéres 1998). It is noteworthy that the foundations of the professional character of statistics and statistical work were laid by the first statistical societies.

Statistical societies in England

According to Westergaard (ibid.), the most noticeable development in statistical societies took place in England because in a short period more societies were founded there than anywhere else. The first statistical society in England was founded in Manchester in 1833[52]. Only few of the founding members were statisticians in the modern sense. The founders were partly driven by alarm over the acute social conditions in Manchester. The population of Manchester had grown by 45 per cent between the censuses of 1821 and 1831 as a consequence of rapid industrialization. This, in turn, caused an expansion in employment and it brought acute housing problems and diseases in its train. The objects of the Society were stated as "The collection of facts illustrative of the condition of Society and the discussion of subjects of Social and Political Economy, totally excluding party politics"[53].

The Manchester Statistical Society was a pioneering organization also in another respect. It was the first institute in Britain to systematically study social problems and to collect statistics for social purposes. In 1834, it carried out the first house-to-house social survey in England. The survey was composed of interviews of 4,102 families of working men in Manchester. One of the first published reports of social surveys was that of Heywood's *Report of an Enquiry, conducted House to House, into the state of 176 Families in Miles Platting, within the borough of Manchester, in 1837* (Heywood 1838). Later, the members of the society carried out several other surveys, including a survey of the state of education in Manchester and the surrounding boroughs.

[52] Statistical Society of Manchester still exists and is working actively.

[53] The early history of the Society has been well documented by Thomas S. Ashton (1934).

The best known statistical society in Great Britain is the London Statistical Society (LSS), founded in 1834. The society grew rapidly, and in 1838 it started to publish a journal which is now known as the *Journal of the Royal Statistical Society (JRSS)*. In 1887, the society was granted a royal charter and the Statistical Society of London became the Royal Statistical Society[54].

Soon after the Statistical Society of London was founded, similar societies were founded in larger industrial cities in Great Britain. This rapidly growing interest in statistical matters was associated with the British Statistical Movement, which was actively taking part in the work of the statistical societies. As an offspring of this activity, Charles Booth (1840–1916) devoted his fortune to surveying poverty in London. Among other things, he carried out a survey. Its results were published in a book entitled "*Life and Labour of the People in London*" (Booth 1889–1903). Benjamin Seebohm Rowntree (1871–1954) adopted Booth's method to study other English towns and compared them to London. The most famous of these is a comprehensive survey that he carried out on the living conditions of the poor in York. It was a complete enumeration during which investigators visited every working class home (see Rowntree 1901). Arthur Bowley has been regarded as a successor of this Statistical Movement (see Chapter 8).

Other famous statistical societies

One of the most famous statistical societies was founded in the Kingdom of Saxony in 1831. Already in the same year, the society published the first issue of its journal, *Mittheilungen des statistischen Vereins für das Königreich Sachsen.* The initiative to form the association came from the government and it got its mandate from the king of Saxony. In addition, a considerable number of private persons, "patriotic citizens", interested in statistics assisted the Society. In 1850, it was converted into a public institution of Saxony.

The American Statistical Association was founded in Boston in 1839. It was originally called the American Statistical Society, but the name was changed to the American Statistical Association (ASA) at its first annual meeting in 1840. In 1888, the society started a new publication that later became the *Journal of the American Statistical Association (JASA)*.

6.2.2 Statistical institutes

During the Era of Enthusiasm, official statistics made considerable steps forward, not only as to the quantity of collected data but also as to its quality. This fact, in combination with the work of statistical societies, made the avalanche of printed numbers possible (Hacking 1990).

In the Era of Enthusiasm, there was also remarkable activity in Germany. The Tariff Union was founded in 1833. It required regular censuses in all the German States that joined the Union, because the income from the tariffs

[54] The history of the society has been documented in great detail for example by Mouat (1885) and Hill (1984).

was distributed according to the number of inhabitants. Therefore, there were regular triennial enumerations until 1866.

Statistical activity in German states was already notable before the Tariff Union. The first statistical bureau in Germany, *Königlich Preußische Statistische Bureau,* was established in 1805 in Prussia. In 1808, the *Statistisch Topografisches Bureau* was established in the Kingdom of Bayern and in the next year it published its first yearbook. In 1807, it had already published a statistical atlas (*Statistische Darstellung der Königlich-Baierischen Staaten*). In 1821, in the Kingdom of Württemberg, a statistical bureau, *Statistisch Topografisches Bureau des Königreichs Württemberg,* was established. In 1826, it published its first yearbook (*Württenbergische Jahrbücher für Vaterländische Geschichte, Geographie, Statistik und Topographie*).

According to Westergaard (1932), noticeable progress was also made in England, where a statistical department was added to the Board of Trade in 1833. Equally important was the establishment of civil registration of vital statistics in 1837 for the registration of marriages, births, and deaths.

In France, the *Bureau de la Statistique generale* was re-establishment in 1833. I had been suppressed in 1812. The bureau had charge of several important subjects, such as population, finance, foreign trade and prices. But there was no decided centralisation, various branches being treated separately. A statistical service was created in 1844 and the following years in the Ministry for Travaux public (Westergaard 1932).

In Belgium, a Statistical Commission was organized in 1841 with the object of controlling the various branches of statistics, and Adolphe Quetelet became its president. The most significant event was the census of 1846, embracing population as well as agriculture and industry. The industrial census had a detailed classification of professions. The report contains the number of working men, their wages, engine horsepower, the number of looms and other utensils employed. The whole census was generally looked upon as a very important step forward and it was held as an example of a perfect census by other countries.

In Russia, the development was slightly different from the other countries in Europe. In 1864, provincial and district *Zemstvo* institutions were created in 33 districts of Russia. They were controlled by the Ministry of the Interior and the respective governors. Many Zemstvos started to organize their own local statistics for their needs, and by the end of the 19th century, 25 out of the 33 provincial Zemstvos had statistical bodies. Kaufman (1918) describes in detail the organisation and statistical activity of zemstvo offices.

Zemstvo administrators also carried out statistical surveys. Mespoulet (2002) argues that the quantity and diversity of statistical data needed by Zemstvo administrators stimulated methodological innovations in the field and influenced the rise and development of sampling in Russia at the end of the 19th and beginning of the 20th century. Mespoulet (ibid.) also argues that Kovalevsky's mathematical treatment on sampling theory and stratified sampling, published in 1924, is a synthesis of the Zemstvo statisticians' sampling practices and Russian academic statisticians' theoretical work.

6.2.3 The International Statistical Congresses

An important step forward was the first International Statistical Congress held in 1853. The chief object of the congress was a practical one, namely to promote the organization of official statistics and to unify the reports from various statistical institutions so as to make the documents comparable, and a cheap and easy exchange of statistical publications was recommended. The tendency in the "Congress Period" was chiefly to establish central statistical bureaus.

The program of the first conference covered the whole field of official statistics of that time, and the result of the meetings was a list of more or less detailed resolutions. As the aim was practical, there was no room for lectures on scientific problems, which later turned to be a fatal problem. Interestingly enough, one of the resolutions recommended a general register of the population in each commune, each family being allotted one page where future changes might be recorded.

As to the theory of statistics, a resolution was passed in Florence, in 1867 at the initiative of Quetelet, that there should be created a special section at future congresses, to deal with statistical questions in direct connection with the theory of probabilities. At the following congress, a corresponding problem was entered, recommending that statistical investigations should not only deal with averages, but with the deviations from the mean.

However, the International Statistical Congresses slowly faded away. The reasons that led to the end of International Statistical Congresses were realized and taken into account when a few years later the **International Statistical Institute** (ISI) was established. The first session of the ISI was held in Rome in 1887. Many of the leading statisticians of that time (e.g., Lexis, von Mayr, and Engel) took part in this meeting.

6.3 Discovery of social phenomena and their stability

People working in the statistical offices were civil servants, and statistical societies were manned by more or less ordinary citizens whose knowledge of statistics was limited. The large amount of numbers appeared incoherent – even chaotic – because there was nothing that tied them together (see Hacking 1990). It was necessary that scientists with vision develop a theory before social research and statistics could gain plausibility and become widely accepted. Characteristic to that era was disbelief that human populations could include such regularities. Theory building became possible when statistical offices and societies began to publish comparable statistics.

6.3.1 Early examples of social research

Before the avalanche of printed numbers started, there were only few examples of social research. William Petty's Political Arithmetic was practiced in some form up to the middle of the 19th century. Political arithmetic was a discipline of empirical collection of population records and preparation of accurate life tables. Its idea was bookkeeping of the population facts, not research. For example, Halley's life tables were mainly used for actuarial purposes. However, there are two famous examples of early investigations in the 18th century by Arbuthnot and Süssmilch.

In 1712, John Arbuthnot (1667–1735) published a paper which discussed the slight excess of male births over female births from a statistical point of view (Arbuthnot 1712). This paper is generally considered as the first application of probability to social statistics (Hacking 1975). Arbuthnot took 82 consecutive years of data on registered births in London and observed that on every recorded year more boys were born than girls. Arbuthnot argues that if there is an even chance for male and female births, the distribution of births should be like outcomes from tosses of a fair coin. He calculated that if his hypothesis were true, there would be an extremely small chance of getting 82 consecutive male years, i.e., $(½)^{82}$.

The first influence of enlightenment philosophy on statistical thinking is claimed to be seen in the works of Johann Peter Süssmilch (1707–1767). Süssmilch published a book in which he gave an extensive presentation of demographic material from a great number of sources, mainly from Germany, but also some from other countries (Süssmilch 1741). The leading motif in Süssmilch's work was the regularity that could be observed in the statistical figures, which were composed of material from larger areas or population groups. The interpretation Süssmilch gave to the statistical regularity has been considered a turning point in the gradual liberation of science from religious influence (see also Hacking 1975).

6.3.2 *Quetelet's contribution to statistics*

Adolphe Quetelet's impact on the emergence of statistics and statistical science was vital in two different areas: as an energetic organizer, he was a key person in forming the statistical institutions in Europe; and as a social scientist, he established a new branch of research which essentially was based on statistics. He appears to have been more innovative, energetic, and influential than any of his contemporaries. In addition, Quetelt's impact on the intellectual atmosphere in Europe was profound.

In the beginning the 19th century, a human population was considered a chaotic mass of individuals. An illustrative example is the reception to Quetelet's attempt to apply Laplace's estimation method. Approximately 25 years after Laplace had estimated the population of France, Quetelet wanted to attempt a similar estimation for the population of the Low Countries (Stigler 1986, p.163). When he had published his plans, baron de Keverberg[55] objected the plans (De Keverberg 1827). De Keverberg was afraid that the sample could never reach full "representativeness" because of the fundamental heterogeneity of the population. He writes

[55] Only little is known about **baron de Keverberg** (1768-1841). According to Stigler (1986), he apparently was serving as an official advisor on state matters in Low Countries.

> "The law regulating mortality is composed of a large number of elements: it is different for towns and for the flatlands, for large opulent cities and for smaller and less rich villages, and depending on whether the locality is dense or sparsely populated. This law depends on the terrain (raised or depressed), on the soil (dry or marshy), on the distance to the sea (near or far), on the comfort or distress of the people, on their diet, dress, and general manner of life, and on a multitude of local circumstances that would elude any a priori enumeration." (see Stigler 1986)

Laplace's estimation and inference had been based on the assumption that the birth and death rates were relatively homogeneous and stable. In essence, de Keverberg argues that the rates were not constant, or more generally, the stability of statistical ratios could not be assumed and hence the urn model could not be applied. Mathematically, it meant that binomial distribution could not be applied. The proportions could not be interpreted as probabilities because there were no homogeneous groups. In the absence of homogeneous groups, there could be no reliable inferences or inductive generalizations from a part to the whole. De Keverberg argued that the only solution was to take a complete enumeration and to describe the entire population. Quetelet accepted de Keverberg's argument about the lack of homogeneity and lost interest in partial investigations.

The debate about stability of statistical ratios, either biological or social, continued throughout the 19th century. Some, for example Poisson, argued that the laws of probability could be applied to human population and its social conditions, implying that he believed in the existence of homogeneous groups (see Stigler 1986). However, many were in favour of de Keverberg's argument that there were an unlimited number of ways of classifying social data after selection, and that homogeneous groups did not exist (Stigler ibid.). For example, in 1843, Cournot argued that there are countless ways to categorize social data.

> "Even a scientist of only average curiosity could classify births by birth order, by parent's age, profession, wealth, or religion, by season of the year, by whether it was a first marriage for both parents, and so forth." (Stigler 1986)

By that argument, striving for full coverage, a sample is simply impossible. It would ultimately mean that the only sample that would fulfil such a requirement would be the population itself. Underlying this debate was the deeper question: are there any stable regularities, or laws, in social science?

Quetelet and social research

Quetelet's interest in social phenomena and statistics grew after his visit to study in Paris in 1824, where Joseph Fourier[56] introduced him to Laplacian mathematics[57]. In the 1820s, Fourier had noticed that statistics on the number of births, deaths, marriages, suicides, and various crimes in the city of Paris had remarkably stable averages from year to year (see Porter 1986).This led Quetelet to think that

[56] **Jean Baptiste Joseph Fourier** (1768 – 1830) was a French mathematician and physicist probably best known for initiating the investigation of Fourier series. He was a student of Lagrande.

[57] There has been controversy on how much Quetelet met with Laplace. It has been documented that they met few times but Laplace was an old man at that time, and not very actively taking part in research anymore. Obviously, Laplace's direct influence was not noteworthy.

social phenomena are governed by laws as is nature. Poisson's Law of Large Numbers was the direct inspiration and indication of the social laws for Quetelet.

Quetelet was convinced that probability influenced the course of human affairs more than earlier generations had believed and more than his contemporaries did. He believed the law of error could also apply to human beings. If the phenomena were part of human nature, Quetelet concluded that it was possible to determine the average physical and intellectual features of a population. He believed that it was possible to identify the underlying regularities for both normal and abnormal behaviour.

Quetelet utilized the outburst of statistics, and in 1835 he published the book "*physique sociale*", or *Social Physics*, with a large number of tables on vital data, moral and criminal statistics, and anthropometry (Quetelet 1835). He did not confine only to presentation of the facts; he also derived new variables[58]. The tables included many different measures, but the most important point was that Quetelet described the distributions of both observed and derived variables – which without exception was the Normal Distribution – and showed that these variables were stable between countries and over time.

In *Social Physics*, Quetelet argued that it was necessary to go beyond the observation of singularities, since they were obstacles to perceiving "the laws of the human species":

> "Above all, we need to lose sight of man taken in isolation, and view him as merely a fraction of the species. By stripping him of his individuality, we will eliminate all that is merely accidental; and the individual particularities that have little or no effect on the mass will disappear by themselves, enabling us to apprehend the general results." (Quetelet 1835)

Quetelet noted that if things are examined at too close a range, it is possible to see only diversity, and observation limited to individual cases does not allow us to identify the "admirable laws". It is important to find the right observational distance to exclude what is accidental. Quetelet argued that this distance makes it possible to develop a science of collective phenomena: by losing sight of individuals, one can unravel, through the social phenomena that dominate the masses, a set of laws. (Quetelet 1835)

Quetelet demonstrated that there existed stability within social phenomena and that there existed regularity, or invariance, which could be called social law[59]. Quetelet's ideas were partly based on his own interpretation of Laplace's error law. Its importance was due to the Central Limit Theorem that Laplace derived mainly to analyze errors of measurements in astronomy. Quetelet was one of the first to apply the error law to human sciences.

[58] An example of the derived variables is the so-called Quetelet index, which in modern form is called the Body Mass Index, indicating obesity. The Body Mass Index, or Quetelet Index, is weight divided by the square of the height of a person ($BMI = W / H^2$).

[59] Behind Quetelet's idea was his aim to show that there were similar laws as the laws of nature that govern social life. However, many Quetelet's contemporaries did not accept the idea that regularities could be interpreted as laws.

The law of error

The error law, which was later named the Normal Distribution[60], became a central concept in Quetelet's analysis. His initial contribution was to show that nearly all features – biological, social, and moral – of human population followed this distribution. In the 1830s, Quetelet invented the concept of *"l'homme moyen"*, the average man. Quetelet justified this idea of the typical by saying,

> "If we do observe a [normal distribution] in nature, it is because nature was aiming for a target and missed due to random errors."

The target, the average man, represents the centre of the population. Quetelet interpreted the normal distribution as evidence that departures from the mean were like errors of measurements, so that the mean value was a 'true mean' which represented a real underlying value or type. The importance that Quetelet (and his followers) gave to the normal distribution led to an exaggerated idea of its prevalence, which was nicknamed "Queteletismus" or "Quetelism".

Eventually, the mean values in a "normal distribution" actually took on the prestige of a social law. Especially Quetelet thought that these statistical regularities were evidence of determinism. Individuals might think marriage was their decision, but since the number of total marriages was relatively stable from year to year, Quetelet claimed that the individuals were determined to marry.

Throughout his work, Quetelet held to the notion that there was no such thing as a chance event. He thought that all phenomena were 'caused' and related. If events have causes that persist through time periods, then the same events can be expected to reoccur. Quetelet claimed that "so long as the same causes exist, we must expect a repetition of the same effects" (Quetelet 1848).

General social conditions influencing the greater part of the social group result in sufficiently constant social phenomena. The study of large numbers suggests that general causes dominate the numerous influences of trivial ones. "The greater the number of individuals, the more the individual is effaced and allows to predominate the series of general facts which depend on general causes according to which society exists and is maintained" (Quetelet 1849). This 'doctrine of probabilities' has been regarded as the essence of Quetelet's statistical analysis.

The emergence of modern social research has been regarded beginning with the Quetelet's works. His *social physics* is often held as the origin of modern empirical sociology[61]. Quetelet published several books touching on the same topic (e.g., Quetelet 1848 and 1869), which subsequently inspired many scientists to develop new theories, thus generating a tradition of statistical research of social affairs. For example, Block devotes a considerable part of his textbook on statistics, *Traité théoretique et pratique de Statistique*, (Block 1886) in explaining the (statistical) regularities observed in different societies.

In addition, Quetelet had a strong influence on criminology. He showed that there was a relationship between crime and social factors. Among his findings

[60] The term"normal distribution" was coined by Galton at the end of the 19th century. Before that, the distribution was called the error law.

[61] Sociology is usually held a creation of the French philosopher August Comte (1798–1857), but he did not accept the statistical approach. Therefore empirical sociology is usually dedicated to Quetelet.

was the relationship between crime and age, as well as the relationship between crime and gender (and poverty, education, and alcohol consumption, etc.). Quetelet's statistical analysis of crimes had far-reaching consequences, especially on social research in German states and in France.

Quetelet also contributed to probability theory, but he did not make epoch-making discoveries. In 1849, he published a book that was in the form of letters, often cited later as Quetelet's *Letters on Probability* (Quetelet 1849). In this book, he outlined the use of probability in statistical research. The probability analysis of Quetelet was based on Laplace's and Poisson's ideas and he was one of the persons who strongly fostered adherence to the Laplace–Bayes paradigm. However, Quetelet did not touch on inverse probability, as he did not undertake any partial investigations.

Monograph surveys

Based on Quetelet's idea of the average man, a new type of survey research was introduced at the end of the 19th century by the French mineralogist and engineer Frederic LePlay. The method was called the Monograph study or the LePlay method. In the second half of the 19th century, the Monograph studies, or surveys, became popular, especially in exploring family budgets[62]. In the monograph method, it suffices to collect information only about typical cases, and investigation of extreme cases was to be avoided. Compared to complete enumeration, in a monograph survey, the amount of collected information per household was enormous. Sometimes the enumerators or observers stayed in the household for many days. Therefore, monograph studies were sometimes called in-depth surveys. The method was partly motivated by Quetelet's propagation of the normal distribution and his idea of the average man (see Desrosiéres 1998 and Hacking 1990).

The monograph design was widely applied at the end of the 19th century, and at the beginning of the 20th century it was still an officially accepted method used by the International Statistical Institute. Especially in Russia, the monographic method was very popular. For example, more than one-third of Tchuprov's textbook on statistical methods dealt directly or indirectly with monograph surveys (see Tchuprov 1910). Also in France, it was still frequently used in the beginning of the 20th century.

6.3.3 Statistics in German states

Statistics was taught at many German universities since the late 18th century. Statistical investigations were originally undertaken by individual scholars in search of the laws of social events, but soon statistical-topographical bureaus took over and statistics assumed a more purposeful orientation to solve problems of policy and crime control (see Tönnies 1925). Operated by professional

[62] Those surveys were the forerunners of the modern Household Budget Surveys, which most national statistical institutes conduct even today. In some countries, the current sampling design still has traces of the method that was applied at the end of the 19th century. For example, the sample is composed of households that have been purposefully selected to be typical households of specific socio-economic classes of that region.

statisticians, the offices collected information regardless of the specific needs of the political and legal administration. Statistical material was considered to benefit the public administration because the topics of information were not determined beforehand.

6.3.3.1 Engel and the first social law

It has been claimed that the first social law was discovered by Ernst Engel[63]. While studying in Paris, he came under the influence of Frederic LePlay who is a pioneer in the study of family budgets. Later, Engel stayed in Belgium for some time and became acquainted with Quetelet, who instilled in him the faith that it was possible to discover quantitative social laws. Hacking (1990) gives a comprehensive account and analysis of Engel's and Quetelet's collaboration.

The basis of Engel's investigations was family budget surveys in which data was collected using the monograph method. **Engel's law** deals with the relationship of expenditures for consumption in households to the income available. It states that the proportion of a consumer's budget spent on food tends to decline as the consumer's income goes up (Engel 1883). Engel's law has been confirmed in many surveys in all parts of the world. The significant point is that Engel demonstrated that general results in social statistics can be obtained from these individual data.

Engel had a strong interest in the development of international statistics, and he was an active participant in the International Statistical Congresses. After the International Statistical Congresses had faded away, Engel was one of the active founding members of the International Statistical Institute (see Hacking 1990).

Engel has been said to be one of the first who conceived statistics in the modern sense as a science on its own, as a structural theory of human societies which he called "demology" (Engel 1871). He was convinced that this science serves to recognize and analyze problems arising from the formation of societies. Sometimes Engel has been called the first statistician, and he has been claimed to point the way to the future of statistics as a science and as an essential tool of applied research (Porter 1986).

He was a prolific writer but his statistical papers are mostly published in the periodicals which he himself established, namely, *Preußisch Statistik*; *Zeitschrift des Statistischen Bureaus*, and *Zeitschrift des Statistischen Bureaus des Königreichs Sachsen* (Engel 1857, 1861, 1863, 1864, 1866).

6.3.3.2 Lexis and stability of statistical series

Probably the most famous of Quetelet's apprentices was Wilhelm Lexis[64]. In the history of statistics, he is best known because of his pioneering work on

[63] **Ernst Engel** (1821–1896) was born in Dresden, Germany. He studied at the Mining academy of Freiberg in Saxony. He held different government positions before he was appointed chief of the statistical department of Saxony. In 1860, he was appointed director of the statistical department of Preuss. Engel was one of the founding members of the ISI.

[64] **Wilhelm Lexis** (1837–1914) was German economist. He graduated from the University of Bonn in 1859. In 1861, Lexis went to Paris to study social sciences. In 1872, Lexis was appointed professor of economics at the University of Strasburg, and in 1874–1876 he acted as professor of statistics in Tartu. In 1876, he was appointed to the chair of economics at Freiburg. In Freiburg, Lexis made his major contributions to statistics

dispersion. Lexis' main topic was the development of mathematical methods in research on the stability of statistical series[65], especially concerning the ratio of sexes at birth (see Lexis 1877). In addition, he elaborated the methods of demography (Lexis1875 and 1903).

Using a binomial urn model to represent the annual number of male births, Lexis derived a dispersion coefficient Q (reportedly in homage to Quetelet), which is the ratio of the empirical variance of the series to the assumed theoretical variance. In the ideal case, Lexis refers to a "normal" dispersion when the fluctuations are purely due to chance, and the coefficient is equal to 1. But in most cases, the coefficient is different from 1, and thus differs from the binomial model. The fluctuations then indicate a "physical" rather than a chance component. Lexis classified these dispersions into two categories, "hypernormal" and "hyponormal", according to the value of Q ($Q > 1$ or $Q < 1$, respectively). He also showed that series of social data usually have a hypernormal dispersion.

Obviously, Lexis' most important specific contribution to statistical social science was the method to assess the stability of statistical series (see e.g. Lexis 1879). According to Porter (1986), the context of this work was social and ideological as well as mathematical. Basically, the measurement of dispersion of statistical series was intended as a critique of statistical determinism and a defence of the autonomy of the human will. Unlike Quetelet, Lexis stressed fluctuations, and in a sense he "corrected" Quetelet's work which aimed to set every series within a unique "normal" model by assuming their homogeneity and stability. However, methodologically, Lexis followed Quetelet in applying urn models to statistical series, but Lexis extended the traditional thought model introducing a model that included several urns.

Lexis' analysis also included certain weaknesses: he required a binomial dispersion for his series to be stable. It applied to the problem of year-to-year fluctuations in the sex ratio among children born in a city, but it also ruled out many interesting series. The problem posed the question of whether an empirical index of dispersion is consistent with the assumption that sex is governed by a simple chance mechanism. Stigler writes:

> "Many scientists attempted to adapt probability-based methods to social science problems, including Quetelet and Lexis, but in the end they were frustrated, Quetelet because his methods were too insensitive to segregate his data into categories amenable to statistical analysis, Lexis because his binomial models were insufficiently rich for interesting applications." (Stigler 1986)

Lexis' contemporaries, such as Tchuprov, Markov, and von Bortkewicz, pointed out the problems in Lexis' theory and attempted to correct them. In publications up to the period between the two world wars, the Continental School of mathematical statistics tended to follow the dispersion theory of Lexis, though. In Russia, Lexis' statistical views did not disappear from the statistical writings before the bolshevist revolution. Especially Tchuprov continued the work of

[65] In this context, Lexis created so-called cohort analysis and the Lexis diagram for it. They are in use even today, though in a modified form.

Lexis and published several articles on the stability of statistical series (see Tchuprov 1919, 1922, 1926).

In England, Lexis had a strongest influence on Edgeworth, even though Edgeworth criticized Lexis' theory. Lexis' analysis of dispersion has also been claimed to foreshadow the statistics of Karl Pearson and even R.A. Fisher's analysis of variance.

6.3.3.3 Von Mayr and criminal statistics

Hundreds of German works in criminal statistics were published in the 18th and 19th centuries but the writings of Georg von Mayr[66] appeared particularly significant. In the first volume of his *Statistik und Gesellschaftslehre*, von Mayr defined the scope of moral statistics as the study "of the circumstances and appearances of ethical life... whose mass-observation is accessible in number and measurement" (Mayr 1895). Moral statistics included the study of suicide, divorce, crime, and ethical aspects of other phenomena of life and society, the obedience to political rule, and the moral qualities of people as indicated by alcohol consumption (Mayr 1914). Research on statistical methods was systematically supported by von Mayr who conceived statistics as an autonomous discipline, with its own methods and objectives (Hertz 2001).

6.3.3.4 Von Bortkewicz and the Law of Small Numbers

Ladisdaus von Bortkewicz[67] was obviously the most influential person in mathematical statistics at the end of the 19th century. In 1898, he published *The Law of Small Numbers* (Bortkewicz 1898), in which he used Lexis' divergence coefficient Q. In this work, he was the first to note that events with low frequency in a large population followed a Poisson distribution even when the probabilities of the events varied. It has been argued that the Poisson distribution actually should have been named the von Bortkewicz distribution.

Von Bortkewicz was one of the main representatives of the "Continental school" in mathematical statistics and its application to statistics, but he left no monographs. Despite writing no monographs, von Bortkewicz wrote over 100 papers, almost exclusively in German. German scientists were only marginally interested in his works, but he was appreciated in Russia.

Von Bortkewicz was critical of the approach of Karl Pearson to statistics. He claimed that Pearson produced formulas to match observed results but with no theoretical reasoning. This, according to von Bortkewicz, was worthless.

[66] **Georg von Mayr** (1841–1925) was the director of the Royal Bavarian Regional Statistical Office at the same time being full Professor at the University of Munich. Von Mayr was the foremost representative of German administrative and bureaucratic statistics (Hertz 2001). In 1890, von Mayr founded *Allgemeines Statistisches Archiv* which is published even today. Von Mayr was also one of the founding members of the ISI and he took an active part in its meetings.

[67] Russian born **Ladisdaus von Bortkewicz** (1868 – 1931) is probably the most famous of Lexis' students. After achieving doctorate (in Göttingen), he spent some time in St. Petersburg teaching statistics. Then in 1901 he was appointed as a professor of statistics at the University of Berlin where he spent the rest of his life. Von Bortkewicz worked on mathematical statistics and applications to actuarial science and political economy.

6.4 Birth of modern Statistical Science

The observed stability of social phenomena and especially the observation that so many features had a regular frequency distribution were important for the further development of statistical methods. Quetelet's book *Letters on Probabilities*, especially the law of error, inspired Francis Galton[68]. Later he introduced the name Normal Distribution to the law of error. Galton argued in his book *Hereditary Genius* (Galton 1869) that the normal distribution would be expected to hold whenever there was a large number of similar events, each the result of the same conditions. Galton's observations on the distribution of human characteristics, both physical and mental, added to the belief on the stability of social phenomena.

At the end of the 19th century, Galton demonstrated that also the laws of heredity were stable. Actually he showed that genetic combinations are governed by the laws of probability, implying stability of inherited characteristics.

Galton's ideas had a strong influence on the development of the methods in statistical science. One of his major findings was the reversion, which was his formulation of regression, and its link to the bivariate normal distribution. Galton was able to place his research on heredity on a scientific basis by applying novel statistical concepts. This paved the way for the development of statistics as a science.

In 1889, Galton published *Natural inheritance* in which he presented a summary of the work he had done on correlation and regression (Galton 1889). He gave a good account of the concepts that he had introduced as well as the techniques that he had discovered. Karl Pearson read the book, and it had a profound influence on his thinking.

Galton had a long collaboration with Karl Pearson[69]. Pearson is generally held as a major innovator in the development of statistics as a serious scientific discipline in its own right. He founded the first statistical department at the University College London in 1911. In the statistical department, he incorporated both the Biometric Laboratories, which he had set up already in 1903, and Galton's Eugenics Laboratories.

Pearson's main focus in statistics was goodness-of-fit testing and later the development of the theory of distributions. Pearson made higher-level mathematics a requisite for doing statistics, and his work was more mathematically complex than Galton's. Galton thought that all data had to conform to the normal distribution, whereas Pearson emphasised that empirical distributions could take on any number of shapes.

[68] **Francis Galton** (1822 – 1911) was an English statistician, explorer, anthropologist, and eugenicist, known for his pioneering studies of human intelligence. Galton was the cousin of Charles Darwin and obviously they influenced each others thinking.

[69] **Karl Pearson** (1857 – 1936) was and English statistician, mathematician, eugenicist and Germanist. He was educated first at University College School, after which he went to King's College, Cambridge in 1876 to study mathematics. He then spent part of 1879 and 1880 studying medieval and 16th century German literature at the universities of Berlin and Heidelberg.

It has been claimed that Pearson created a new type of statistics in response to the conviction, held by many statisticians, that the normal distribution was the only feasible distribution for the analysis and interpretation of statistical data. At the end of the 19th century, most statisticians assumed that no other curve than the normal distribution could be used to describe data. This view was challenged by Pearson, and his derivation of the χ^2-distribution has been seen as a response to the "tyranny of the normal distribution".

Pearson's book, the *Grammar of Science* (Pearson 1892), became very famous and influential. This book represents his philosophy of science but does not reveal much about Pearson's thinking and ideas of the modern theory of mathematical statistics. Porter (2004) gives a comprehensive review of Karl Pearson's life and work.

Up to the beginning of the 20th century, only the large sample theory was studied within statistical science. W.S. Gosset (1876–1937) was developing quality control methods at the Guinness Brewing Company of Dublin. Sample sizes available for experimentation in brewing were necessarily small, and Gosset knew that a correct way of dealing with small samples was needed. He consulted Karl Pearson about the problem, and Pearson told him the current state of knowledge was unsatisfactory. The following year Gosset undertook a course of study under Pearson. An outcome of his study was the publication in 1908 of Gosset's paper (under the pseudonym "Student") on "*The Probable Error of a Mean*", which introduced a form of what later became known as Student's t-distribution (Student 1908a). The modern form of Student's t-distribution was later derived by R.A. Fisher and it was first published in 1925.

Gosset's derivation of t-distribution was a significant development for statistical inference. Its purpose was to form a tool for quality control, and in quality control, the problem is basically the same as inverse inference: the causes (of deviations in production) are inferred from observations. Later Fisher adapted Student's t-distribution as a central building block in his fiducial inference.

Gosset – as all other statisticians in England at that time – worked from the Laplace-Bayes paradigm, but he appeared somewhat hesitant about its validity (see Student 1908b). Zabell (2008) has thoroughly analyzed Gosset's statistical philosophy.

7 Emergence of the Representative Method

7.1 Introduction

Methodologically, sample surveys involve two different but intimately related and equally important questions: (1) how the sample should be selected from the given population, and (2) how to draw conclusions about the population based on the data that are obtained from the selected sample. Laplace gave one answer to the second question when he introduced the method to calculate probabilistic estimates, but the first question did not receive much attention during the 19th century.

In the second half of the 19th century, national statistical institutes had grown into the major organisations that produced statistics about socio-economic phenomena. Both in Europe and North America, the harmonised decennial censuses were established as the main sources and the "officially" accepted data collection schemes for population statistics (see Porter 1986 and Hacking 1990). In a census, neither of the basic questions of sample surveys is relevant: all population units are selected and the results do not include sampling errors. Therefore, the basic problem of survey research did not receive much attention.

The total amount of data collected in a census is enormous due to the size of populations, and it requires a lot of effort to gather and to process the data. Consequently, the information content of a census, i.e., the information collected from each household, always remains fairly modest. Even with a small amount of data per unit, the processing time for the census results took several years in the 19th century. The need for more detailed and timelier information than a census could provide grew rapidly at the end of the 19th century along with the development democratic societies and also because of the profound changes in society which industrialization and urbanisation had set forth.

Monograph surveys were widely conducted to reveal in-depth information about households. However, monograph surveys did not aim at the same goal as to what censuses were supposed to do. By a monograph survey, it is possible to explore economic and social facts of specific families and have a description of (e.g.) a typical working-class family budget, but it is not possible to disclose population distributions. For example, by a monograph survey, it is not possible to tell how many households (in a country) live in poverty, or what the total consumption of meat is, or what the average family size is, or any other general fact about the population. A monograph survey does not address the same questions as a census.

Obviously, neither probability theory nor Laplace's method of inverse inference had a noticeable influence on the data collection methods in statistical offices or on statistical thinking in general (see Westergaard 1932). In a way, this was expected because of the small amount of partial investigations in the 19th century; and in those investigations that were carried out, inference was done on an intuitive basis. Statisticians working for the national statistical institutes did not pay much attention to the mathematical aspect of statistics, and the thought that probability theory could have a role in statistics had not been articulated strongly enough. In addition, as Porter (1986) claims, officers who worked at national statistical institutes actually had little or no mathematical training.

Porter (ibid.) concluded that there is only a little mystery in the attitudes of statisticians towards probability theory. At the end of the 19th century, statisticians were almost unanimously distrustful of sampling and emphasized at every opportunity the importance of complete enumeration. Porter (ibid.) also claims that the scepticism of statisticians about inference from samples was not wholly unjustified, for in the absence of reliable information about the population as a whole, it was difficult to know if a particular sample was adequately representative. Also Westergaard (1932) claims that the calculus of probability had less influence than expected because its authors chiefly confined themselves to abstract theories that had little or nothing to do with reality.

7.2 The Representative Method

In texts that touch on the history of survey sampling, it is commonly held that the idea of the Representative Method was first developed by Anders Kiaer[70] at the very end of the 19th century. Obviously, however, the idea of using representative sampling is older (see Didier 2002). For example, Jensen in the report to the ISI (Jensen 1926) claimed:

> "The method recommended by A. N. Kiaer in the nineties was, by the way, previously used in enquiries of various kinds, as for instance in an enquiry regarding housing conditions and rent undertaken by the Municipal Office of Statistics of Copenhagen in 1885."[71]

In the survey to which Jensen refers, the sampling method was not exactly the same as what Kiaer had used in Norway, but the methods did not differ in essence. It is possible that Kiaer was aware of the partial investigation carried out in Denmark, although he did not refer to it.

In agricultural research, methods resembling the representative method were frequently applied in many countries. Especially in Russia, agricultural surveys and surveys on peasants' living conditions were common (see e.g. Zarkovic 1956, 1962 or Kohn 1922). Mespoulet (2002) has found out that Russian textbooks on statistics usually state that in Russia the first survey "on parts of the whole" was already carried out in 1875. Mespoulet (ibid.) also claims, referring

[70] **Anders Kiaer** (1838-1919) was one of the founders and first director of the Statistical Central Bureau of Norway. He was also responsible for decennial censuses of the population and agriculture.

[71] In this investigation, the sample consisted of 36 streets distributed all over the town. In these 36 streets, there were altogether 9,366 dwellings, which was one-seventh of all the dwellings in the town; and the number of inhabitants in these was 39,350, which was one-seventh of the whole population of the town. The final sample consisted of all the dwellings on the selected streets (Jensen 1926, p. 407).

to Kaufman (1922), that A. Kaufman already carried out a sampling survey in Russia between 1887 and 1890. In that survey, sampling relied on random selection. For example, the statistician was "... ordered that every tenth or twentieth person taken in alphabetical or other mechanical order should be questioned" (Kaufman 1918). According to Didier (2002) the United States Department of Agriculture conducted partial surveys beginning in 1863 for the purpose of measuring the country's agricultural production.

Even though the method may have been used earlier, Kiaer's first public appearance may be considered a turning point in the history of sample surveys on human populations. He was a central pacemaker and advocate of the Representative Method for several years. With the first public presentation of his ideas in 1895, he started the process that ended in the development of modern survey sampling theory and methods. In a review on the history of the survey method, Kruskal and Mosteller (1980) noted that Kiaer was the first man ever to use *analytically* the term '*la Méthode Représentative*' in the 9th meeting of the ISI in 1895. Carroll D. Wright[72] had earlier used the same expression, but according to Kruskal and Mosteller (ibid.), the term he used was so shallow and used in a less influential way so that it cannot be considered as the starting point. Also Jensen concluded in the report to the ISI:

> "...it must doubtless be admitted that the official statistics in both the United States and Canada have made wide use of partial investigations as substitutes for complete statistics, but in the opinion of the author the methods used there are on the whole not of such a kind that they can be termed «representative» in the narrower sense which this expression, in our opinion, really ought to symbolise ... This applies to the numerous partial investigations which have been made by the United States Department of Labor and by Canada's Dominion Bureau of Statistics, and it also applies, for instance, to the interesting efforts for the promotion of the economic use of Canada's natural resources made under the leadership of the so-called «Commission of Conservation»." (Jensen 1926)

Nevertheless, Wright obviously was an important person in developing the method. He undertook partial investigations in the U.S., and he also was in correspondence with Kiaer. In the report of the Bern meeting, Kiaer gives an allusive comment about the work done in America (Kiaer 1895).

7.2.1 Kiaer's Representative Method

Kiaer was one of the first who used the Representative Method for collecting data independently of the census. He carried out several purely sampling investigations for the Statistical Central Bureau of Norway, including both the first *Income Distribution Survey* (see Kiaer 1897a) and the first *Survey on Level of Living (Living Conditions Survey)* in Norway (see also Jensen 1926 and Seng 1951).

When Kiaer presented his idea on partial investigations for the first time (in the ISI meeting in 1895) he had already carried out two surveys. However, Kiaer

[72] **Carroll D. Wright** (1840 -1909) served as a professor at several universities in the U.S., and he was the founder and director of the U.S. Department of Labor. He also served as the president of the American Statistical Association from 1897 to 1909. He took active part in the meetings of the ISI.

had only partly processed the results before his first presentation and therefore he was able to present the idea mainly in theory. Two years later, in 1897, when he gave the next speech about the *Representative Method* to the Norwegian Academy of Science and Letters, he was able to prove his arguments by empirical facts (Kiaer 1897a).

Kiaer was methodologically oriented, but obviously he was not mathematically oriented. All his presentations were verbal without any formal description of the methods. This was not unusual at that time, though. Most of the presentations in the ISI meetings were only verbal. Apparently, Kiaer was unsure about giving a presentation of the method. He had asked Professor Harald Westergaard from Copenhagen, "*dont les connaissances théoriques dans cette matière sont universellement connues*", to give a more general and international presentation on the method but Westergaard was not able to attend the conference in Bern (see Kiaer 1895). Instead of a theoretical presentation, Kiaer gave a detailed description on how the data collection had been carried out.

Later, Kiaer's address to the Academy of Sciences and Letters of Norway (Kiaer 1897a) was more detailed than the presentation at the ISI meeting, and it gives much better insight in Kiaer's thinking. He started the address by explaining the ideas of the method:

> "The characteristic feature of this method is that in connection with the general and complete information provided by the established statistics for the field of study as a whole, more penetrating, more detailed and more specialized surveys are instituted, based on certain points or limited areas, distributed over the domain of study in such a way and selected in such a manner that they will yield a sample that might be assumed to constitute a correct representation of the whole." (Kiaer 1897a)

Kiaer continues by taking examples from natural sciences where partial investigations had been used already for a long time. He mentions mineralogy, surveys of the flora and fauna of a country, meteorological observations, etc., and concludes that many more examples could be quoted, but social phenomena are so diversified, widespread, and complex that they do not fully justify the comparison with the flora of a country. He noted that, in order to obtain a deep understanding of the social phenomena, it is necessary to "go deep in details and to formulate a series of special questions" to such an extent that it would become too expensive to carry out a full enumeration in a country or even in a large town. Finally, Kiaer came to the conclusion:

> "We are forced therefore, in social research as in the natural sciences, to conduct partial investigations, and it is obvious that these will give the best results if they are designed so that the scattered fields of observation together form a representative picture of the whole field of study." (Kiaer 1897a)

In his address to the Academy of Norway, Kiaer described the first survey in the following manner:

> "I am referring to the very extensive and detailed investigation of the living conditions of the various classes in the community, in particular the one concerning workers, that has been initiated by the Parliamentary Commission appointed by

> the Storting[73] in 1894 for considering disablement and old-age pensions, in which comprehensive statistical material was collected that not only supplemented and completed the information already collected on personal income and property, but also provided new information in many important respects. This statistical material consists of individual returns on a number of economic and other personal characteristics — including also what might be described as a historic record of the economic life of the individuals — for a representative sample of 80 000 adult men and women from various walks of life, of which, for the working classes in particular, comparable returns from between 40 000 and 50 000 persons were collected." (Kiaer 1897a)

Enumerators, hired and trained only for this purpose, filled in a total of 80,000 forms on the adult population in Norway according to the rules which Kiaer had laid out. Simultaneously with this data collection, another 40,000 forms were collected for another survey by a slightly different Representative Method in the areas where members of the working class lived. This survey did not get much attention in Kiaer's reports, however.

For the sample of 80,000 respondents, the households in Norway were divided into two main strata (Kiaer did not use the term 'strata') based on the 1891 census. Approximately 20,000 respondents were selected from cities and the rest from rural areas in accordance with the population distribution of the country in the previous census. The actual sample was selected by a slightly different method in cities and in the rural areas.

Out of the 61 cities of Norway, 13 "representative" cities were selected: All the five big cities having more than 20,000 inhabitants were included, and the remaining eight cities represented the medium sized (6) and small towns (2). The proportion of respondents (of the total population) in cities varied: in the middle-sized and small cities, the proportion was greater than that in the big cities. In the capital of Norway, Kristiania[74], the proportion was 1/16; in the medium-sized towns, the proportion varied between 1/12 and 1/9; and in the small towns, it was 1/4 or 1/3 of the population. This was motivated, according to Kiaer, by the fact that the middle-sized and small cities did not only represent themselves but a larger number of similar cities. He concluded that "Taken as a whole this is expected to supply a fairly correct *miniature of the urban population in the whole of the country.*" (Italics by the author of this study.)

In Kristiania, a census was carried out every year. Therefore, it was more accurately known how many people lived on each of the 400 streets of the city. The streets were sorted into four categories according to the number of inhabitants on them, and a total of 62 streets were selected. After the streets were selected, their distribution over the city was taken into consideration to ensure the largest possible dispersion and the 'representative character' of the enumerated areas. A specific selection scheme for respondents was then specified for each category: every house on the smallest 6 streets was included (i.e., the whole adult population was enumerated). On the next larger streets, 20 were included in the sample, and on these, every second house was included. Amongst the second largest category of streets, 20 were included in the sample, which was every fourth

[73] The parliament of Norway.

[74] Kristiania is the former name of the capital of Norway. Currently, it is called Oslo.

street in this category, and in the selected streets, every fifth house was enumerated. And among the category of the biggest streets (streets on which lived more than 1,000 inhabitants), the adult population was enumerated on half of the streets, and on these, one out of ten houses was included in the sample.

The booklet where Kiaer's speech to the Academy is published also includes a map of Kristiania as an annex, in which the selected streets and houses are marked with red spots (Kiaer 1897a). The map was an expedient in designing the enumeration. Kiaer did not explain how the houses on streets were selected. Obviously, the selection was "mechanical", or a systematic selection in modern terms.

In the medium-sized towns, the sample was selected using the same principles, though in a slightly simplified manner. In the smallest towns, the whole adult population in three or four houses was enumerated.

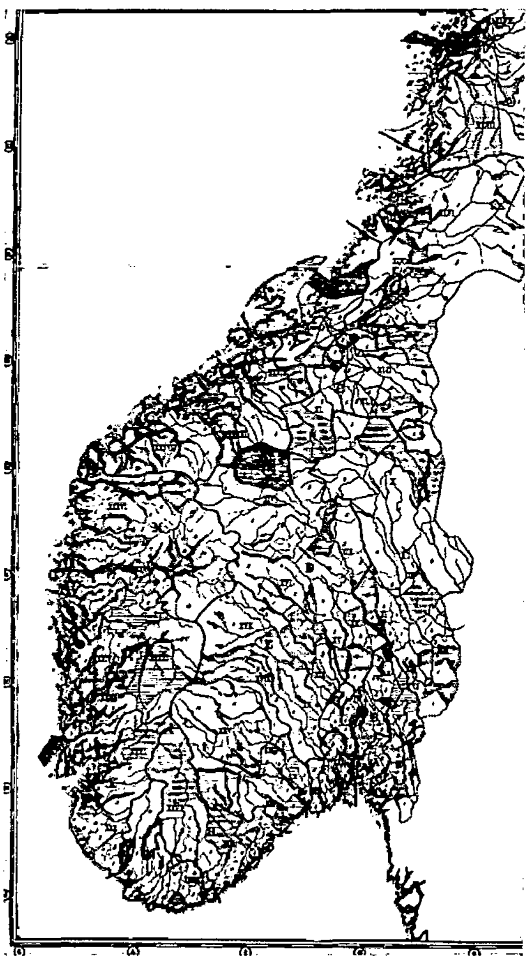


Figure 7.1:
Copy of a part of the map that Kiaer used in stratification of municipalities. Different shading indicates different categories. (Source Kiaer 1897a)

Also, in the rural area, the number of informants in each of the 18 counties in Norway was decided on the basis of the 1891 census. Information about population was used to determine the number of forms to be collected from each county so that the proportions in the sample were the same as in the census. To obtain "as far as possible, a correct representation of the population within each county", the local government districts (the municipalities) in each county were classified according to their main industry either as predominately crop-farming, or predominately livestock-farming, or forestry, or fisheries, or shipping and manufacturing municipalities. In addition, the geographical distribution was taken into account (see Figure 7.1). In relation to the population as a whole, the representative municipalities in each category and also the number of informants were decided so that each industry attained a correct weight. If this was not to be the case, e.g., if the selected districts showed too high a proportion of the population to some branch of industry, Kiaer attempted to correct it by adjusting the number of forms allocated to the districts. He gave a few detailed examples to illustrate how this was done

The total number of the representative municipalities was 109, an average of six in each county. The number of parishes involved was estimated by Kiaer to be more than 200. The total number of municipalities in Norway was 498, hence 21.9% of the municipalities were included.

The procedure for distributing the required number of forms within the districts was also designed on the basis of the census information: the number of forms within each parish in the district was determined separately for built-up areas and heterogeneous areas. Furthermore, the enumerators were instructed, according to the information available, "to select respondents in a representative manner within the locality". Kiaer expressed his worry that the responsibility had to be left with the enumerators and to be based on their judgment. Enumerators were instructed to follow distinct routes and while doing so, to visit houses of different types in the same neighbourhood and in particular check that not only typical middle-class houses were visited but also the more well-to-do and the poor-looking houses, both for families and single persons. Kiaer concluded

> "This task was obviously not always an easy one, and one would expect that the enumerators in a large number of cases would tend to collect forms from houses that would be more easily accessible and this would have had some effect on the correctness of the representativeness of the survey. It was attempted to overcome this difficulty through the instructions for the enumerators." (Kiaer 1897a)

It is quite obvious that the element of "human choice" was unavoidable in the selection of respondents, but Kiaer tried to reduce it as much as he could. Obviously, a great number of enumerators were needed in the data collection (Kiaer does not tell how many), and therefore the "human choices" probably did not cause any bias in the results.

Kiaer does not tell in his reports how the estimates were calculated. Probably the reason was that the representative sample was constructed as a miniature of the population and therefore the calculation is trivial: the sample mean is the estimate of the population mean, and the estimate of the population total could be attained simply by multiplying the sample total by the inverse of the sampling fraction. Being a miniature of the population was a central requirement in Kiaer's idea of representativeness and the sampling design aimed at reaching it on the basis of the knowledge that censuses provided about the population. Kiaer put a lot of effort into proving that the sample distributions of some important background variables are 'close enough' to the population distributions.

The aim in the selection of households was to select them so that the obtained sample would cover all types of social classes and in the correct proportions, i.e., represent the population and its variation. Subjective selection was applied in the selection of the areas. Laplace already applied the same principle. The final selection of units within the areas was more or less haphazard but not strictly random. The major innovation in Kiaer's method, as compared to monograph studies, was that the variation in population was considered an essential characteristic and the sample was selected in such a manner that the variation within the population was covered in correct proportions. Obviously, Lexis' analysis of dispersions was one source of ideas in designing the sampling method.

Kiaer's Representative Method did not include a method for estimating the accuracy of estimates, but he was aware of sampling variation. Later, he even suggested a method to assess the stability of the obtained samples by a method that was based on the idea of sample re-use (Kiaer 1901, p. 68). It must be emphasised that Kiaer was aware of the probability laws concerning sampling. In the report, he says:

> "One must of course expect that the more detailed the classification considered, the larger the deviation between the sample and the census. This is a natural consequence of the law of large numbers which, according to conditions, requires a larger or smaller number of observations in order to obtain acceptable results."

The main reasons why Kiaer chose a method that did not employ random selection were practical, not the lack of knowledge or the lack of reliance. This can be concluded from the fact that earlier, Kiaer had applied an explicitly random selection of informants in an investigation concerning income distribution in Norway (see Kiaer 1897a).

In the other survey described in the reports, a survey on personal income and property in Norway, the representative sampling method was employed in direct connection with the population census of 1891. Kiaer used the opportunity that had emerged during the processing of the census forms to extract the information from them to select a representative sample of the male population in Norway.

> "It is easy to see that the original census-material, for a survey of this kind, would be unmanageable if complete coverage had been attempted, so it has been subjected to a three-way reduction by the following method. Firstly the material was reduced through the sample of local government districts, next by the restriction to every 5th year of age, and lastly by the final reduction through including only persons with names starting with certain letters.
>
> Whether the representation can be considered an approximate miniature of the whole field of study obviously depends in part on a correct method being employed for the selection of the sample and partly on the sample being of sufficient size. "

Kiaer considered the selection of the sample more important than the size of the sample. He selected government districts that were included in a previous government investigation because "... it was considered desirable that the results of the survey could be directly related". Kiaer estimated that these districts (128 rural local government districts and 23 towns and cities) have a sufficient geographic distribution over the whole country, and therefore they could be assumed to provide, at least approximately, a correct representation of the whole country.

The second criterion to reduce the census material was to include only persons according to age at 5 year intervals (i.e., men who in 1890 reached the ages of 17, 22, 27, 32, 37, and so on). Kiaer states that "it seems to be fully in accordance with the representative principle. I have pointed out in a paper [reference to Kiaer 1897b] that no particular reason can be found for a sample drawn in this manner not to provide a true miniature of one fifth of the size of the whole."

The third criterion was to include only persons whose names start with certain letters. The letters were: A, B, L, M, and N for the rural districts and smaller towns, and the three last ones for the nine largest towns. Kiaer stated that they

> "...seem to be related to the persons from amongst whom the sample should be selected in a haphazard and random way, so that a sample selected in this manner would turn out in the same way as would have been the case had the sample been selected through the drawing of lots in order to avoid, in the most stringent manner, any procedure that could give preference to persons in certain occupations or belonging to particular social strata."

This citation shows that Kiaer was aware of the merits of a random selection and its implications. Practicalities in a large-scale survey hindered its use.

The reason why Kiaer used only half the letters for the nine largest towns was an attempt to reduce the effect that the largest towns would have had in the results concerning incomes: If the same proportion had been selected from all the largest towns and only some of the smaller towns had been included in the sample, the weight of the larger towns in the data would have been too high. It was known that the average personal income was higher in the larger tows than in the smaller ones.

> "The total number of forms that was collected for this survey was 11 427 of which 7 162 came from the rural districts and 4 267 from the towns and cities. The sample amounted to 7.85 % of the male population in the selected age groups in the rural districts and 15.7 % in the urban districts, and 1.54 and 3.1 % of the total adult male population in rural and urban districts respectively. In preparing the tables for the whole country the figures for the rural districts were given double weight."[75] (Kiaer 1897a)

Kiaer concluded, after various comparisons of the obtained sample with the census data, that their compatibility "proved to be very satisfactory" (see Kiaer 1897a). The purpose of Kiaer in presenting the other form of the representative method was to show that representative samples could be obtained in many ways. Jensen (1926) described this method as "purposive selection of groups combined with random selection of units". In fact, Kiaer's method to select a respondent within a household is similar to the so-called Kish method, which is currently applied to form a haphazard sample when the frame does not provide enough information for more a specific selection of respondents. Kiaer's method also resembles the so-called closest birthday method[76] applied occasionally to select a respondent within a household.

A central idea in estimating the representative nature of an obtained sample was to compare it to the latest census data. Kiaer (1897a) pointed out that the results of a partial investigation could be controlled to a certain degree even if general statistics were not available. For example, regularity of the observed phenomena was one kind of a control. In addition, results could be controlled by comparing

[75] The weighting of the sample that Kiaer describes was based on intuition and common sense, but it was already taken into account in designing the data collection.

[76] In the closest birthday method, the member of the household whose birthday is closest at the time of interview is selected as the respondent.

them to the results of other partial investigations obtained by different representative designs. Kiaer concluded that if one obtains approximately the same results by various methods, greater reliability can be placed on the results.

Approximately 40 years later, in the context of a population census in the U.S., additional information was collected with a partial investigation. In that roughly 12% of respondents were asked an extra set of questions which could not be included in the census forms. Stephan with his colleagues reported the lines of thoughts that finally led to the applied method (see Stephan et. al. 1944). Interestingly enough, the considerations had many parallels with those of Kiaer, and one of the possible methods was close to the Kiaer's Representative Method. Also the method which was finally selected had similarities to Kiaer's method.

7.2.2 Discussion on Kiaer's Representative Method

Kiaer closed his presentation in Bern by concluding that (1) the Representative Methods could be of great importance for the development of statistics, especially if representative investigations are arranged in such a manner that they can be controlled, concerning the main points, with the aid of general investigations; (2) depending on the circumstances, there may be different methods between which one is able to choose; (3) as a consequence, the advantages and the inconveniences of diverse methods deserve to be recommended for study and to be discussed by statisticians (Kiaer 1895).

With his visions, Kiaer was ahead of his time[77], which can be concluded from the reactions his paper raised at the ISI meeting. Obviously, in that meeting he was not able to properly defend his ideas, but later in the speech to the Academy of Norway he could (Kiaer 1897a). According to Kiaer, Professor G. von Mayr remarked that partial investigations may have some limited value "but it is a value restricted to terrain already illuminated by full coverage." Von Mayr continued by saying that for legislative and administrative purposes, restricted surveys might be useful but they could never replace complete statistical surveys. Von Mayr added still that it would be particularly dangerous to express views to the contrary in an assembly of statisticians. Kiaer (1897a) interpreted that this was some kind of warning for the "noticeable tendency" amongst the mathematicians to replace observations by calculations. The last sentence of von Mayr's comment almost became a catch phrase: *«Il faut rester ferme et dire: pas de calcul là où l'observation peut être faite.»* («We must remain strong and say: no calculation when observations can be made.») [78].

Luigi Bodio, the director-general of the Italian Statistical Bureau, supported von Mayr's views. The Austrian statistician Herr Rauchberg stated that further discussion of the matter was unnecessary because in statistics there would never

[77] Kiaer was a pioneer in other respects, as well. For instance, he used a punch-card machine, a Hollerith machine, in processing data of statistical surveys as early as 1894 – only a few years after Hollerith had invented it.

[78] G. von Mayr was one of the most prominent statisticians of that time and a founder of the ISI. He was also critical of the use of the monograph method (Hertz 2001). His main interest was moral statistics, including the studies on suicide, divorce, and crime, which cannot be covered by partial investigations,

be a question of anything but complete surveys covering the whole field of study. Herr G. E. Milliet from Switzerland said that the type of *pars pro toto* statistics recommended [by Kiaer] might at some times provide interesting information but he demanded that incomplete surveys should not be granted equal status with the statistical ideal and with *"la statistique serieuse"*.

Later in his address to the Academy Sciences of Norway, Kiaer said that those who opposed the representative method in the ISI meeting had implicitly also attacked the Monograph method, which was already accepted by the ISI. Therefore, in Kiaer's opinion, the French statistician Cheyson had expressed his hope that the result of the discussion would not be to prejudice the Monograph method. (The monograph method was widely used in France.) In Cheyson's mind, the methods complement each other. After Cheyson, another French statistician and the Vice-President of the Institute, E. Levasseur, emphasised that there were actually three methods: (1) general statistics (*les statistiques generales*), which in practice means censuses; (2) monographies, which were concerned with detailed descriptions of an object of phenomenon; and (3) statistical explorations (*les explorations statistiques*), which were virtually the same as the Representative Method. Levasseur continued by proposing that "the Institute would do well to promote discussion of the matters relating to the third method". This proposal was accepted by a very narrow margin against Herr Rauchberg's proposal that the question should not be taken up by the Institute anymore[79].

7.3 Developments after the meeting in Bern

The criticism at the ISI meeting was almost shattering, and the critics were the most eminent statisticians of the time and leading figures of the ISI. Kiaer was not completely insensitive to the critic (see Kiaer 1897a). Nonetheless, the method was accepted on the agenda of the next meeting and Kiaer continued to elaborate the method and to defend and promote it. He also wrote an extensive article in the *Algemeines Statistisches Archive* (1899) in which he showed that the Representative Method could also be used in agriculture and forestry, not only for social and economic enquiries. He also recommended that the questions in the inquiry should be as close as possible to the ones used in the census so that the results could be controlled.

Kiaer gave presentations on his method at following meetings of the ISI, in St. Petersburg (1897), in Budapest (1901) and Berlin (1903). The presentations on the Representative Method were usually given in the same session with the Monographic method. In the Budapest meeting, during the discussion on Kiaer's method, von Bortkewicz[80] said that he had used "formulas deduced for analogous cases by Poisson, to find out if the differences between two numbers was

[79] According to an unofficial protocol done by the Swiss Statistical Society the assemply decided to accept Herr Rauchberg's proposal! (Malaguerra 2000)

[80] Von Bortkewicz was the leading figure in the continental mathematical statistical school at that time.

was fortuitous or not" and that he concluded that the difference between census data and partial data was too big. He did not explain in detail how he came up to this conclusion using probability theory, though. Von Bortkewicz' comment did not raise any discussion and Kiaer did not react on that, either (see also Desroières 1998).

At the St. Petersburg meeting in 1897, the ISI nominated – on Kiaer's proposal – a sub-committee to consider the limits of the application of partial investigation and to give recommendations of the best "representative typological methods". The members of the committee were J. Bertillon (of France), L. Bodio (of Italy), J. Körösi (of Hungary), G. von Mayr (of Germany), C. Wright (of the United States.), and Kiaer who was nominated as the reporter[81].

The sub-committee published its proposal for resolution in the meeting in Berlin, in 1903. The resolution said that the method could be used in certain specific cases provided that it is done using strict guidelines. The proposal also said that the question will be kept on the agenda, so that the report on the applications of the method in practice and on the results obtained by it can be presented. For some reason, however, the Representative Method disappeared from the agenda of the ISI meetings for twenty years. Probably, the reason is why Kiaer did not touch the topic since 1906 was the hard criticism that emerged in Norway (Lie 2002).

7.3.1 Final approval of the Representative Method

Kiaer's representative survey method was approved as a valid statistical method only at the ISI Rome meeting in 1925, five years after Kiaer's death. At the same meeting, the ISI nominated a commission to study the applications of the Representative Method in Statistics. Mr. Adolph Jensen of Denmark was appointed as reporter to the commission and the other members were Arthur Bowley, Corrado Gini (from Italy), Lucien March (from France), Coenraad Alexander Verrijn Stuart (from Holland)[82], and Frantz Zizek (from Germany).

The report was presented at the meeting in Rome in 1925 and published in 1926 (Jensen 1926). The report starts out "... Three decades have elapsed since our late lamented colleague, the Norwegian A.N. Kiaer, placed this matter for the first time on the agenda for the session of the institute...". Later in the report, Jensen writes, "The investigations made by A.N. Kiaer in the nineties, which form the starting point for the discussion on the Representative Method at a number of meetings of the International Institute of Statistics, were representative in the truest sense of the word."

[81] **Jacques Bertillon** was the head of the Paris bureau of vital statistics. **Luigi Bodio** was one of the founders of the Italian Statistics. He was the first General Secretary of the International Statistical Institute (ISI) and among the first presidents of the ISI. **Joseph de Körösy** was the director of the Budapest communal bureau of statistics.

[82] **Corrado Gini** was an Italian statistician who developed the Gini coefficient. He held several chairs in Statistics in different universities and founded Metron in 1920. **Lucien March** was the superintendent of the Bureau of the Statistique Générale de la France. **C.-A. Verrijn Stuart** was the director of the Statistical Central Bureau of Holland.

As an annex to the report was Bowley's memorandum on the accuracy of estimates obtained by the Representative Method. It included two different methods for sample selection: random selection and purposive selection. Random selection meant, according to the memorandum, a method of selecting for investigation **a number of units** using some mechanical principle or other which is not connected to the subject or with the purpose of the inquiry and the selection arranged in such a way that every unit in the population had an equal probability of inclusion. The report lists many benefits for the random selection but concludes that "it is subject to great difficulties of carrying it out in practice".

Purposive selection was defined as a method of selecting **a number of groups of units** in such a manner that the selected groups together yield as close as possible the same averages or proportions as are found in the population with respect to those characteristics that are already "a matter of statistical knowledge" (see Chapter 8).

The Representative Method, as defined in Jensen's (Jensen 1926) report, was composed of these two methods: random and purposive selections. However, neither of them was exactly the same as Kiaer's method anymore. On the use of the Representative Method in practice, Jensen (ibid.) listed three methods[83]: random selection of units, usually by systematic sampling; random selection of groups, which in modern terms would be cluster sampling; and purposive selection of groups. Jensen placed Kiaer's method in the last category, a purposive selection of groups, because in his mind the sample consisted of all the dwellings in certain purposively selected streets of a town. Obviously, this was a mistake, though. In Kiaer's method, all dwellings were enumerated only for a few of the smallest streets. In other streets, the sampling fraction varied between ½ and 1/10, and the selection was "mechanical", i.e., systematic. Interestingly enough, Jensen considered the sampling method which Bowley applied in the survey in Reading as random (see next chapter), although in practice, Kiaer's and Bowley's methods were close to each other.

In the report to the ISI, Jensen (1926) argued that Kiaer's influence was strongest in Norway[84], and to some degree, it had spread to the neighbours of Norway, Denmark, and Sweden. Outside Scandinavia, the method had been utilized, especially in Germany and England. It raised the least interest in the Latin countries. Jensen concluded, "... this despite the fact that France is no doubt the country in which the representative method, in an undeveloped form, was earliest applied". Jensen gathered examples from 15 different countries altogether, and concluded that "isolated examples of the application of the representative method may be found everywhere where statistics have arrived at any particular stage of development in a methodical sense." (Jensen 1926)

According to Jensen's account, the domains of statistical research where the representative method had been markedly applied showed an almost equal distribution between the three great groups: 1° Demographic Statistics, 2° Agricultural Statistics, and 3° Social Statistics. Outside of these three groups, Jensen

[83] This was annexed to the report; it was written by Adolph Jensen alone, not by the commission.

[84] Here Jensen was not very precise. Because of the hard critic Representative Method was not applied in Norway since the beginning of 1900s (see Lie 2002).

had found comparatively few representative investigations, but he had found examples of representative investigations on "the important domains which lie on the boundary between economics and social policy". (Jensen 1926)

7.4 Discussion

Kiaer was ahead of his time. He had statistical, or scientific, insight about the value and reliability of partial investigations, and he realized the principles by which a sample should be selected. Kiaer emphasized that there are two important conditions of a successful sampling investigation: proper representation and rational selection of units.

In modern terms, proper presentation can be obtained by a well-designed stratification. Kiaer's stratification factors were geographic, social, and economic. He also introduced proportional selection of units within each stratum, based on the population information of the latest census. At the end of the 19th century, there was no theory for stratification, and Kiaer had to design it intuitively using common sense reasoning aiming at a proper representation of population characteristics. The actual definition of strata was rational or purposive, as it still is in modern sampling practice.

In modern terms, Kiaer's 1895 design can be described as a multi-stage stratified area sample with systematic sampling of households in the final stage in the urban areas. In rural areas, the final stage of data collection had to be organised differently. Houses were selected from routes that Kiaer had defined and enumerators had instructions that were aimed at producing a representative sample.

There were three important principles involved in the accepted approach: The first is the representativeness of the sample. It was vital that information about the population structure was used in the design of the sample. The second principle was that the selection of units for observation should be made objectively, and that enumerators' subjective judgment should not influence the selection (Kiaer 1897a). The third principle was that for every survey, the reliability of the results should be assessed: Each survey should be divided into a number of distinct parts, using for each a different representative method. The comparison of the results of these parts would provide evidence as to how much faith could be placed in the results of the survey (Kiaer 1901). This procedure can be seen as a rudimentary form of a replication method of variance estimation.

To a large extent, Kiaer's method was dictated by the possibilities and facilities that were available at the time. He was aware of the merits of a random selection of units, but at the end of the 19th century, a randomly selected sample from a human population was not possible for two reasons: such sampling frames did not exist where it had been possible to draw a sample from the population of a country; even if it had been possible, collection of data from a random sample had become very difficult and expensive and probably also too time consuming.

There is no indication that random sampling had been considered in the context of estimation, and there is no documentation on attempts to formulate probabilistic statistical inference at the end of the 19th century. Basically, Laplace's estimation of French population and Laplace's method had provided some tools for estimating accuracy. Interestingly enough, a French statistician, Lucien March, made a remark on the use of mathematical methods in the ISI meeting in Berlin (Kiaer 1905, pp. 129-131). He noted a method which had been known for a long time and which involves "partly the science of mathematics and partly the techniques of statistics". He was referring to Laplace's estimation. He also noted that it was possible to determine *"l'erreur á craindre"*, Laplace's version of standard error, provided that observations were obtained randomly from the group that was studied. The report of the meeting continues by saying that "Nobody found this condition necessary in correct application of the method." The topic of the discussion that followed was how the expression "take randomly" should be understood.

It was pioneering work what Kiaer did in developing the Representative Method, even though the method in some form had been already used elsewhere[85]. The most important achievement of Kiaer was to adapt it to social research, and to raise it on the agenda of the ISI meetings, and to defend it persistently despite the criticism.

After Kiaer had raised the issue at several ISI meetings, the method could not be disregarded, and the ISI had to take a stand on the Representative Method. At that time, the International Statistical Institute was a central organ where the officers of statistical institutions and researches from universities met regularly to present results and to learn about new methods. The ISI did not have an explicit role, but as a leading forum of scientific debate, its resolutions had a significant impact on national statistical institutes.

Kiaer's efforts were rewarded first at the Berlin session of the ISI in 1903, when the ISI adopted a resolution which recommended the use of the Representative Method, subject to the provision that the conditions under which the selection of the observations was made were completely specified. The sample survey had become an acceptable method of data collection. The final acceptance took place only in 1925 at the Rome meeting of the ISI.

Thinking back, the Representative Method could not be developed much earlier than it was. Its central prerequisite is that there is sufficient information available to design the data collection and to assess whether the obtained sample can be considered representative. In addition, stratification in general, or as Kiaer performed it, is not possible without sufficient knowledge of the population structures. Only after census data had become available was it possible to determine the size of a sample and to design stratification for data collection and instruct enumerators to collect data.

Another reason why the method could not be developed earlier was the inadequate understanding of the structure of human society. Only after Quetelet's and Engel's research had shown that the structures and the laws governing society were stable and regular, a partial investigation became viable and could be generally accepted (see also Porter 1986). One important factor in using survey research is trust in the results obtained by it. If the knowledge of the population under study is insufficient or non-existent, it is difficult to gain reliance on such a method.

[85] Lie (2002) provides evidence that another Norwegian scientist, Jakob Mohn, discovered the representative method.

A new paradigm

An important reason behind the emergence of partial investigations was the growing need for information on various aspects of society and economy. This was partly caused by the enthusiasm for statistics and also because of social movements that pushed to gather more information about the society, especially on the working class and its living conditions (see Westergaard 1932 and Bellhouse 1988). National statistical offices only trusted (decennial) censuses, which were insufficient in providing timely information about social issues. A growing need emerged for a method that would provide information faster and with fewer costs than a census. It was also found important to obtain more detailed information, i.e., to include more and more detailed questions.

The Representative Method was not meant to replace censuses, but to provide another method to obtain more detailed and timelier information about society than what was possible in censuses. Kiaer's Representative Method can be seen as a completely new method, but it also started a new paradigm for statistical data collection (see also Bellhouse 1988). When the method was presented for the first time, some statisticians tried to reject the proposed method altogether and even to prevent further discussion on it (see also Porter 1986). The documented discussions at the ISI meetings to which Kiaer's presentation gave rise to show that it was intellectually violent in the sense Kuhn defined it.

The development that Kiaer initiated eventually resulted in a new branch in statistical science. Even though randomization and probability did not have any role in the method[86], it can be considered as one starting point in the development of modern survey sampling. Inference in Kiaer's method was intuitive without any theoretical analysis. However, Professor A. L. Bowley[87] of England became at an early stage attracted to Kiaer's method, and began to apply a kindred method in social surveys in England. Even more importantly, Bowley started to develop a method of statistical inference for sample surveys that was based on probability. The result was published in the annex Bowley wrote to Jensen's report for the ISI (Jensen 1926), entitled "*Measurement of the Precision Attained in Sampling*" (Bowley 1926). In that, Bowley described a method on how to estimate the accuracy of estimates (see next chapter). That method was based on the Laplace–Bayes paradigm.

[86] Kiaer was aware of the random selection of units and the virtues of the method, but because of practical reasons, random selection was not an option at that time.

[87] Bowley was accepted as a member of the ISI in 1903.

8 Arthur Bowley and statistical inference for finite populations

8.1 Introduction

In the beginning of the 20th century, statisticians in England were mainly interested in biological (biometrical) topics and eugenics, and they were mainly affiliated with universities and academic research. The continental school showed more interest in official statistics, including social statistics. It was mainly composed of statisticians working at national statistical institutes, or they were academics closely related to the official statistics and statistical institutions.

Arthur Bowley[88] was an exception in the English statistical school because his main interest was first in economics, and from around 1910, Bowley started to take an interest also in social statistics. Bowley undertook several studies on British economical statistics, on trade, and on wages and incomes. Bellhouse (1988) argues that Bowley should be considered as a descendent of the British Statistical Movement of the 19th century. In its aim, Bowley's work can be seen as a continuation of surveys of social conditions, such as Charles Booth's "*Life and Labour of the People in London*" (1889-1903) and Rowntree's "*Poverty, A Study of Town Life*" (1901).

The *New Survey of London Life and Labour* (Llewellyn-Smith 1929) has been held the most significant of Bowley's social surveys, even though he acted mainly as an advisor. Later, Bowley conducted, together with R.G.D. Allen, one of the first empirical studies concerning consumption behaviour (Bowley et. al. 1935). Noteworthy in all these surveys were the methodological innovations: the use (and reporting) of sampling techniques and the careful undertaking of data collection.

In his statistical thinking, Bowley drew from the ideas and works of Quetelet and Lexis, but mathematically, many of Bowley's ideas were derived from

[88] **Arthur Lyon Bowley** (1869–1957) was an English statistician and economist who worked mainly on economic statistics. After school, he received a scholarship to Trinity College, Cambridge, to study mathematics. After leaving Cambridge, Bowley taught mathematics for a short while. Meanwhile, he was publishing in economic statistics; his first article appeared in 1895, the same year the London School of Economics (LSE) opened. Bowley was appointed as a part-time lecturer at LSE, and he was connected with the School until he retired in 1936. He has been considered one of the School's intellectual fathers. At LSE, he became Reader in 1908 and Professor in 1915. In 1919, he was appointed to the newly established Chair of Statistics.

In statistical theory, Bowley was no innovator, but he drew on the writings of Karl Pearson, Udny Yule, and most importantly, F. Y. Edgeworth. Edgeworth was Bowley's first teacher in probability theory, and their collaboration, starting at the end of 19th century, lasted up to Edgeworth's death.

Bowley received many honours. In 1922, he became Fellow of the British Academy, and in 1950, he was knighted. He served on the council of the Royal Economic Society and was president of the Econometric Society in 1938-9. The Royal Statistical Society awarded him its Guy Medal in Silver already in 1895 and the Guy Medal in Gold in 1935, and he served as its president in 1938-40 and was vice-president in 1907–1909 and 1912–1914.

Edgeworth's contributions[89]. Bowley himself had only a few new contributions directly on probability theory, but his achievements in teaching, application, and promotion of sample surveys are remarkable. He summarized his early statistical lectures in two books: *Elements of Statistics*[90] in 1901 and *Elementary manual of Statistics* in 1910. Both books are non-mathematical in nature when compared to modern statistical textbooks. The former was the first English textbook on statistics, and it became widely known and cited all over Europe and North America. Also Russian statisticians frequently referred to it (see Kaufman 1913 and Kovalevsky 1924). However, from the perspective of modern statistical science, Bowley's most important contribution was the detailed analysis of statistical inference for sample surveys (62 pages), published in 1926 by the ISI. It is the first analytical study on statistical inference for fixed populations since Laplace. This monograph was also the first treatment of sampling theory[91], but it was written within the Laplace–Bayes paradigm that was soon replaced by a new paradigm for statistical inference (see Chapter 10). Obviously, this is one reason why it has never attracted wider attention.

8.2 Bowley's presidential address in 1906

In his early 30's, Bowley followed Edgeworth on the chair of the British Economic Society. His presidential address in 1906 is interesting, not only from the perspective of survey research and statistical inference, but also as to its general statistical and political message. At the time when the address was given, surveys or partial investigations were rare. The idea of the use of the Representative Method had been raised (by Kiaer), and it was discussed and criticized at several ISI meetings[92]. Some surveys following the example of Kiaer had already been conducted in Europe (see Jensen 1926). The motive of the address seems to have been Bowley's concerns about the state of social statistics and social research in England.

The beginning of the address was devoted to contemplating the question: "Have we any guarantee that the public service, whether official or unofficial, will be supplied with a sufficient number of persons who are qualified to handle statistics expertly, to follow rapid mathematical developments which alone can get the full significant records, and to inform the public with reasoned knowledge of the measurable phenomena in life?" As an answer, Bowley gives a fairly pessimistic account of the statistical activities in the UK at that time.

[89] It has been claimed that Edgeworth did not have any students and that Bowley was his only follower.

[90] "Elements of statistics" was very popular textbook and Bowley published several revised editions of it. Fifth edition was published in 1925,

[91] Actually, Kovalevsky published the first analytical text already in 1924 (Kovalevsky 1924) but it was in Russian and it was not known outside Soviet Union.

[92] Bowley was accepted as a member of the ISI in 1903 and took part in the meetings on regular basis.

He continues by urging the need for a mathematical approach to statistics in addition to the "arithmetical statistics"[93].

> "It must be recognized that most statistics are necessarily approximate; and just as in other scientific measurements the quantity is given as correct to so many significant figures, so in statistics the possible and probable limits of error should be estimated, and the false show of so-called mathematical accuracy given up." (Bowley 1906)

By "the false show of mathematical accuracy", Bowley meant the habit to give estimates of economic and social parameters that were not based on observations.

After a thorough account of the development of mathematical statistics from Gauss and Laplace, to Quetelet, and to Edgeworth and Karl Pearson, Bowley takes up the application of probability in partial investigations, saying:

> "... the region to which I am devoting particular attention is that where the theory of probability is invoked ... because this is of the greatest importance and least generally understood."

First he touches on the accuracy of estimates leaning primarily on Edgeworth's recent results about the "probable error" and provides justification for the use of the Normal Distribution.

> "... by applying this method ... we are able to give not only a numerical average, but also a reasonable estimate for the real physical quantity of which the average is a local or temporary instance. ... It is the main weakness of statistical estimates... that no measures of precision is given, and consequently that no determination can be made as to whether observed differences (in wages, in death-rates, in diet, in prices) are accidents of observation or are really significant." (Bowley 1906)

Bowley describes how he had selected a random sample from the Investor's Record (including 3,878 companies) to find an average of interest rates by sampling: First, he had numbered the list of companies consecutively. Next, he continued reading down a column in a table in the Nautical Almanac. He took the last digits in groups of four and included all numbers below 3,878. Lastly, he searched the corresponding entry in the Investor's Records, and wrote down the interest value from the table. This way he collected a sample of 400 items.

Probably this is the first documented application of a sampling frame to select a random sample (see also Stephan 1948 and Hansen et al. 1985). It is noteworthy that already in 1906, Bowley emphasized the importance of random selection of units in sampling[94]. He said:

[93] By "arithmetical statistics" Bowley meant the tabulation and classification of statistical information.

[94] Some years earlier when Bowley published his "results which follow those obtained in connection in the Newmarch Lectures, 1897" (Bowley 1897) he drew attention to the probable error due to omission of part of the data because part of the data is inaccessible. He suggested that accuracy would be expressed by using a quantity which was close to standard error but he did not touch the selection of the units.

> "...It was necessary to make certain in some way as this, that the chances are the same for all the items of the group to be sampled, and that the way they are taken is absolutely independent of their magnitude." (Bowley 1906)

In the same context, he sought to give a simple empirical verification for the validity of the Central Limit Theorem in a context that today would be called simple random sampling: He tabulated the means of 40 samples (from the same "universe") and observed that the distribution of the 40 sample means was approximately bell-shaped, i.e., normal. Bowley also showed that the accuracy of estimates does not depend on the size of the population, but only on "its nature" and on the size of the sample. He continues by showing that the accuracy can be increased – and the probable error decreased – by increasing the size of the sample.

Bowley emphasized the importance of a frame and the problems that its absence might introduce to social research and to sample surveys. Therefore, at the end of his presidential address, he made a plea to establish a household or population registry in the UK, which could be used as a sampling frame in social research.

> "To learn the actual economic condition of all the 40 000 000 persons of the United Kingdom, or even of those who are not obviously above any poverty line, seems at first sight an impossible task; and so indeed it is, but only because of the general apathy to the subject. We must, therefore, proceed by some method of samples. Before we can get sound information from samples we must have a method of numbering or classification by persons or by districts. If we had a definite system of registration and identification, as in Germany, it would be easy to choose, say 1 in 100 or 1 in 1000 at random from among all the persons whose records satisfied certain conditions, and then to investigate more carefully the history and circumstances of those chosen. A similar method could be applied to any particular district. There is no need to make a house-to-house visitation to learn the conditions of a district; it is sufficient to enumerate the houses, to choose a certain proportion at random, and investigate carefully the status of their inhabitants." (Bowley 1906)

The speech was fairly long and covered a variety of topics. Bowley concluded the part about sampling by emphasizing the importance of probability theory and the random selection of units:

> "The method of sampling is not only one of many instances of the application of the theory of probability to statistics. I have taken it at length because the method is so persistently neglected, and even when it is used the test of precision is ignored. We are thus throwing aside a very powerful weapon of research. It is frequently impossible to cover whole area, as the census does, ... , but it is not necessary. We can obtain as good results as we please by sampling, and very often quite small samples are enough; the only difficulty is to ensure that every person or thing has the same chance of inclusion in the investigation." (Bowley 1906)

It is noteworthy how well Bowley was aware of the central questions of survey research already in 1906 when surveys were very rare and sampling was not recognised as a part of statistical theory.

8.3 First social surveys in England

An important characteristic in Bowley's work was his interest in a practical statistics. He carried out several social surveys in England, mostly dealing with living conditions, poverty, and wages, notwithstanding the fact that the feasibility of partial investigations was questioned. An example of the distrust is the discussion after Bowley's lecture in 1908 before the Royal Statistical Society entitled "The Improvement of Official Statistics". Among the many topics, he spoke again about "the scientifically chosen samples". In the debate that followed, Yule said that he was not convinced that "a sampling can be truly random and representative of the total population".

In a survey concerning the living conditions of working-class households in Reading, the sample selection method resembled that of Kiaer's. Random sampling, as Bowley described it in his presidential address, was not an option in a large-scale social survey: Adequate sampling frames were not available in England, and probably also the random selection itself would have presented problems. In addition, the arrangement of the fieldwork for a large sample had been a difficult task and expensive if simple random sampling had been used.

8.3.1 Survey in Reading

The report of the survey which Bowley carried out in Reading in 1912 is a significant contribution for survey sampling itself and for the use of survey methods in social research. It is an example of Bowley's deep interest in social problems, but an equally important incentive seems to have been the promotion of the research method. This article was not the first of this kind, though. Yule had published a similar article 15 years earlier on the history of pauperism in England (Yule 1896), which had a similar structure as Bowley's in the sense that a great part of it was devoted to the presentation of the methods and their application to social research. However, Yule's point was not sampling but fitting distributions to data, which was the main method of statistical analysis at the end of the 19th century.

The beginning of the Bowley's report reveals his motives:

> "An investigation was made in Reading, in the autumn of 1912, into the general economic conditions of the working class, by a small unofficial committee. The results are of much more than local interest, since they prove that an inquiry adequate for many purposes can he made rapidly and inexpensively by a proper method of samples. In particular, by classifying earners in relation to dependants, we show the relative numbers of men and of women who have to support families of various sizes (data which hardly exist on any large scale); and further a comparison is afforded in essentials with Mr. Rowntree's well-known study of York[95].

[95] Bowley referred to Rowntree's study on "the extent and depth" of poverty amongst the wage-earning families in York (Rowntree 1901). Rowntree had become to conclusion that "all questions to be answered with any fullness and accuracy, nothing short of a house-to-house inquiry extending to the whole of the working class population would suffice." Consequently he tried to obtain information regarding the housing, occupation, and earnings of every wage-earning family in York, together with the number and age of the children in each family. Rowntree's survey actually was not the first of the kind in England. Charles Booth had carried out similar survey more than ten years earlier in London (Booth 1889),

> It is already arranged that a similar inquiry shall be made in another town, and it is hoped that, when the simplicity of the method and the importance of the results are appreciated, a sufficient number of people will be interested to carry out investigations in other towns and in rural districts, till we have general knowledge of the economic conditions of the households of Great Britain. It is not generally realised that the only information we have at present is that given by the Census as to the number of persons and number of rooms." (Bowley 1913)

Rowntree's "study" was a comprehensive survey into the living conditions of the poor in York. Enumerators visited every working class home, which amounted to data on 11,560 families or 46,754 individuals (Rowntree 1901). While Rowntree carried out a total enumeration, Bowley decided to use a sample of households. Bowley describes the method how the sample was selected and which factors influenced the decisions in the following manner:

> "A sample was selected from the whole of the present borough of Reading as follows: One building in ten was marked throughout the local directory in alphabetical order of streets, making about 1,950 in all. Of these about 300 were marked as shops, factories, &c., institutions and non-residential buildings, and about 300 were found to be indexed among Principal Residents, and were so marked. The remaining 1,350 were working-class houses, and a number of volunteers set out to visit every one of these. It was presently found that the scale taken was beyond their powers, and it was decided to take only one house in 20, rejecting the incomplete information as to the intermediate tenths. The visitors were instructed never to substitute another house for that marked (unless the house was unoccupied, in which case the next door was to be taken), however difficult it proved to get information, or whatever the type of house. In the end we failed to learn anything as to 32 households out of 677, and substituted for these 32 of the surplus tenths, without, so far as can be judged, introducing any bias. Information was entered on cards by the visitors, and a great deal of supplementary description was written on the back of the cards." (Bowley 1913)

The described method is close to the one which Kiaer had applied in Norway when sampling houses in the cities, except that Bowley did not use stratification as Kiaer did. In modern terms, Bowley's method was based on systematic sampling (every tenth house in an alphabetical list of streets). Bowley regarded the obtained sample as random because "it did not involve any purposive elements".

After obtaining the data, Bowley compared the sample to the latest Population Census (of 1911), in the same way as Kiaer had done. Obviously, his aim was to convince the readers of the representative nature of the sample, although he did not say so explicitly.

Inference from the sample to the population was intuitive and straightforward because the sample was supposed to be a miniature of the population. Therefore Bowley could write:

> "At the date of the investigation there were probably about 18,000 inhabited dwelling-houses, of which our first table deals with 840, that is 1 in 21. The multiplier twenty-one is then to be applied to all the sample data to give estimates for the whole of Reading." (Bowley 1913)

Then he continues with giving guidelines for a more accurate interpretation of the results:

> "It may appear to persons who are not familiar with processes of sampling that a proportion of 1 in 21 is too small for any conclusion, and that in any case not more than a vague probability can be obtained. The theory and method of sampling is discussed in [reference to his presidential address in 1906]. It is there shown that the precision of a sample depends not on its proportion to the whole, but on its own magnitude, if the conditions of random sampling are secured, as it is believed they have been in this inquiry. It is demonstrated mathematically that if in our sample 622 working-class households we find respectively 5, 10, 20, 40, 50 per cent. of cases, we may expect that the percentage in the whole are within 5±1, 10±1, 20±1½, 40±2, 50±2 and may be nearly certain that they are within 5±3,10±4, 20±5, 40±6, 50±6." (Bowley 1913)

In the footnote, Bowley still specified that

> "Here the standard deviation is used; the change is about 2 to 1 in favour of the true being within the limits for the first set, and 1 to 250 for the second set." (Bowley 1913)

Bowley also reported probabilistic interval estimates, which are conceptually close to confidence intervals. Bowley calculated the limiting values by using the Central Limit Theorem. Bowley devoted a considerable part of the report to explaining the principle, although the normal approximation in large samples was already well established in textbooks on statistics (e.g., Bowley 1910 and Yule 1911).

8.3.2 Other surveys in England

A consequence of the survey in Reading was that it set off a boom of several similar surveys in the UK (as Bowley obviously had hoped for). Bowley's associates carried out surveys for three other cities in 1913 and two the next year. The survey in Reading was soon followed also by a systematic sampling of census schedules in 1915 (Bowley and Burnett-Hurst 1915). Ford conducted a survey in Southampton in 1927, which was similar to Bowley's first survey, both methodologically and contextually (Ford 1934). And the London School of Economics under Bowley's direction carried out a similar survey London in 1929 (Bowley 1929). In that survey, the "House Sample" involved stratification in two ways. First, the method of selecting households within a given administrative area was based on systematic sampling by streets arranged in alphabetical order. Secondly, for the whole London area, the stratification was done by administrative area, in which the sampling ratios were constant. A few years later, Caradog-Jones carried out two similar surveys, one in Liverpool in 1930 and one in Merseyside in 1931 (Caradag-Jones 1931 and 1934).

The reliance on partial investigations was not high in the first quarter of the 20th century, except amongst only a few statisticians. Examples of the difficulties are shown in two papers by John Hilton, director of statistics at the Ministry of Labour in the UK, concerning the studies about workers in the unemployment insurance system (Hilton 1924 and 1928):

The problem was that the Ministry of Labour needed information about the more than one million "persons who were being returned week by week as 'insured workpeople unemployed' ". In the first effort, the ministry had ordered the selection of every third claim of Unemployment Benefits to be investigated.

Interviews of the claimants were used to enhance the previously registered data. Informants were selected systematically from the files of the Labour Exchanges

However, processing that amount of data (372 875 persons) in tables had required too much labour and time in order to serve the purpose of the survey. Therefore, the administration asked Bowley's advice for a better design. Bowley suggested systematic sampling with a sampling fraction of 1 to 1000, but the office considered that it would not suffice to carry out as detailed analysis as required. After some consideration, it was decided to use 1% as the sampling fraction.

The first of Hilton's two papers (Hilton 1924) comprehensively treated both the problems in selection of claimants and the difficulties of arranging the interviews. Eventually, a design close to quota sampling was applied. Methodologically, the most important information, however, was obtained by comparing the distributions of certain background variables of the 'one percent' sample to previous distributions: distributions were close to each other.

Certain deviations from a strictly systematic selection introduced biases, which were reduced by subsequent improvements of the method, which were described in the second paper (Hilton 1928). Hilton found a sample of only one per cent quite satisfactory to meet the practical administrative and policy-making purposes for which the studies were made. The reduction in expenses that was achieved by sampling such a small proportion of the records was impressive. Despite all the benefits which the method yielded, it was not imitated by other government bureaus, although the results Hilton showed were very favourable (see also Stephan 1948).

Hilton read the paper before the Royal Statistical Society. The invited discussants were Bowley, Edgeworth, Yule, and Greenwood, who all found Hilton's method to be close to Kiaer's and encouraged Hilton to further develop the method he had applied.

Bowley's work had also an impact on the development of sample surveys in the United States (see also Jessen 1942 and Stephan 1948). Margaret Hogg, who had worked under Bowley's direction on some of the British surveys, moved to the United States to work for the Russell Sage Foundation. She made a critical study of employment and unemployment statistics in the United States. In an article published in JASA, she made a plea for rigorous methods of sampling and also cast some doubt on the value of surveys that had been made in the U.S., in which the sample was selected by judgment rather than random procedures (Hogg 1930). In 1931, Hogg conducted a survey of unemployment in New Haven, partly to test the practical difficulties of applying a random sampling method, and also to develop better schedules and statistical categories for unemployment surveys (Hogg 1932). Hogg's contributions had a strong influence in the development of survey methods in the U.S. (see Chapter 12).

In addition to carrying out sample surveys and promoting the sampling method, Bowley also laid the foundations for the mathematical approach to sampling theory. Based largely on Edgeworth's contributions, he elaborated and summarized his ideas in the report he wrote to the ISI, which was published as an appendix to the report of the committee evaluating Kiaer's Representative Method (Bowley 1926).

Bowley often took part in the ISI meetings and thus became aware early on about the discussion concerning the partial investigations and Kiaer's Representative Method. Bowley quickly saw the potential that the Representative Method provided in shedding light on living conditions of the working class (Bowley had left wing sympathies), and he started to develop the method into a statistically more valid form. Since Laplace, nobody before Bowley had tried to apply the Laws of Errors on a (randomly selected) sample from a finite population. Edgeworth and Yule dealt with sampling "fluctuation", but they treated only infinite populations, or if they sampled the real population, the method of sampling was obscured (see Edgeworth 1906, 1907, 1908, 1909 and Yule 1911).

Edgeworth, Yule, Pearson, Tchuprov[96], and a few others had already published most of the related mathematical theory. Bowley summarized this and his own contribution in a report to the ISI, concluding "... As far as I can ascertain, no one has brought together these formulae so as to give ..." (Bowley 1926).

Already in the early 1920s, Bowley wrote an article entitled *The precision of measurements estimated from a sample* (Bowley 1923). It was published in Metron in 1923 already before the ISI session in which the committee was nominated "to study the applications of the Representative Method in statistics". The article treated the "inverse problems in statistics", drawing from the method given by Edgeworth in 1908. It was partly motivated by the article that Pearson had written a few years earlier on a similar problem (Pearson 1920), but Bowley's approach was different. In the introduction to the article, he wrote

> "I have not considered here the problem 'if in a sequence *m* things have and *n* things have not a certain attribute then what is the chance that in the future sequence of $r + s$ events *r* shall have this attribute' for that involves logical questions and further definitions of some complexity; but I have had in view simply the ordinary and practical problem of the precision with which the characteristics of a group of considerable size can be ascertained by examination of a sample chosen from it." (Bowley 1923)

Bowley referred to the principle of learning from experience, which was the subject of Pearson's article, but he does not describe what the "logical questions" and "complex further definition" are, which he wanted to avoid. This citation indicates that Bowley aimed at addressing the practical problems of survey research for which the principle of learning from experience is not suitable.

8.4 The precision of measurements estimated from samples

The memorandum to the ISI was based on the 1923 paper, but Bowley elaborated considerably his approach from the original. In addition, the memorandum contained an extensive analysis on general survey methodology. The mathematical, or inferential, part was based on Laplace's principle of the inverse probability method – or it was based on the Laplace–Bayes paradigm. The origin of Bowley's

[96] Bowley refers to Tchuprov's papers published in Biometrika in 1918 and 1920.

inference model is shown, for example, in how he describes sampling without replacement as "not replacing the balls in the urn". The idea of balls in an urn indicates that he had the urn model in his mind, and he derives his formulas implicitly using Bernoulli trials as his inference model. Since Laplace, Bowley's two papers seem to be the first systematic treatments on the accuracy of estimates obtained from a sample (see also Hald 1998). Also, Bowley himself argues that his contribution is the first that "brings together the old formulas in simple and stratified sampling, in a slightly enhanced form" (Bowley 1926). It may also be the last text on sampling theory within the Laplace–Bayes paradigm.

The mathematical part of the memorandum is divided into two sections: the first one deals with estimation under random sampling, and the second with estimation under purposive selection. In estimation under random sampling, Bowley analyzed both estimation under "unrestricted sampling" and "restricted" sampling. In modern terminology, unrestricted sampling means simple random sampling, and restricted sampling refers to stratified sampling. In addition, Bowley treated separately sampling for three different parameters: the prevalence of one attribute; distribution of alternative attributes; and sampling for "the magnitude of an average". Bowley's paper is very extensive and only a brief description is given here.

8.4.1 Inference under random sampling

Bowley's approach was comprised of two parts: first he described the direct problem (what kinds of distributions the random selection from a known population can produce); the second and more important part dealt with the inverse problems (which are the characteristics of the population that may have yielded the obtained results). The most important distinction to his previous report (Bowley 1923) is that in the memorandum to the ISI, Bowley analysed sampling from **finite** populations (of size N).

The approach to the problem follows the lines of thought that were already present in Laplace's works, and he acknowledges it: "So far as the direct problem is concerned, the expression of E_x [see formula (8.1)] in the case of purely random sampling from an infinite universe, including the unsymmetrical term that involves $1/\sqrt{n}$, have been known since the time of Laplace, Gauss, Bernoulli and Poisson." In deriving estimates, Bowley applied the same mathematical methods that Laplace had used, i.e., applying Stirling's formula to elaborate factorials, Taylor's expansion, and ignoring terms that became negligible in large samples. He also applied the Method of Moments[97], which Edgeworth and Karl Pearson applied frequently [98].

Before touching on the main topic, Bowley makes a puzzling comment "... the problem before us is to make inferences from a given sample to an unknown universe ... and we are therefore obliged to go on to the doubtful ground

[97] The Method of Moments is an old method for point estimation of population parameters by equating sample moments with unobservable population moments and then solving those equations for the quantities to be estimated.

[98] In practice, Bowley embraced many of the ideas and methods of Edgeworth and Karl Pearson who both considered that the inverse probability is the basis of (then) modern statistical theory (see e.g. Pearson 1920)

of inverse probability." (Bowley 1926). However, he does not explain what he means by the "doubtful ground of inverse probability". He may have referred to R.A. Fishers' paper published in 1922, in which Fisher implicitly attacked the method of inverse probability.

8.4.1.1 Estimation under simple random sampling

Bowley started with the simplest case: estimation of the frequency of one attribute by random sampling. By random sampling, Bowley meant a selection in which all population units have equal inclusion probabilities. He stated the problem in the following manner: "Given that in a sample on *n* persons or things, drawn at random from a universe containing *N*, *pn* possess a certain attribute, what can we infer about the prevalence of the attribute?" The novelty here was that in the earlier 1923 paper, Bowley treated infinite populations.

The solution to the problem was given in two parts: "in one, the chances[99] that the sample would be drawn from hypothetical universes are compared" (the direct problem); "in the other, it is considered under what circumstances it is possible to make any inference about the relative chances that in fact the universe containing given proportions" (the inverse problem). Bowley adds that the second part involves the theory of inverse probability. In this, Bowley leaned on Egdeworth's version of the Central Limit Theorem (Edgeworth 1908 and 1909).

8.4.1.1.1 Sampling for prevalence of one attribute

The direct problem for one attribute case was presented as follows: In a population containing *N* persons or things *PN* have a certain attribute. From the population whose members have been "numbered or otherwise indexed", *n* persons are selected at random. In the direct problem, it is required to find the probability that the number of sampling units in the sample which possess the attribute is *pn*.

Below are a few examples on how Bowley solves the problem, to give an idea about his techniques:

There are ${}_NC_n$ equally probable combinations of *n* units chosen out of *N* units, with ${}_NC_n = \frac{N!}{n!(N-n)!} = \binom{N}{n}$.

The proportions in the population and sample are denoted $Q = 1 - P$, $q = 1 - p$, respectively. If the bias in estimate due to sampling is denoted by *x*, then the number of sampling units with and without the attribute are $pn = Pn + x$ and $qn = Qn - x$, respectively.

Bowley sought to find the probability E_x:

$$E_x = {}_{PN}C_{pn} \times {}_{QN}C_{qn} / {}_NC_n = {}_{PN}C_{Pn+x} \times {}_{QN}C_{Qn-x} / {}_NC_n \qquad (8.1)$$

That is, Bowley analyses hypergeometric probability, which is a consequence of the assumption that the population is finite and the sample is drawn without

[99] It should be noted that Bowley used the concept of "chance", which in modern terminology is synonym to probability.

replacement. Bowley concluded that if *Pn* is so large that it is possible to neglect *1/Pn* in comparison to unity, then approximately

$$E_x = \frac{1}{\sigma\sqrt{(2\pi)}} e^{-\frac{x^2}{2\sigma^2}\left\{1-\frac{Q-P}{2\sigma}\left(1-\frac{2n}{N}\right)\left(\frac{x}{\sigma}-\frac{x^3}{3\sigma^3}\right)\right\}} \tag{8.2}$$

where $\sigma^2 = PQn\left(1-\frac{n}{N}\right)$

The (direct) probability that p should not differ from P by more than z ($=x/n$) is approximately

$$\int_{-z}^{z} \frac{1}{\sqrt{\left\{2\pi PQ\left(\frac{1}{n}-\frac{1}{N}\right)\right\}}} e^{-\frac{z^2}{2PQ\left(\frac{1}{n}-\frac{1}{N}\right)}} dz \tag{8.3}$$

The inverse problem in one attribute casė is following: Given that in a random sample of n units from a population of N units, *pn* units possess the attribute. What can be inferred about the prevalence of the attribute in the population?

To solve the first part, Bowley modified the preceding formulas so that they depend only on the observed p (instead of the population parameter P). He showed that under simple random sampling, the expectation that $pn + x$ would be found from a population in which the proportion is P would be, if $1/pn$ is negligible

$$E_x = \frac{1}{s\sqrt{(2\pi)}} e^{-\frac{x^2}{2s^2}\left\{1-\frac{q-p}{6s}\left(2-\frac{n}{N}\right)\frac{x^3}{3s^3}\right\}} \tag{8.4}$$

where $s^2 = pqn\left(1-\frac{n}{N}\right)$ is the sample variance.

If $1/\sqrt{n}$ is negligible, the formula (8.4) reduces to

$$E_x = \frac{1}{s\sqrt{(2\pi)}} e^{-\frac{x^2}{2s^2}} \tag{8.5}$$

Bowley gave several tables showing that this probability falls rapidly as x increases. He also gave another table supposing that the *a priori* probability that P should have certain values is constant over small ranges and adds by integration the probabilities over these ranges.

To be able to continue with the inverse probability model, Bowley incorporated *a priori* probability distribution for parameter P. Unlike Laplace, Bowley did not agree to the assumption of uniform distribution of priors. Instead, he assumed that the *a priori* probability $F(P)$ (that a population should contain a proportion P of units that have the attribute) to be continuous and derivable in the neighbourhood of $P = p$, where p is the proportion observed in the sample.

The 'double chance' that P was the proportion in the population and that then p should be found in the sample is $F(P) \times E_x$. Hence, "the inverse chance" that p being found in the sample and P was the proportion in population is

$$\frac{F(P) \times E_x}{\sum \{F(P) \times E_x\}} \tag{8.6}$$

summation over all possible values of P. Bowley showed that under some fairly general assumptions on $F(P)$, the probability that P does not differ from p on either side by more than x/n is independent of F (see also Hald's (1998) comments on this). This he had already shown in the earlier article (Bowley 1923).

Bowley derived the result leaning on the results that Edgeworth had published earlier (Edgeworth 1908)[100]: If the assumptions hold, the probability that P is within the limits $p \pm z$, with $z = x/n$, is approximately

$$P(p - z \leq P \leq p + z) = \int_{-z}^{z} \frac{1}{\sqrt{\left\{2\pi\, pq\left(\frac{1}{n} - \frac{1}{N}\right)\right\}}} e^{-\frac{z^2}{2pq\left(\frac{1}{n} - \frac{1}{N}\right)}} dz \tag{8.7}$$

In the memorandum, Bowley derived this result by the method of moments but added that the same proof could also be done using Stirling's formula and Taylor's expansions (as Laplace had done).

In 1923, Bowley observed that in sampling from an infinite population, the total probability that the proportion P was within limits $[p - x, p + x]$ is

$$P(p - x \leq P \leq p + x) = \int_{-x}^{x} \frac{\sqrt{n}}{\sqrt{2\pi\, pq}} e^{-\frac{x^2}{2pq/n}} dx \tag{8.8}$$

In modern terms, Bowley's approach can be regarded as based on a superpopulation model: The actual and observable population is not considered as fixed or stable, but as constantly changing. The observable population is one realization of all potential populations and therefore its parameters can not be constants but have an (unknown) random distribution. The *a priori* probability distribution of the population parameter gives the probability that in the population from which the sample was selected, the parameter has value P. He showed that (8.8) holds whatever is the form of *a priori* distribution F, however. The method is like the one Laplace used to derive his estimates. It was a characteristic method in the Laplace–Bayes paradigm.

[100] Edgeworth based his results on the Laplace's Inverse Probability in the form it was published in *Théorie Analytique des Probabilités* (Laplace 1812)

8.4.1.1.2 Estimation of average

The direct problem in the estimation of a mean was the following: From a population of size N, a random sample of size n is selected. The sampling fraction is k = n/N. The population mean of the variable is $\bar{u}$, σ^2 is the population variance, and the observed (sample) mean is $\bar{u} + x$. The standard deviation of x is

$$\sigma_a = \frac{\sigma}{\sqrt{n}}\sqrt{(1-k)} \tag{8.9}$$

When N is large, the term $\sqrt{(1-k)}$ can be disregarded because it will be $\simeq$ 1

.

Bowley showed that the probability that $\bar{u} + x$ will be observed is

$$E_x = \frac{1}{\sigma_a\sqrt{(2\pi)}}e^{-\frac{x^2}{2\sigma_a{}^2}\left\{1-\frac{\kappa_1}{2}\left(\frac{x}{\sigma_a}-\frac{x^3}{3\sigma_a{}^3}\right)\right\}} \tag{8.10}$$

where $\kappa_1 = \frac{1}{\sqrt{n}}\frac{1-2k}{\sqrt{(1-k)}} \times \frac{\mu_3}{\sigma^3}$

and μ_3 is the third moment about the mean. If n is so large that $1/\sqrt{n}$ is negligible, the term that contains κ_1 is also negligible.

If also k is so small, and N large, that "½k is negligible", then σ_a can be written $\sigma/\sqrt{n}$ and the probability then will be given approximately by

$$E_x = \frac{\sqrt{n}}{\sigma\sqrt{(2\pi)}}e^{-\frac{x^2 n}{2\sigma^2}} \tag{8.11}$$

Based on the formula (8.10), Bowley concluded that the probability that the sample mean does not differ from the population mean more than x is

$$p(\bar{u} - x \le \bar{u} \le \bar{u} + x) = \int_{-x}^{x} \frac{\sqrt{n}}{\sigma\sqrt{(2\pi)}}e^{-\frac{x^2 n}{2\sigma^2}}dx \tag{8.12}$$

The solution to *the inverse problem* Bowley derived in the same way as in the previous problem:

In a sample of size n, drawn at random from a population containing N units, the population mean is $\bar{u}$ and the second moment about it is $\mu_2{}'$. $F(\bar{u})$ is the *a priori* probability that the population mean is $\bar{u}$. Bowley did not assume that $F(\bar{u})$ is rectangular but that it is continuous and that its "derived functions" in the neighbourhood of $\bar{u} = \bar{x}$ are finite. $f(\bar{x})$ is the probability that the sample mean will be $\bar{x}$.

He continued the proof by denoting $\bar{x} = \bar{u} + x$.

The probability that, given $\bar{x}$, the population mean is $\bar{u}$

$$Q_x = \frac{F(\bar{x} - x) \cdot E_x}{\int_a^b F(\bar{x} - x) \cdot E_x dx} \tag{8.13}$$

where $a = min(x)$ and $b = max(x)$. The term $1/\sqrt{n}$ is "neglected". Based on the previous paper (Bowley 1923), he concluded that $E_x = \frac{1}{s\sqrt{(2\pi)}} e^{-\frac{x^2}{2s^2}}$ and s is the observed standard deviation.

According to Laplace's principle, the probability that the population mean is in the range $(\bar{x} - x \leq \bar{u} \leq \bar{x} + x)$ will be

$$C_x = P(\bar{x} - x \leq \bar{u} \leq \bar{x} + x) = \int_{-x}^{x} Q_x dx = \frac{\int_{-x}^{x} F(\bar{x} - x) \cdot E_x dx}{\int_{a}^{b} F(\bar{x} - x) \cdot E_x dx} \tag{8.14}$$

Then denoting $\tau = x/s$ and $s^2 = \mu_2\left(\frac{1}{n} - \frac{1}{N}\right)$ where μ_2 is the second moment from the sample, s is the standard deviation.

If $1/\sqrt{n}$ is disregarded, the numerator in 4.14 becomes

$$\int_{-x}^{x} F(\bar{x} - x) \cdot E_x dx = F(\bar{x})\int_{-\tau}^{\tau} \frac{1}{\sqrt{2\pi}} e^{-\frac{\tau^2}{2}} d\tau$$

and the denominator becomes

$$\int_{a}^{b} F(\bar{x} - x) \cdot E_x dx = F(\bar{x})\int_{a}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{\tau^2}{2}} d\tau = F(\bar{x})$$

because "the total probability of τ is unity"[101]. The a priori probabilities, $F(x)$, cancel each other out, and finally

$$C_x = P(\bar{x} - x \leq \bar{u} \leq \bar{x} + x) = \int_{-x}^{x} \frac{1}{s\sqrt{(2\pi)}} e^{-\frac{x^2}{2s^2}} dx \tag{8.16}$$

Lastly, Bowley gave an example of estimation of the average size of working-class households in Northampton (see also Bowley 1915). The sample was composed of 693 households resulting in sampling fraction $k = 1/22.7$. The observed average $\bar{x}$ was 4.342 persons and the standard deviation s was .073. Using the preceding formulas, Bowley calculated that with probability .0027, the average of the number of persons in the working-class houses was outside $\bar{x} \pm 3s$, i.e., above 4.562 or below 4.122. Bowley also gave examples of other interval estimates with different 'confidence' probabilities. In addition, he compared these results to the census of 1911 for the whole region where the average was 4.44.

[101] Strictly speaking, the total probability of τ is unity only in case the integration is from $-\infty$ to $+\infty$. In practice, the total mass of variable τ equals unity because Bowley assumed that the lower limit is *min(x)* and upper limit *max(x)*.

8.4.1.2 Stratification

In the report to the ISI, Bowley also treated stratification in sampling, but not in such a detailed manner as estimation under random sampling. By stratification, he meant a method where an equal proportion of units are selected at random from each stratum. In modern terms, the method is called proportional stratification. In some older textbooks, it was also called Bowley stratification, but that name has disappeared from modern literature. Bowley shows that in every case, the accuracy of estimation increases by stratification, and in some cases, the improvement is considerable.

Interestingly enough, Bowley did not consider stratified random sampling as truly random sampling. The only truly random sampling was defined to be a scheme in which every unit in "the universe has the same chance of being selected in the sample".

8.4.2 Estimation under purposive selection

Basically, Bowley did not regard random and purposive selections as too different methods, and he accepted both. However, his analysis concerning random selection is more detailed and also more advanced than that of purposive selection. In the discussion on Neyman's paper for the Royal Statistical Society in 1934 (Neyman 1934), Bowley said that he did not give equal importance to the methods and continued that purposive selection is very difficult to formulate, difficult to carry out, and that it is difficult to get a good estimate of the precision of the result, except in rather unusual cases.

Bowley begins his analysis of purposive selection in the report with the following introduction:

> "The problems presented by purposive selection differ in emphasis, rather than in kind, from those already discussed when the selection is random. In both methods we are concerned with the proportion, with the average, or with the distribution of some quantity or attribute. In both methods the two fundamental factors in the measurement of the precision of the observations are the dispersion from their mean of the proportions or averages through the universe under consideration, and the number of entries (in random selection the number of individuals, in purposive selection that of districts) that are included in the sample. In each method the precision may be increased by stratification." (Bowley 1926)

The essential difference between random and purposive selections (as Bowley defined it) is that in purposive selection, the sampling unit is an aggregate, such as a whole district, and the sample is "an aggregate of these aggregates", while in a random selection, the sampling unit is a person or a thing. Consequently the analysis is based on weighted averages instead of unweighted averages.

> "Further the fact that the selection is purposive very generally involves intentional dependence on correlation, the correlation between the quantity sought and one or more known quantities. Consequently the most important additional investigation in this section relates to the question how far the precision of the measurement is increased by correlation, and how best an enquiry can be arranged to maximize the precision." (Bowley 1926)

Bowley's definition of purposive selection is different from how it is defined nowadays. His approach has to be seen from the background of the availability of statistical data at that time. In the presidential address nearly 20 years earlier, he had urged the establishment of a population registry in the UK, but it had not taken place. Obviously, only aggregate level statistics were available to be used for the calculations of accuracy of estimates.

8.4.2.1 Notation[102]

Bowley started assuming that the population, or "universe", under investigation consists of N districts[103]. The s^{th} district in the population consists of a_s units, and the population consists of a total of A units, so that

$$A = \sum_{s=1}^{N} a_s \tag{8.17}$$

The aim of the survey is to find P, the proportion of the units in A having the attribute of interest or X, the average of some variable that every unit has. If p_s is the proportion and x_s is the average in the s^{th} district, the corresponding values in the population are

$$AP = \sum_{s=1}^{N} a_s p_s;\; AX = \sum_{s=1}^{N} a_s x_s \tag{8.18}$$

Bowley says that the N values of the p's, or the x's, can be regarded as "frequency groups", whose unweighted means are $\bar{p}$ or $\bar{x}$ and standard deviations $\sigma_p$or σ_x, respectively. In the following section, the analysis will be illustrated only for the variable X. In the memorandum, Bowley also included an analysis for the proportion, P, but it is skipped here because the derivation is analogous with that of the variable.

8.4.2.2 Estimation of an average

Bowley assumes that there are one or more associated variables, whose values are known in every district. In the s^{th} district, the values of these "controls" are written $y_{1s}, y_{2s}, y_{3s} \dots y_{ts}$, and their corresponding population values are, Y_1, Y_2, $Y_3 \dots Y_t$, so that

$$AY_1 = \sum_{s=1}^{N} a_s y_{1s},\; AY_2 = \sum_{s=1}^{N} a_s y_{2s},\; AY_3 = \sum_{s=1}^{N} a_s y_{3s} \cdots \tag{8.19}$$

Bowley regarded the N values of y_i, $i=1,\dots, t$, as 'frequency groups' whose unweighted averages are $\bar{y}_i, i = 1,\dots,t$, and standard deviations $\sigma_{y_i}, i = 1,\dots,t$,

The correlation coefficients between x and y_i, $i = 1,\dots, t$, are denoted by r_{xi}, $i = 1, \dots t$, and correlations between y_i and y_j, $i, j = 1,\dots , t$ are $\rho_{ij}, i,j = 1,\dots,t$.

[102] The notation in this chapter is slightly changed from that of Bowley's, to make it easier for a modern reader.

[103] In modern statistical language the district would be nearly the same as a cluster.

In addition, Bowley assumed that the partial regression equation between x with y_i, $i=1,\ldots, t$, is linear, so that it can be written in the form

$$(x - \bar{x}) = \sum_{i=1}^{t} \beta_i (y_i - \bar{y}_i) \tag{8.20}$$

where the values of the regression coefficients (in case of one control) are $\beta_i = r_{xi} \dfrac{\sigma_x}{\sigma_{y_i}}$.

For district s, e_s is the error resulting from calculating x_s from the regression equation, i.e.,

$$e_s = (x_s - \bar{x}) - \sum_{i=1}^{t} \beta_i (y_i - \bar{y}_i) \tag{8.21}$$

For the purposive selection, Bowley assumed that the number of districts, n, is selected in such a way that the average for each control variable 'is the same in the aggregate of them as it is in the universe', that is

$$Y_i \sum_{s=1}^{n} a_s = \sum_{s=1}^{n} a_s y_{is}, i = 1,\ldots,t \tag{8.22}$$

This requirement was essential to Bowley's definition of purposive selection. It involves the assumption that if the averages match, then the selected districts compose a representative sample from the population.

The value of the unknown parameter computed from the selected districts (if X is the true value) is denoted by X_n, so that

$$X_n = \frac{\sum_1^n a_s x_s}{\sum_1^n a_s} \tag{8.23}$$

The problem is to find the accuracy of the estimate X_n, or to estimate its sampling error.

From the previous results (8.23 and 8.21) and the selection of the districts, it follows

$$X_n = \frac{1}{\sum_1^n a_s} \cdot \sum_1^n a_s \left[e_s + \bar{x} + \sum_1^t \beta_i (y_{is} - \bar{y}_i) \right] \tag{8.24}$$

Applying the definition of the purposive selection given in (8.22), the formula (8.24) reduces to

$$X_n = \frac{\sum_1^n a_s e_s}{\sum_1^n a_s} + \bar{x} + \sum_1^t \beta_i (Y_i - \bar{y}_i) \tag{8.25}$$

Therefore

$$X = X_n - K - \frac{\sum_1^n a_s e_s}{\sum_1^n a_s} \tag{8.26}$$

where

$$K = -(X - \bar{x}) + \sum_1^t \beta_i (Y_i - \bar{y}_i) \tag{8.27}$$

Bowley assumed that the adjustment factor K would be small unless there is considerable correlation between the sizes of the districts and the variables. All the terms in 8.27 can either be calculated (exactly) from the population data or values from the sample if they involve x.

In order to determine the error term, $\frac{\sum_1^n a_s e_s}{\sum_1^n a_s}$, Bowley wrote

$$n\bar{a} = \sum_{s=1}^n a_s\,,\ n\sigma_a^2 = \sum_1^n (a_s - \bar{a})^2 = \sum_{s=1}^n a_s^2 - n\bar{a}^2,$$

which gives

$$n\bar{a}(X_n - K - X) = \sum_{s=1}^n a_s e_s$$

Letting σ_e stand for the standard deviation of all error terms e_s, i.e., they are assumed to have the same standard deviation, and error terms are also assumed to be uncorrelated. If σ_n stands for the standard deviation of the error made in estimating X by $(X_n - K)$, then

$$n^2\bar{a}^2\sigma_n^2 = \sum_1^n a_s^2 \cdot \sigma_e^2 = n\sigma_e^2(\bar{a}^2 + \sigma_a^2)\,.$$

Therefore, the variance of the error is

$$\sigma_n^2 = \sigma_e^2 \cdot \frac{1}{n}\left(1 + \frac{\sigma_a^2}{\bar{a}^2}\right)$$

Using Yule's results on partial correlations (Yule 1911), Bowley shows that the variance of error is

$$\sigma_e^2 = \frac{R}{R_t} \cdot \sigma_x^2 \tag{8.28}$$

where t is the number of controls and R is a matrix of correlation coefficients where the first row (and column), that is r_{xi}, $i = 1, \ldots t$, are the correlations between the variable of interest, x, and the control variables, y_i, $i=1,\dots, t$; the other rows (and columns), that is $\rho_{ij}, i,j = 1,\dots,t$, are correlations between the control variables; R_t is a matrix of the correlation coefficients between the t control variables (see Bowley 1926). The correlations r_{xi}, $i = 1, \dots t$, are not known and therefore they have to be estimated from the sample. The correlations between control variables, $\rho_{ij}, i,j = 1,\dots,t$, may be calculated from the population. Standard deviation σ_x must be estimated from the sample and its standard deviation (standard error) is $\sigma_x/\sqrt{n}$.

Hence, the standard deviation of the error in estimating X from $X_n - K$ is

$$\frac{\sigma_x}{\sqrt{n}} \cdot \sqrt{\left(1+\frac{\sigma_a^{\;2}}{\bar{a}^2}\right)} \cdot \sqrt{\frac{R}{R_t}} \qquad (8.29)$$

Bowley concluded that the advantage obtained of the control variables depends only on the value of the last term, $\sqrt{R/R_t}$, and the advantage is maximised when the ratio is minimised.

If there are no control variables (t = 0) then $\frac{R}{R_t} = 1$.

If there is only one control variable, then $\frac{R}{R_t} = 1 - r_{x1}^{\;2}$, etc.

Bowley continues to show that the advantage of increasing the number of controls is in ordinary cases quite small.

> "It is clear that the standard deviation of the error of the results is in ordinary cases dominated by the value of σ_x and by n the number of observations, rather than by the controls exercised in purposive selection." (Bowley 1926, p. 50)

8.4.2.3 An Example

Bowley gave the following example on the estimation of a mean in the context of purposive selection: From an official report, the wage rates in 1912 of compositors, masons, and engineering labourers were extracted for 47 towns. The problem was to find the average wage rates of iron moulders in these towns using a purposive selection of 12. (Bowley ibid.)

No weights were used in the example, and therefore $a_1 = a_2 = \dots$; $\sigma_a = 0$, $n = 12$.

The towns were selected so that the averages of the three occupations used as controls were approximately the same as in the 47 towns together.

Parameters of the control variables were:

Compositors: $Y_1 = 33.49$, $\sigma_{y_1} = 2.326$, $\bar{y}_1 = 33.50$ shillings per week.

Masons: $Y_2 = 9.120$, $\sigma_{y_2} = .529$, $\bar{y}_2 = 9.20$ pence per hour.

Labourers: $Y_3 = 20.33$, $\sigma_{y_3} = 1.464$, $\bar{y}_3 = 20.33$ shillings per week.

Iron moulders: $\bar{x} = 38.17$, $\sigma_x = 2.01$ shillings per week.

Correlations between control variables were $\rho_{12} = .712$, $\rho_{13} = .176$, $\rho_{23} = .477$

Correlations between the study variable, x, and control variables were $r_{x1} = .54$, $r_{x2} = .38$, $r_{x3} = .002$

For all three control variables, $R = .247$, $R_3 = .354$, $\sqrt{R/R_3} = .835$

For control variables y_1 and y_2, $R = .349$, $R_2 = .493$, $\sqrt{R/R_2} = .841$

For variable y_1 only, $R = .7048$, $R_1 = .1$, $\sqrt{R/R_1} = .842$

Therefore $\sigma_n = \sigma_e \times \frac{1}{\sqrt{12}} = \frac{1}{\sqrt{12}}\sigma_x \times .835 = \frac{1}{\sqrt{12}} \times 2.01 \times .835 = .484$, or .488 for y1 only (because $a_1 = a_2 = \ldots = a_{12}$ and hence $\sigma_a = 0$).

This gives K = –.0316 and the "forecast" becomes $\bar{x} - K \pm \sigma_x = 38.21 \pm .484$, and the true value obtained from the census was 39.12. Bowley concluded:

> "The difference between the forecast is 1.9 times the standard deviation for the error, which is greater than would be anticipated. But not much dependence can be placed on a sample based on only 12 districts, since the errors of the terms involving x are considerable." (Bowley, ibid.)

Bowley also gave another example to estimate the number of males occupied in transport by road, in England and Wales, using as a control the proportion of the rural population to the whole population. He selected 12 Administrative Counties from a total of 61. In this case, the estimate was .0117 ± .0010 while the value calculated from the latest census was .0115. Bowley also drew a random sample of 12 counties and obtained an estimate .0097 ±.0010.

8.4.2.4 Stratification in purposive selection

The stratification in purposive selection as Bowley defined it becomes fairly complicated both to accomplish in practice and to analyze mathematically. Therefore, Bowley gives only a rough description on what it might be.

At the end of this section, Bowley concludes that "The advantage obtained by stratification, though it exists, may be expected to be slight. It depends on the non-rectilinearity of the regression between the control and the quantity sought in the divisions." Divisions are the parts of the districts that are used to construct the strata.

8.4.2.5 Purposive selection vs. balanced sample

Bowley's study on purposive selection in sampling is interesting in many respects: It has similarities with the basic form of so-called balanced sampling; and there is some similarity to modern regression estimation. There is also some similarity with the method that Laplace applied in the estimation of the population in France.

Royall and Herson (1973) defined a **balanced sample** as a sample that is a miniature of the original population. More specifically, they defined a balanced sample of order T as one for which the sample mean $\bar{x}_s^{(t)}$ of variable x_i' was equal to its population mean $\bar{x}^{(t)}$, for $t = 1, 2, \ldots, T$. Also, the authors noted that the notion of balanced sampling was in essence the purposive selection defined in the ISI report and the purposive selection used, e.g., by Gini and Galvani (1929). The basic idea is the same in both: a sample is made representative of the population by purposive selection of sampling units. Balanced sample differs from Bowley's purposive selection in what are regarded as sampling units. In balanced sampling, they are single observations or measurements, but in Bowley's purposive selection, they are aggregate values of 'districts' or clusters.

Bowley's approach was probably enforced by practical matters. He knew that only aggregate-level information was available, and there was no sense in defining the purposive selection on an individual level. The selection of a purposive sample had not been possible using individual-level data, especially if there were several variables. The method of Royall and Herson requires a computer to carry out the selection of the sample. At Bowley's time, the selection had to be done by hand. Even aggregate-level data produces problems. Gini and Galvani (1929) found it too difficult to select by hand a 15 percent purposive sample based on seven variables from 8,354 communes (see also Neyman 1934).

Jensen (1926) classified Kiaer's Representative Method as an example of purposive selection (in the sense Bowley defines it). Obviously that was a misunderstanding. In a way, Kiaer can be said to use 'control variables', but they were not used in the selection of sampling units as in Bowley's purposive selection. Kiaer's 'control variables' were used to assess the representative nature of obtained sample and in few cases for weighting the collected data. Kiaer's sampling design was based on an intuitive application of the knowledge of the population that census data and common sense provided. That is the same method with which stratification is done even today.

8.5 Bowley's contribution to survey methodology

Although the main purpose of Bowley's report to ISI was to create a mathematical apparatus for estimation in sample surveys, it also addresses more generally the problems in survey research. Already the presidential address in 1906 indicated that Bowley had insight into the practical issues of survey research. After that, Bowley carried out several surveys and thereby became aware of the more practical problems of surveys, especially the potential influence of non-sampling errors. The beginning of the report to the ISI shows how well he realized the problems.

> "It is necessary to define exactly the population or "universe" in question. Only those populations can be treated in which there exists or can be made an adequate directory or list of members, every one of which is theoretically accessible to observation. The attribute or variable [of interest] must also be adequately defined.
>
> It being decided, from considerations discussed below, how many persons or things should be observed, a number of them is selected at random in such a way that a priori every person or thing has equal chance of being selected. The universe, which is sampled, is in fact limited by this condition. If, for example, observations of children of school age were to made, the universe might be either children present on a certain day in state-supported schools, or in any schools, or children on the register of schools whether present or absent, or all children in the country between certain ages whether on school register or not. Which of these universes is in fact represented depends on the answers to the question: for which

> was it that we had or could make a list, and from which was it that we selected children at random, each with an equal chance of inclusion?"
>
> . . .
>
> "Minute precautions are necessary to ensure that the method of selection is completely uncorrelated with the presence of the attribute or the size of the variable.
>
> The selection being made, every person or thing selected must be observed, if possible. Where observation is impossible or inaccurate the resulting unknown element must be retained and exposed in the final report.
>
> Any breach of these conditions, however, slight, introduces an unknown element of error in the result, and destroys the relevance of the formulae. It is naturally to be understood that very small departures from the rule in large samples cannot have any great effects, but in general the magnitudes of the resulting errors cannot be estimated."
>
> . . .
>
> "A common and very injurious departure from the rules is to ignore persons of things in which observation is difficult, e.g. when no one is present in selected house when the investigator calls. Another and even more obvious mistake is to define the universe loosely, and to be content with answers from people who happen to willing to give them." (Bowley 1926)

Probably this citation was the first description of the practical problems in survey research, including non-sampling errors, but it could be part of a modern text on survey methodology. A noteworthy feature is that Bowley held acceptable only samples that were drawn from a sampling frame with adequate coverage of the population. A sample should be selected in such a manner that every unit has an equal chance of being selected and to collect data from all selected units, not only from those units that are willing or easy to observe.

8.6 Conclusions and Discussion

Bowley's impact on the historical development of statistical science is indisputable, although his works are rarely mentioned in the modern statistical literature. His scientific achievements were twofold: the research concerning the statistical inference in finite populations was partly started by his contributions. It was Bowley who, for the first time, brought together survey sampling and statistical inference (see Smith 1976). He is the link between the tradition of social research and the application of probability theory.

As early as in 1906 in his presidential address to the British Economic Society, Bowley had emphasised the importance of random selection. He believed (and probably had tried it experimentally) that the random selection of units would provide a sample that is a miniature of the 'universe', thus enabling a reliable estimation of population characteristics. In addition, he noted that random selection justified the application of the Central Limit Theorem and calculation of the standard error.

Bowley's most significant publications deal with the problems of survey data collection and the inference from these data. Since Laplace, he seems to be the first one who explicitly and systematically treated the problems of statistical in inference for fixed populations. Bowley leaned extensively on Edgeworth's contributions on the theory of probability, but his philosophy in statistical inference derives its origin from Laplace.

The theory for statistical inference Bowley created was based on the Laplace-Bayes paradigm, and obviously Bowley's memorandum was the last contribution of the kind. Bowley's adherence to the Laplace-Bayes paradigm may be the reason why he has not been acknowledged in the modern statistical literature.

Nevertheless, Bowley was an outstanding statistician. First of all, he was a practicing statistician who carried out several economic and social sample surveys in the UK on living conditions, wages and poverty, etc. thus partly starting the social survey tradition. This still has bearing on the work of modern national statistical institutes. Apart from the work on the theoretical basis of surveys, he was also an advocate of the survey method, both towards statisticians and towards social scientists.

Bowley played a decisive role in the history of survey sampling: he was one of those who persuaded the ISI to endorse Kiaer's ideas in a resolution in 1903. Soon after Kiaer's appearance at the ISI meetings, he started to elaborate on the merits of the Representative Method in the context of large-sample surveys, both by conducting surveys in England and by developing a mathematical theory for sampling. Already in his presidential address in 1906, he expressed modern views on survey methods. The report to the ISI included central principles of survey undertaking in addition to formalism for statistical inference.

Bowley's monographs were well known and often cited by Central European and Russian statisticians. His memorandum to the ISI was one of the reasons that a few years later led Jerzy Neyman to develop the modern mathematical theory for the Representative Method. Bowley was also a central figure in another respect: in the first quarter of the 20th century, he was the second in fame amongst statisticians in England, after Karl Pearson. They were both faithful to the Laplace-Bayes paradigm. For both of them, the new ideas of Neyman and especially those of R.A. Fisher were difficult to accept (see Lehman 2008). Still, in 1935, in a speech to the Study Group of the Royal Statistical Society, Bowley sticks to the old paradigm in explaining the method to infer the population from the sample:

> "The problem is strictly analogous to that of estimating the proportions of various colours of balls in a limited urn on the basis of one or more trial draws." (Bowley 1936).

The speech reveals that the only concession Bowley made to Fisher and Neyman was to use "population" instead of "universe".

9 Revolution in statistical inference

9.1 Introduction

In 1920s, a profound change took place in statistical science. The change was so profound that in the contemporary literature on the history of statistics, the time before it is briefly passed over or disregarded altogether. The change happened gradually in a decade when the new theories and methods emerged, became widely known and accepted.

At the end of 19th century, there were only vague ideas of statistical inference and the term did not exist within statistical science. Only few new contributions on inverse probability had appeared since Laplace and his ideas and methods were accepted throughout the statistical world. For example, well-known mathematicians and statisticians, such as Gauss, Quetelet, and Lexis, did not touch the topic. Poisson seems to be the only person who had an interest in developing Laplacian mathematics, but he mainly refined Laplace's methods to a more mathematically manageable form and did not renew it.

In the beginning of the 20th century in the UK, the main focus in statistical science was on "Pearsonian" statistics. Karl Pearson was the dominant person within academic statistics and his research gave direction to the development of mathematical statistics. He founded the department of "Applied Statistics" at University College in 1911, which at that time was the only place where one could study for a degree in statistics. The "Pearsonian" statistics included the analysis of distributions and correlation, and statistical analysis meant fitting distributions to data and calculation of correlations. Also, regression analysis had emerged in the repertoire.

Arthur Bowley was another significant figure besides Karl Pearson, but Bowley understood statistics as comprising two interrelated dimensions: the arithmetical and the mathematical. The former was concerned with statistical techniques as they relate to the measurement, compilation, interpolation, tabulation, and plotting of data, as well as the construction of index numbers. The mathematical part was comprised of the application of probability theory to statistics, especially the application of probability theory to evaluate the errors associated with estimation. In probability theory, Bowley leaned on Edgeworth's contributions to the law of error in the beginning of the 20th century (Edgeworth 1906, 1907, 1908, and 1909).

Interest in the application of probability to statistics started to increase in the beginning of 1900s. An example of the growth is the contents of Bowley's textbook "*The Elements of Statistics*". In its first edition, published in 1901, the subject of "mathematical statistics" took up 74 pages. In the fifth edition, published in 1925, the subject had expanded to 210 pages.

The prevailing idea of statistical inference was Laplace's inverse probability principle, and its validity was not questioned. Using Kuhn's terms, it was an era of normal science when the development takes place within a framework defined by the universally accepted postulates, values, and principles. There is no direct evidence of the existence of a paradigm, but it can be inferred indirectly from the texts of the leading mathematical statisticians of that time. For example, in 1920, Karl Pearson described the "Fundamental problem of practical statistics":

> "...The problem I refer to is that of 'inverse probabilities' and in practical statistics it takes the following form:
>
> An 'event' has occurred p times out of $p + q = n$ trials, where we have no *a priori* knowledge of the frequency of the event in the total population of occurrences. What is the probability of its occurring r times in a further $r + s = m$ trials?
>
> In statistical language, a first sample of n shows p marked individuals, and we require a measure of the accordance which future samples are likely to give with this result. For example, a medical treatment is found to be successful in p out of n cases, we require some measure of the probable stability of this ratio. It is on this stability and its limits according to the size of the first sample that the whole practical doctrine of statistics, which is the theory of sampling, actually depends. We usually state the "probable errors" of results without visualising the strength or weakness of the logic behind them, and without generally realising that if the views of some authors be correct our superstructure is built, if not on a quicksand, at least in uncomfortable nearness to an abyss." (Pearson 1920)

This problem is nearly the same, which Laplace presented in his *Memoir on the Probability of the Causes of Events* (see Chapter 4). Inference is seen as a method that is based on learning from experience. A key element in it was the assumption of the stability of statistical ratios. In this article, Pearson analyzed the inverse probability theory and pointed out that Laplace had made mistakes. However, the article fostered the method and did not try to prove its foundations (the paradigm) erroneous or to replace it by another theory.

In the early 1920s, Bowley was a professor of statistics at the London School of Economics. While Pearson was a famous biometrician, Bowley, like Edgeworth, was as much an econometrician as a statistician. Bowley also held great interest in social research and survey sampling. Another example of the customary thought model of that time is the following piece of text by Bowley in 1923:

> "One of the inverse problems of statistics, that of estimating the value of frequencies, averages, etc., in a universe from similar quantities measured in a sample, has again become prominent ... It must be freely admitted that no general solution is possible, that if we know nothing at all about the universe except what we learn from the sample then no «principle of indifference» can lead us to valid knowledge. The object of this note is to define certain conditions of preliminary knowledge under which inference can be made from the known to the relativity unknown. The method is that indicated by Professor Edgeworth in the «Journal of the Royal Statistical Society», *1908 circa* p. 387. If we are concerned, to take an example, with the correlation coefficient and obtain $r = .5$ in a sample of 1000 instances, we can readily calculate the chance that this value would be obtained if in the universe $r = .4$ or $r =. 5$ or any other assigned value; but we cannot add together these chances or proceed to any statement as to the chance that in the universe r was (e.g.) between .4 or .6, without some hypothesis about the distribution of universes with respect to r; the hypothesis that every value from 0 to 1 is equally probable is not only baseless, but also inconsistent with an equally plausible hypothesis that all values of *arcsin* r from 0 to 1 are equally probable. As is shown in the sequel, however, we are only in fact concerned with a small range of possible values of r (the smaller as n, the number of cases, increases), for values outside this range give negligible chances of obtaining the value of the sample. All we have to assume is that in a certain small range there is a continuous function representing the a priori chance of the occurrence of assigned values of r in the

> universe; then it is shown that the exact form of the function is indifferent and that it need not even be symmetrical. If *F(r)* is the function in question and the second and higher derived functions carry coefficients *1/n*, the first derived function disappears on integration, and the function itself appears in numerator and in denominator and is cancelled." (Bowley 1923)

This citation reveals the problem in statistical inference that statisticians were struggling with: to be able to estimate population parameters (which were not assumed constants), some information about the population, or "the universe", would be needed. That was brought in by an assumed *a priori* distribution of the population parameter. Unlike Laplace, Bowley assumes that the principle of indifference (see Chapter 4) cannot lead to any valid knowledge.

9.2 Theory for statistical estimation

R.A. Fisher unexpectedly established a new theory of statistical estimation in two papers[104] in 1922 and 1925, while he was working as a statistician at the Rothamsted Experimental Station (Fisher 1922 and 1925a). The theory of estimation in the modern sense did not exist before Fisher's contributions. The paper published in 1922 included a great number of completely new ideas. Stigler (2005) regarded it as an astonishing piece of work because "It announces and sketches out a new science of statistics, with new definitions, a new conceptual framework and enough hard mathematical analysis to confirm the potential and richness of this new structure." In hindsight after all these years, it is easy to see that this article was a watershed in the development of statistical science[105].

[104] **Ronald Alymer Fisher** (1890–1962) graduated in 1913 from Cambridge. After that, he taught mathematics and physics at several different schools but pursued research in both genetics and statistics and published his first major papers on these topics. He published his first paper already in 1912, before he had graduated. His scientific activity led first to a temporary statistical position in 1919 at Rothamsted Experiment Station and soon to a more permanent one (see Box 1978). Fisher remained a resident statistician in Rothamsted until 1933. In 1933, after Karl Pearson's retirement, Fisher was appointed as Galton Professor of Eugenics at the University College in London. Actually, Karl Pearson's chair was divided in two: Egon Pearson was appointed to be professor of statistics and Fisher to be professor of eugenics. This was Fisher's first academic position. Ten years later, in 1943, he was appointed to be Arthur Balfour Professor of Eugenics at the University of Cambridge, and he held that position until his retirement in 1957.

[105] Fisher read the paper to the Royal Society already in the autumn of 1921, but it was published in 1922. All of Fisher's ideas were not fully developed in 1921 when the paper was read to the Royal Society, but in the next paper, published in 1925, the ideas were more elaborated and established. Fisher was aware of the fact that all the proofs were not completed in 1921 and noted that "... the number and variety of the new results which the method discloses press for publication". Edwards (1997) points out that there was some vagueness in the concepts that Fisher used. The origin of the first paper is not completely known. Stigler (2005) claims that it is a genuine puzzle and concludes that at least one reason why Fisher wrote the paper was in reaction to Karl Pearson who did not publish Fisher's comment on an article in Biometrika.

In later papers, published in 1930s, Fisher presented his famous fiducial argument to replace inverse probability principle, together with a new mode of statistical inference, which he called inductive reasoning. These papers dealt with statistical inference for hypothetical populations but fiducial argument was also instrumental in the development of statistical inference for finite populations.

The purpose of Fisher's 1922 paper was to analyze the theoretical foundations of statistics because "... the basic principles of this organ of science are still in a state of obscurity, and it cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved." This paper is the first serious attempt to formalize statistical estimation and hence to establish a theory of estimation in the modern sense. By doing this, Fisher sets the foundations for modern mathematical statistics. He did not treat statistical inference in this paper, except in criticizing inverse probability[106]. The paper is very long (59 pages) and so rich in content that it can only be covered here superficially.

Fisher begins the article with definitions of fifteen basic concepts of statistical science (such as centre of location, distribution, estimation, consistency, efficiency, sufficiency, etc.) which he defines more accurately later in the paper. Many of the concepts are central to modern mathematical statistics but they had not been given a precise meaning before Fisher. Some were new statistical concepts for which he gave a precise meaning[107]. For a modern statistician, they belong to the arsenal as self-evident parts and they sound so commonplace that it is a bit surprising that they did not exist earlier. Fisher continues with the formulation of statistical problems:

> "... the object of statistical methods is the reduction of data. A quantity of data, ... , is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data." (Fisher 1922)

Fisher recognized three types of problems that arise in the reduction of data:

1. Problems of Specification. These arise in the choice of the mathematical form of the population. In Fisher's vocabulary, this means the distribution function.
2. Problems of Estimation. These involve the choice of methods for calculating from a sample statistical derivates, or **statistics**, which are designed to estimate the values of the **parameters** of the hypothetical population.
3. Problems of Distribution. These include discussions of the distribution of statistics derived from samples, or generally any functions of quantities whose distribution is known.

[106] In the very beginning of the paper, Fisher sharply criticized Karl Pearson and Edgeworth because of their adherence to Bayes Theorem and especially because they had defended the use of a priori probabilities. Actually, it was slightly misleading because Pearson and Edgeworth worked from the Laplace-Bayes paradigm.

[107] Hald (1998) noted that today we cannot discuss statistical theory without making use of Fisherian terminology.

The concepts of "statistics" and "parameter" were completely new in statistical science[108]. These concepts appeared to be central components in the development of estimation theory. Fisher contemplated reasons for why "the fundamental problems" were still unsolved: First of all, he argued that a distinction is seldom drawn between the sample and the population, and

> "... in statistics a purely verbal confusion has hindered the distinct formulation of statistical problems; for it is customary to apply the same name, *mean, standard deviation, correlation coefficient*, etc., both to the true value which we should like to know, but can only estimate, and to the particular value at which we happen to arrive by our methods of estimation; so also in applying the term probable error, writers sometimes would appear to suggest that the former quantity, and not merely the latter, is subject to error." (Fisher 1922)

Fisher suggested that this confusion had led to the survival of the "fundamental paradox of inverse probability, which like an impenetrable jungle arrests progress towards precision of statistical concepts".

On several occasions Fisher expressed that another obstacle to the development of statistics was the fact that statisticians dealt only with discrete distributions, mostly binominal distribution, and their normal approximations. Statisticians ended up with binominal distribution because of the frequent use a Bernoulli trial as their thought model (see Pearson 1920). Fisher clarified his point by saying:

> "The concept of a *discontinuous frequency distribution* is merely an extension of that of a simple dichotomy, for though the number of classes into which the population is divided may be infinite, yet the frequency in each class bears a finite ratio to that of the whole population. In *frequency curves*, however, a second infinity is introduced. No finite sample has a frequency curve: a finite sample may be represented by a histogram, or by a frequency polygon, which to the eye more and more resembles a curve, as the size of the sample is increased. To reach a true curve, not only would an infinite number of individuals have to be placed in each class, but the number of classes (arrays) into which the population is divided must be made infinite. Consequently, it should be clear that the concept of a frequency curve includes that of a hypothetical infinite population, distributed according to a mathematical law, represented by the curve. This law is specified by assigning to each element of the abscissa the corresponding element of probability. Thus, in the case of the normal distribution, the probability of an observation falling in the range *dx*, is
>
> $$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}$$
>
> in which expression x is the value of the variate, while m, the mean, and σ the standard deviation, are the two parameters by which the hypothetical population is specified. If a sample of n be taken from such a population, the data comprise n independent facts. The statistical process of the reduction of these data is designed to extract from them all relevant information respecting the values of m and σ, and to reject all other information as irrelevant." (Fisher 1922)

[108] Stigler (2005) notes that in this paper of Fisher, the word "parameter" is the first time it was used in its current meaning in statistics. Stigler calculated that in this article, the words "parameter" or "parameters" appeared a total of 57 times.

This citation shows Fisher's central ideas in the field of applications of statistical methods and how he perceived population and a sample from it: A population is defined by a (continuous) distribution function, such as a normal distribution, but a finite population cannot have a frequency curve. Reduction of data is performed by interpreting the available observations as a sample from a hypothetical infinite population. This idea did not exist in the writings of other statisticians before Fisher.

Next, Fisher defined the well-known criteria of estimation: *Consistency*: A statistic[109] satisfies the criterion of consistency if it is equal to the parameter when calculated from the whole population. *Efficiency:* the criterion is satisfied by those statistics that, in large samples, tend to normal distribution with the "least probable error"[110]. *Sufficiency* requires that the statistic summarises the whole of the relevant information supplied by the sample. For a sufficient statistic, no other statistic that can be calculated from the same sample can provide any additional information as to the value of the parameter. Fisher continued by saying that a statistic that fulfils the criterion of sufficiency will also fulfil the criterion of efficiency.

9.3 Method of maximum likelihood

To solve problems of estimation, Fisher sought a method that, for each particular problem, would lead automatically to the statistic by which the criterion of sufficiency would be satisfied. He was anxious to publish his idea already in 1922, although his work was not completed:

> "Such a method is, I believe, provided by the *Method of Maximum Likelihood*, although I am not satisfied as to the mathematical rigour of any proof which I can put forward to that effect. ... For my part I should gladly have withheld the publication until a rigorously complete proof could have been formulated; but the number and variety of the new results which the method discloses press for publication . . . " (Fisher 1922)

He wanted to communicate the idea to other statisticians already at that time, obviously in hope that it would catch their attention. If this was Fisher's plan, it did not come off very well. His paper did not raise any noticeable discussion, and other statisticians of the time did not understand the ideas or even the terms in it. According to Stigler (2005), the 1922 paper remained almost unnoticed for many years.

Fisher (1922) defined the maximum likelihood method as follows:

> "If any distribution involving unknown parameters $\theta_1,\theta_2,\theta_3,\ldots$ the chance of an observation falling in the range dx be presented by $f(x,\theta_1,\theta_2,\theta_3,\ldots)dx$, then the

[109] Fisher coined the term "statistic" for a function of the sample, designated to estimate the value of a parameter. Consequently, Fisher also speaks about the sampling distribution of a statistic.

[110] "Probable error" was a frequently used term in the beginning of the 20th century; it usually had the same meaning as "standard error" in modern terms.

> chance that in a sample of n, n_1 fall in the range dx_1 , n_2 in the range dx_2, and so on, will be
>
> $$\frac{n!}{\Pi(n_p!)}\Pi\{f(x,\theta_1,\theta_2,\theta_3,\ldots)dx_p\}^{n_p}$$
>
> The method of maximum likelihood consists simply in choosing that set of values for the parameters which makes this quantity a maximum, and since in this expression the parameters are only involved in the function f, we have to make *S(log f)* a maximum for variations of θ_1, θ_2, θ_3,... In this form the method is applicable to the fitting population involving any number of variates, and equally to discontinuous as to continuous distributions." (Fisher 1922)

Hald (1999) argues that Fisher created the modern version of the method of maximum likelihood single-handedly between 1912 and 1922. In recent years, there has been a lot of research on the history of maximum likelihood and its origins (see, for example, the recent discussions by Edwards 1997, Aldrich 1997, and Edwards 1974). All the authors agree that the method of maximum likelihood occurs in various rudimentary forms before Fisher, but not under this name. Notably, to some extent this method is present also in Laplace's ideas (see Chapter 4). Hald (1999) concluded that Fisher did not know these results when he wrote his first papers on maximum likelihood. In any case, Fisher put the principle on the place in statistics where it currently stands.

After defining the maximum likelihood principle, Fisher immediately explained the conceptual difference between inverse probability and likelihood, and concludes that the word "probability" is wrongly used in connection with inverse probability. Fisher defined probability as a ratio of frequencies. He said that it is not possible to know anything about the frequencies of values of the kind of events which are dealt with in the determination of inverse probability. He explained why he has chosen to use the word "likelihood":

> "We must return to the actual fact that one value of p, of the frequency of which we know nothing, would yield the observed result three times as frequently as would another value of p. If we need a word to characterise this relative property of different values of p, I suggest that we may speak without confusion of the *likelihood* of one value of p being thrice the likelihood of another, bearing always in mind that likelihood is not here used loosely as a synonym of probability, but simply to express the relative frequencies with which such values of the hypothetical quantity p would in fact yield the observed sample." (Fisher 1922)

The solution for the calculation of estimates of parameters, put forward in the method of maximum likelihood, simply consists of choosing values of these parameters that have the maximum likelihood. Fisher explained that therefore it formally resembles the calculation of the mode of an inverse frequency distribution and noted that this resemblance is quite superficial:

> "if the scale of measurement of the hypothetical quantity be altered, the mode [of distribution] must change its position, and can be brought to have any value, by an appropriate change of scale; but the optimum, as the position of maximum likelihood may be called, is entirely unchanged by any such transformation." (Fisher 1922)

Fisher argued that scale invariance is a central criterion in the maximum likelihood. In addition, a central difference between likelihood and probability is that likelihood is not a differential element and cannot be integrated. Fisher explained that there is an "absolute measure of probability" so that the elementary probabilities add up to unity, but there is no such absolute measure of likelihood. The sum of the likelihoods of admissible values will always be infinite.

9.4 A new theory of statistical inference

The Laplace–Bayes paradigm in statistical inference was not markedly questioned before Fisher. This was not only the case in the UK, but also in continental Europe and in Russia. In 1936, Fisher described the situation in Great Britain a few years earlier in the following way:

> " . . . In the latter half of the nineteenth century the theory of inverse probability was rejected more decisively [than Boole] by Venn and by Chrystal, but so retentive is the tradition of mathematical teaching that I may myself say that I learned it at school as an integral part of the subject, and for some years saw no reason to question its validity. Mathematicians were averse from abandoning a theory, which often led to plausible conclusions, and, above all, which they had nothing to replace. Its acceptance as orthodox effectively concealed from majority the fact that, not a mere restatement in more accurate terms, but a fundamentally new approach, was required. As late as 1908 we find Edgeworth, vague but definitely defensive: 'I submit that very generally we are justified in assuming an equal distribution of *a priori* probabilities over that tract of the measurable with which we are here concerned' . . . to take another example, should Karl Pearson, a few years later (1920) put forward what he, and I believe he alone, regarded as a *proof* of the disputed axiom. Such stubborn unwillingness to abandon a false position, to admit ignorance, and to start again, can only be due to mathematicians having so seldom experience of situation which call for an orderly retreat!" (Fisher 1936, p. 248)

Inverse probability had been part of Fisher's training, but his critical stand on it had emerged fairly early in his career. In the paper of 1925, he wrote:

> "... I must indeed plead guilty in my original statement of the Method of Maximum Likelihood [Fisher 1912] to having based my argument upon the principle of inverse probability; in the same paper, it is true, I emphasized the fact that such inverse probabilities were relative only." (Fisher 1925a)

Fisher argued that Bayes made the first attempt to rationalize the process of inductive reasoning. Fisher seemed to appreciate Bayes' ideas and the attempt to formulate the method, although he considered it as faulty[111]. Fisher said:

> "Bayes perceived the fundamental importance of this problem and framed an axiom, which, if its truth were granted, would suffice to bring this large class of inductive inferences within the domain of the theory of probability; so that, after a sample had been observed, statements about the population could be made, un-

[111] In his characteristic style of expressing his views, Fisher said that he appreciated Bayes and considered him as one of the greatest scientists – because he never published his paper.
"There is one point for which Bayes is seldom given enough credit. He had doubts as to the necessary truth of his axiom. So serious were these doubts that he withheld his entire treatise from publication..."

> certain inferences, indeed, but having the well-defined type of uncertainty characteristic of statements of probability. Bayes' technique in this feat is ingenious. ... His problem was: given a particular kind of sample, to state with what probability a particular type of population might have given rise to it. He imagines, in effect, that the possible types of population have themselves been drawn, as samples, from a superpopulation, and his axiom defines this superpopulation with exactitude. His problem thus becomes a purely deductive one to which familiar methods were applicable." (Fisher 1936)

Fisher had studied Bayes' procedure with care, but when compared to the original Bayes' text (see Chapter 3), the original is not immediately recognizable in Fisher's interpretation. An interesting point is that he does not refer to Bayes' procedure as an example of inverse probability. On the contrary, he says, "In a less obtrusive form the same species of arbitrary assumption underlies the method known as that of inverse probability".

Fisher gave practically no credit to Laplace for any of his contributions to statistical science. Already in his first papers on estimation (Fisher 1922), Fisher launched attacks on Laplace's theory of inverse probability, claiming that it was the greatest fallacy in modern science. Only rarely did Fisher refer to Laplace, but when he did, it was almost always critical. The critique on the principle of inverse probability is in some way present in many of Fisher's writings starting in 1920s; and still in 1960s he touches it. The first explicit attack on inverse probability and Fisher's own method for statistical inference appeared only in 1930.

In 1930, Fisher described inverse probability as follows [$\varphi(x,\theta_1, \theta_2, \theta_3,\ldots)$ is the distribution from which x is a random sample]:

> "Suppose that we know that the population from which our observations were drawn had itself been drawn at random from a superpopulation of known specification: that is, suppose that we have *a priori* knowledge that the probability that [parameters] $\theta_1, \theta_2, \theta_3,\ldots$ shall lie in any defined infinitesimal range $d\theta_1, d\theta_2, d\theta_3,\ldots$ is given by
>
> $$dF = \Psi(\theta_1, \theta_2, \theta_3,\ldots)d\theta_1, d\theta_2, d\theta_3,\ldots$$
>
> then the probability of the successive events (a) drawing from the superpopulation a population with parameters having the particular values $\theta_1, \theta_2, \theta_3,\ldots$ and (b) drawing from such a population the sample values $x_1, x_2, x_3, \ldots$ will have a joint probability
>
> $$\Psi(\theta_1,\theta_2,\theta_3,\ldots)d\theta_1,d\theta_2,d\theta_3,\ldots \times \prod_{p=1}^{n} \left(\varphi(x_p,\theta_1,\theta_2,\theta_3,\ldots)dx_p\right)$$
>
> If we integrate this over all possible values of $\theta_1, \theta_2, \theta_3,\ldots$ and divide the original expression by the integral we shall then have a perfectly definite value for the probability (in view of the observed sample and of our *a priori* knowledge) that $\theta_1, \theta_2, \theta_3,\ldots$ shall lie in any assigned limits." (Fisher 1930)

Strictly speaking, this is not the traditional inverse probability, as Fisher also noted.

Fisher criticizes several characteristics of the inverse probability principle, most notably the concept of probability in it, sampling from a superpopulation, and *a priori* probabilities.

9.4.1 The meaning of probability

Fisher did not accept that the notion of probability in "inverse probability" would be conceptually equivalent to that in "direct probability". He explained the difference by describing the difference between likelihood (Fisher's counterpart for inverse probability) and probability:

> "... while we might speak of one value p as having an inverse probability three times that of another value of p, we might on no account introduce the differential element dp, so as to be able to say that it was three times as probable that p would lie in one rather than the other of two equal elements. ...probability is a ratio of frequencies, and about frequencies of such values we can know nothing whatever ... one value of p, of the frequency of which we know nothing, would yield the observed result three times as frequency as would another value of p. If we need a word to characterize this relative property of different values of p, I suggest that we may speak without confusion of the *likelihood* of one p being thrice the likelihood of another, bearing always in mind that likelihood is not here used loosely as a synonym of probability, but simply to express the relative frequencies with which such values of the hypothetical quantity p would in fact yield the observed sample." (Fisher 1922)

A few years later in 1922, Fisher explained the conceptual discrepancy in describing the difference between probability and likelihood as follows:

> "If A and B are mutually exclusive possibilities the probability of "A or B" is the sum of the probabilities of A and of B, but the likelihood of A or B means no more than "the stature of Jackson or Johnson"; you do not know what it is until you know which is meant. I stress this because in spite of the emphasis that I have always laid upon the difference between probability and likelihood there is still a tendency to treat likelihood as though it were a sort of probability.
>
> The first result is thus that there are two different measures of rational belief appropriate to different cases. Knowing the population we can express our incomplete knowledge of, or expectation of, the sample in terms of probability; knowing the sample we can express our incomplete knowledge of the population in terms of likelihood. We can state the relative likelihood that an unknown correlation is + 0.6, but not the probability that it lies in the range 0.595 – 0.605." (Fisher 1930)

Fisher defined probability and likelihood as two different numerical measures of our *rational beliefs*, which are appropriate in different situations. On several occasions, he emphasized that likelihoods do not obey the rules of probability calculus.

9.4.2 Superpopulation and a priori probability

Fisher introduced the word "population" to replace "universe" and gave it a new definition. Statisticians at that time perceived "a universe" as a set of distinct units, which possessed attributes, or "values of magnitudes". Fisher defined the population to be hypothetical and infinite, "the resultant of the conditions we are studying". That means that the population was defined only through rules and distributions. Fisher defined the object of statistical methods to be the reduction of data, and to accomplish this, a statistician must construct a hypothetical infinite population of which the actual data are regarded as constituting a random sample. This definition of population sets a different framework or thought model for inference. Fisher defined the parameters of a population to be constants.

Before Fisher, the observable "universes" were implicitly supposed to be samples, or realizations, of a superpopulation. The superpopulation was supposed to be constantly changing, like the real universe. This led to the model of estimation, "probability that *P* is the proportion in the universe and that then *p* should be found in the random sample from the universe". The changing universe was described by a distribution of the parameters.

The difference between these two conceptions of population may seem slight, but conceptually it was a significant opening and gave a completely new perspective to statistical inference: there was no need to refer to the *a priori* probability of the population parameter anymore. In Fisher's approach, *a priori* probability is not feasible at all.

Fisher argues that the different structure of population and the nature of distributions it induces had hindered the earlier mathematicians form observing the possibilities of statistical inference. He described the problem in the following manner:

> "This form of argument leads in certain cases to rigorous probability statements about the unknown parameters of the population from which the observational data are a random sample, without the assumption of any knowledge respecting their probability distributions *a priori*. For such a deduction we need to know the exact sampling distributions of statistical estimates, calculable from the observations only, of the unknown parameters, and these distributions must be continuous. It was probably these restrictions which stood in the way of the recognition, by the early writers on probability, of a form of argument having both theoretical interest and practical value; for the problems of distributions of which they possessed the exact solutions were nearly all discontinuous, being, like the binomial expansion, and the many similar generating functions given by Laplace, distributions of frequencies, rather than of continuously variable measurement, or functions calculated from these." (Fisher 1935a)

9.4.3 Fiducial argument

Fisher called his substitute for inverse probability a **fiducial probability** that he derived from his fiducial argument. He first described it in 1930, in the following words:

> "In many cases the random sampling distribution of a statistic, *T*, calculable directly from the observations, is expressible solely in terms of a single parameter, of which *T* is the estimate found by the method of maximum likelihood. If *T* is a statistic of continuous variation, and *P* the probability that *T* should be less than any specified value, we have then a relation of the form
>
> $$P = F(T,\theta)$$
>
> If now we give to *P* any particular value such as .95, we have a relationship between the statistic *T* and the parameter θ, such that *T* is the 95 per cent value corresponding to a given θ, and this relationship implies the perfectly objective

> fact that in 5 per cent. of samples T will exceed the 95 per cent value corresponding to the actual value of θ in the population from which it is drawn. To any value of T there will moreover be usually a particular value of θ to which it bears this relationship; we may call this the "fiducial 5 per cent. value of θ" corresponding to a given T. If, as usually if not always happens, T increases with θ for all possible values, we may express the relationship by saying that the true value of θ will be less than the fiducial 5 per cent. value corresponding to the observed value of T in exactly 5 trials in 100. By constructing a table of corresponding values, we may know as soon as T is calculated, what is the fiducial 5 per cent. value of θ, and that the true value of θ will be less than this value in just 5 per cent. of trials. This then is a definite, probability statement about the unknown parameter θ, which is true irrespective of any assumption as to its *a priori* distribution." (Fisher 1930)

Also implicit in the description was a new thought model for inference, i.e., repeated sampling from the same distribution (or population). Probably the most illustrative description of Fisher's fiducial argument can be found in a paper he wrote in 1935 (Fisher 1935a), partly as a reaction to Neyman's paper of 1934[112]:

> "In a series of papers from 1930, the author has called attention to a form of argument, which seems to have been entirely overlooked by the classical writers of probability, but which arises naturally from the exact tests of significance, when the variate is tabulated in terms of the probability. This form of argument leads in certain cases to rigorous probability statements about the unknown parameters of the population from which the observational data are a random sample, without the assumption of any knowledge respecting their probability distributions a priori."
>
> "If a sample of n observations, $x_1, \dots, x_n$, has been drawn from a normal population having a mean value μ and if the sample we calculate two statistics $\bar{x}$ and s^2. "Student" (1925) has shown that the quantity t, defined by equation
> is distributed in different samples in a distribution dependent only from the size of the sample, n. It is possible, therefore, to calculate for each value of n, what the value of t will be exceeded with any assigned frequency, P, such as 1 per cent. or 5 per cent. ...
>
> It must now be noticed that t is a continuous function of the unknown parameter, the mean, together with observable values $\bar{x}$, s, and n, only. Consequently the inequality
>
> $$t > t_1$$
>
> is equivalent to the inequality
>
> $$\mu < \bar{x} - st_1 / \sqrt{n}$$
>
> so that the last inequality must be satisfied with the same probability as the first.
> ...
> Since the right-hand side of the inequality takes, by varying t_1, all real values, we may state that probability that μ is less than any assigned value, or the probability that it lies between any assigned value, or, in short, its probability distribution, in the light of the sample observed." (Fisher 1935a)

Fisher called the derived probability fiducial "to distinguish it from any of the inverse probability distributions derivable from the same data". In several occa-

[112] Fisher prepared this paper because he thought that Neyman had not completely understood his fiducial argument and had applied it in an incorrect manner.

sions, he explained the difference between fiducial and inverse probabilities. For example,

> "The inverse probability distribution would specify the frequency with which μ would lie in any assigned range $d\mu$, by absolute statement, true of the aggregate of cases in which the observed sample yielded the particular statistics $\bar{x}$ and s. This can be found by Bayes' procedure, if the prior distribution of μ is known. The distribution we have obtained [fiducial probability distribution] is independent of all prior knowledge of the distribution of μ, and is true of the aggregates of all samples without selection. It involves $\bar{x}$ and s as parameters, but does not apply to any special selection of these quantities." (Fisher 1935a)

The central idea in the fiducial argument was that pivotal quantities permitted the derivation of probability statements concerning an unknown parameter independent of any assumption concerning its a priori distribution. The idea behind the fiducial argument was not new. Simpson had already presented it in 1755 and Lambert presented it in 1760 (see Stigler 1986). Fisher coined the term in statistics, however. It was borrowed from astronomy where fiducial point meant a fixed point. If e represents the error, O the observation, and P the point to be observed, then the equation $O = P + e$ can also be written $P = O - e$. In the fiducial argument, the symmetrical difference, or the error, $e = O - P$ is treated as randomly distributed.

The meaning and interpretation of the fiducial argument and fiducial probability has give rise to wide scientific debate that is still continuing. The vast amount of papers seeking to find the correct interpretation indicates that its basic idea is not crystal clear. Some authors even consider fiducial inference as Fisher's one great failure (see Zabell 1992). Nevertheless, the fiducial argument was the central tool for Neyman leading to his theory of confidence intervals (see Chapter 10).

9.4.4 Design of experiments

Fisher's contribution to experimental design is considered to be among his most important contributions to statistical science[113]. Mahalanobis (1939, p. 271) noted that "the Fisherian technique was something in the nature of a revolution, which altered the subsequent course of agricultural experiments throughout the world." For experimental design, Fisher also introduced the principles of randomization and replication. Replication is the main source of the estimate of error, while randomization ensures that the estimate will be unbiased. Although these are essential in experimental research, indirectly these principles also had bearing on the inference for finite populations.

One of the sources for the ideas in survey sampling originates from agricultural research. That was also one of Kiaer's (1895) rationales when he introduced the Representative Method. Aside from data collection, Fisher's principles of experimental design pointed out the importance of randomization for

[113] Fisher's work on experimental design is summarized in his book, *The Design of Experiments* (Fisher 1935b). This was foreshadowed in his earlier book, *Statistical Methods for Research Workers* (Fisher 1925b), in which he presented hypothesis testing and analysis of variance.

statistical inference. In addition, the replication in experiments is a kindred idea with drawing repeated samples from the same population.

Some of the most important developers of sampling theory worked on agricultural research at the Rothamsted Agricultural Station. Rothamsted Station and Fisher's work there were taken as examples when a statistical laboratory in Iowa State University was established (see Chapter 12).

9.5 *Reception of Fisher's inference model*

It is well known that Karl Pearson had a long dispute with Fisher, which started almost in the beginning of Fisher's career. It has been claimed that some of Fisher's most important contributions were corrections to Pearson's work (see Stigler 2005). The dispute began already in 1917, and it was bad enough to have Fisher turn down the post of Chief Statistician at the Galton Laboratory in 1919 since that would have meant working under Pearson (Box 1978). Pearson also claimed that Fisher had done a disservice to statistics by widely publishing erroneous results. The Royal Statistical Society then refused to publish Fisher's papers and he resigned from the Society in protest.

There was also tension between Bowley and Fisher, although Bowley was a professor of statistics and Fisher was a young statistician at the Rothamstead Experimental Station. Obviously, there existed disagreements partly because of Fisher's view of economic and social statistics, which Bowley represented. In his 1925 book, Fisher clarified his view:

> "Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences. This particular dependence of social studies upon statistical methods has led to the unfortunate misapprehension that statistics is to be regarded as a branch of economics, whereas in truth methods adequate to the treatment of economic data, in so far as these exist, have mostly been developed in the study of biology and the other sciences." (Fisher 1925b, p. 2)

The tension between the different views on statistical science in the UK surfaced during Fisher's presentation to the Royal Statistical Society in 1934 when he read a paper entitled "*The Logic of Inductive Inference*" (Fisher 1935a). In that paper, Fisher attempted to explain his published work since his 1922 paper by reformulating the problem of statistical induction. The presentation was an outline of his contributions and contained only a few novel items. One point Fisher wanted to make in the 1935 paper was the inductive nature of statistical inference. Fisher suspected that the mathematicians were trained mainly in the technique of deductive reasoning and therefore were not capable of inductive reasoning.

> "... it would not be surprising or exceptional to find mathematicians of this class ready to deny at first sight that rigorous inferences from particular to the general were even possible." (Fisher 1935a)

The content of this paper is not particularly interesting, but the discussion to which it gave rise is interesting[114]. The prominent members of the "old guard" statisticians were invited to participate in the discussion. In the beginning of the discussion, Bowley was assigned to move the traditional vote of thanks and open the discussion. After some more or less hesitant thanks for Fisher " ... not so much for the paper...as for his contributions to statistics in general", he went on to belittle Fisher's new approach to statistical inference based on the likelihood function.

> "The essence of the method of 'likelihood', and its relation to earlier ways of approaching the problem of estimating properties of a universe from those of a sample, can be sufficiently appreciated by all those interested by studying Dr. Neyman's paper and the discussion on it in the last *Journal*. Both methods have their importance; the newer one, I think, in choosing the best arrangement of experimental work. Dr. Neyman says that 'if all we need consists in the chance that, in the universe which we are sampling, the proportion is within given limits, we certainly cannot go any further than is already known' (p. 624). He also says, 'we are interested in the probability of committing an error when applying constantly a certain rule of behaviour' (p. 624). But Professor Fisher claims (p. 40) that 'a mathematical quantity of a different kind, which I have termed *mathematical likelihood*, appears to take its place as a measure of rational belief, when we are reasoning from the sample to the population.' And in an earlier place (p. 562) Dr. Neyman said that an approach to problems of this type, where the population is not known *a priori*, 'removes the difficulties involved in the lack of knowledge of the *a priori* law.' 'It is superfluous to make any appeals to Bayes' theorem.'
>
> We are therefore left very much where we were, and I must confess that the new method appears to me to tell us only one-half of what we really need, for that *is* to determine 'the chance that in the universe, which we are sampling, the proportion is within given limits.' That seems to me the fundamental problem; but I had hoped that this subject would not have come up for discussion again to-day." (Fisher 1935a, p.55)

Bowley describes Fisher's paper as abstruse, arbitrary, and misleading. His comments were predominantly sarcastic and discourteous – even childish – and lastly, he accused Fisher of giving insufficient credit to Edgeworth (Fisher 1935a, pp. 55-57). The last comment he made clearly illustrates the difficulty Bowley had in trying to understand Fisher's ideas on the basis of the Laplace-Bayes theory:

> "Finally, I should wish everyone to consider the 'claim that mathematical likelihood supplies a measure of rational belief.' If, in fact, we knew nothing about a universe except that the variance measured in a particular way corresponded to a certain point on the normal curve of error, should we have any grounds for any rational belief, let alone a measurement of it.' (Fisher 1935a, p. 57)

The rest of the old guard statisticians continued giving more pertinent comments and remarks: first Isserlis, then Irwin and the philosopher Wolf, who was brought in by Bowley obviously to undermine Fisher's philosophical discussion

[114] As for the background for the discussion, one should bear in mind that a few months earlier, Neyman had read his famous paper to the Royal Statistical Society, meriting Fisher's developments for statistical inference (see Chapter 10). Another point to remember is that many years earlier, Fisher's 1925 book, *Statistical Methods for Research Workers*, had made him famous worldwide, but it obviously was overlooked by the academic circles in the UK.

on induction. Characteristic to these was that no one had really understood Fisher's approach. In addition, Jeffreys complained about Fisher's criticisms of the Bayesian approach (ibid., pp. 70-72).

The younger statisticians, Egon Pearson, Neyman, and to some extent Bartlett, came to Fisher's support. Pearson argued that: "When these ideas [on statistical induction] were fully understood . . . it would be realized that statistical science owed a very great deal to the stimulus Professor Fisher had provided in many directions." Neyman was equally supportive, praising Fisher's path-breaking contributions, and explaining Bowley's reaction to Fisher's critical review of the traditional view of statistics as an understandable attachment to old ideas (ibid., p. 73).

Fisher, in his reply, was equally blunt and contemptuous of Bowley and the old guard:

> "The acerbity, to use no stronger term, with which the customary vote of thanks has been moved and seconded, strange as it must seem to visitors not familiar with our Society, does not, I confess, surprise me. From the fact that thirteen years have elapsed between the publication, by the Royal Society, of my first rough outline of the developments, which are the subjects of to-day's discussion, and the occurrence of that discussion itself, it is a fair inference that some at least of the Society's authorities on matters theoretical viewed these developments with disfavour, and admitted them with reluctance. The choice of order in speaking which puzzles Professor Bowley, seems to me admirably suited to give a cumulative impression of diminishing animosity, an impression which I should be glad to see extrapolated.
>
> In his fourth paragraph Professor Bowley provides a medley of remarkably disconnected quotations, and of this I need only say that he is mistaken in thinking that Dr. Neyman's paper was based on the use of likelihood, or discussed the same topics as that which he had just heard. However true it may be that Professor Bowley is left very much where he was, the quotations show at least that Dr. Neyman and myself have not been left in his company.
>
> For the rest, I find that Professor Bowley is offended with me for 'introducing misleading ideas.' He does not, however, find it necessary to demonstrate that any such idea is, in fact, misleading. It must be inferred that my real crime, in the eyes of his academic eminence, must be that of 'introducing ideas'." (ibid., pp. 76–82)

In a way, Fisher's sarcastic reference to "his academic eminence" was understandable. Bowley had a long career in academia and was awarded with almost all the honours that a statistician could receive in Britain. He became a member of the Council of the Royal Statistical Society as early as 1898, served as its Vice-President in 1907–8 and in 1912–14, and President in 1938–40. He was awarded the Society's highest honour, the Guy Medal in gold, in 1935; he had received the Guy in silver as early as 1895. In contrast, Fisher did not associate much with the academic statisticians in Britain and had no academic position until 1933. Even then, it was granted to him with the condition that he would not teach statistics from his position of Professor of Eugenics at University College (see Box 1978).

The situation described above indicates that Fisher set off an intellectually violent revolution within statistical science. Typically, Fisher came from outside the academic establishment and created a totally new approach to statistical problems. The old guard in the UK did not understand the ideas in Fisher's inference theory and did not accept it (see Lehman 2008).

9.6 Discussion

There is no doubt that Fisher pioneered a recasting of statistics, moving away from the reliance on large sample approximations and on inverse probability, which was the approach of (e.g.) Karl Pearson and Bowley. Pearson used large samples which he measured and tried to deduce correlations. Fisher, on the other hand, focused on the use of small samples and finding causes by designing experiments rather than deducing correlations. For this purpose, Fisher in 1925 (in *Statistical Methods for Research Worker)* introduced significance tests and an analysis of variance and provided tables for *t*- and *z*-distributions.

Modern mathematical statistics started from the contributions of Fisher and it is inherently different that the "Pearsonian" statistics which dominated statistical science. Karl Pearson was the dominant person within academic statistics and his research gave direction to the development of mathematical statistics. He founded the department of "Applied Statistics" at University College, which at that time was the only place where one could study for a degree in statistics[115]. The "Pearsonian" statistics included the analysis of distributions and correlation, and statistical analysis meant fitting distributions to data and calculation of correlations. Also, regression analysis was in the repertoire. Fisher's contributions thoroughly changed the fields of interest in statistical science.

Statistical inference as a branch of statistics science started from Fisher's contributions in the 1920s and 1930s and the "old guard" did not contribute to it. Statistical estimation and statistical inference in the modern sense did not exist before Fisher. Statistical inference existed only in the form of inverse probability or inverse inference, but it was based on the conceptually different approach of the Laplace–Bayes paradigm.

Fisher's inference model was based on fiducial probabilities and sampling distributions. This was partly inspired by Gosset's (Student 1908) derivation of Student's *t*-distribution for a given sample size *n*. In Gosset's method, sample sizes were assumed to be small so that the large sample theory could not be applied. Fisher made the new basis of statistical science explicit in his 1922 paper in which he created estimation theory and sharply criticized the inverse probability approach. By doing this, Fisher put statistical science on a totally new track. In current terminology, all Fisher's precursors, including Gosset, were "Bayesians". However, a more accurate description would be "Laplacians".

Jerzy Neyman adopted Fisher's fiducial argument as the basic element when he constructed his confidence intervals in the 1930s (see Chapter 10). Thus, Fisher had an important role in the development of inference methods for finite populations, although he did not contribute directly to it.

Fisher worked nearly all of his most productive years at the Rothamsted Experimental Station. There he wrote the papers in which he established the theory of estimation and set the principles of estimation and the criteria for estimates. He also created the theory for experimental design, introduced sig-

[115] In 1919, Fisher was offered a post in the Galton's laboratory which was closely linked with the Department of Applied Statistics but Fisher refused the appointment because he recognized that nothing would be taught or published without Pearson's approval (Box 1978).

nificance tests, and eventually set up a new paradigm of statistical inference. Characteristic to Fisher's work as a statistician was that he developed the theory to meet the needs of scientific work in the real sciences.

Fisher had a great interest in epistemic questions starting from his first contributions. On several occasions, he said that the problems of statistical inference are more connected to the logic of scientific conclusions than to mathematics. Most of his publications support this argument, and even the disagreement Fisher had with Neyman was partly due to the difference between the two approaches to scientific induction (Fisher's inductive reasoning vs. Neyman's inductive behaviour).

The new paradigm that Fisher eventually started enabled the later development of modern statistical inference for both hypothetical and finite populations. The change in statistical science that Fisher set forth may be compared to the one Albert Einstein a few years earlier had done to Newtonian physics (Einstein's theory replaced Newton's theory, and Fisher's theory replaced Laplace's theory). Hald (2007) regarded Fisher's 1922 paper as revolutionary and it initiated the latest revolution in parametric statistical inference. In addition, Mahalanobis (1939), speaking of experimental design, noted that

> "... Fisher's techniques are something in the nature of a revolution and alter the subsequent course of agricultural experiments throughout the world."

One of Kuhn's arguments was that the "old guard" scientists are so deeply indoctrinated to the prevailing paradigm that they are not capable of throwing it out. The discussion after Fisher's presentation in the meeting of the Royal Statistical Society is an indication of this unwillingness. According to Kuhn, the training of new scientists aims at teaching the paradigm so that they would continue to foster the established tradition. From the very beginning, new scientists are indoctrinated to further develop the science with accepted methods and accepted aims. Only young scientists who are not yet so deeply indoctrinated into accepted theories, such as Newton and Einstein were in physics, can manage to sweep away an old paradigm (Kuhn 1962).

There are obvious similarities in the careers of Einstein and Fisher. Einstein was unable to find a teaching post after graduation in 1901 and therefore accepted a position as technical assistant in the Swiss Patent Office. During his stay at the Patent Office, in his spare time he produced much of his outstanding work. Only in 1908 was he appointed to his first academic position in Bern. Fisher also worked for many years outside the academic circles. He was appointed to his first academic position after he had made his most significant contributions to statistical science.

10 Statistical inference for finite population

10.1 Introduction

It is commonly held that Jerzy Neyman[116] in the 1930s set the foundations of the current mode of the statistical inference for finite populations, while working at the University College in London. Neyman was not trained in England, however. He obtained his education first in Russia, in the city of Kharkov, and after that in Poland. While studying at Kharkov, Neyman's teacher was S.N. Bernstein[117], who introduced him to Karl Pearson's *Grammar of Science.* Later Neyman has said that this influenced his development considerably, although it did not coincide with his main interests during his early career.

In 1925, while he stayed in Warsaw, Neyman got a government grant to study for a year in London with Karl Pearson in his laboratory. In London, Neyman began the long-lasting collaboration with Karl Pearson's son, Egon Pearson. Before his arrival in London, Neyman had sent some of his statistical publications to Karl Pearson, including two papers on statistical methods in agricultural experimentation. Pearson suggested that Neyman republish part of the second paper in Biometrika (Spława-Neyman 1925).

After the visit to London, Neyman returned to Poland, holding different teaching positions. Egon Pearson sent him material from England so that Neyman was all the time aware of what was going on in statistical science. In 1934, Neyman moved to England to join Egon Pearson at the University College, where he got a permanent position as a senior lecturer and later as a reader.

Fienberg and Tanur (1966) concluded that until his 27th birthday, Neyman had lived in isolation from western influence in his social life. On the other hand, it is not known how much "western influence" existed in the universities of Rus-

[116] **Jerzy Neyman** (1894–1981) was born in Bendery, which was in southeast Poland at the time. (Neyman's father was Czezlaw Spława-Neyman, whose name Jerzy Neyman used during his early years.) In 1912, he entered the University of Kharkov (later named Maxim Gorki University) in southwest Russia at that time to study mathematics. After finishing his undergraduate studies in 1917, Neyman remained at the University of Kharkov and was appointed to be lecturer at the Kharkov Institute of Technology. In 1920, he passed the examination for a Master's degree and became a lecturer at the university. In 1921, he moved to Bydgoszcz in northern Poland and started to work as a "senior statistical assistant" at the National Agricultural Institute. While working there, he wrote his first two scientific papers, which dealt with statistical methods in agricultural experimentation. The papers were published in 1923 in Polish, but both papers were also later published in English (Spława-Neyman 1923 [1990], and 1925). In 1924, Neyman obtained his doctor's degree from the University of Warsaw, using the work done in Bydgoszcz as his thesis.

[117] **Sergei Bernstein** (1880–1968) was a Russian mathematician and statistician. After graduating from high school, he went to Paris where he studied at Sorbonne and at École d'Electrotechnique Supérieure. After he had returned to Russia, he taught at Kharkov University for 25 years, beginning in 1907. His main interest was in elliptic functions, but some of Bernstein's most important work was in the theory of probability. He attempted an axiomatisation of probability theory already in 1917. He also generalised Liapounov's conditions for the Central Limit Theorem, studied generalisations of the law of large numbers, and worked on Markov processes and stochastic processes. It was Bernstein who coined the term 'Markov chain'.

sia where he had studied. At the beginning of the 20th century, statistics and sampling methods were widely studied both in continental Europe and in Russia. At that time, the continental school of statistical science, composed mainly of German and Russian[118] statisticians, was very active and productive.

Chang (1976) argues that Russia was the first centre of the modern mathematical theory of sampling at the end of the 19th century. Zarkovic (1956) and Seneta (1985) also claim that the foundations of statistical inference for sample surveys were an extensively studied branch within statistical science in Russia. Zarkovic (ibid.) gives a thorough but brief account of the activity in sampling theory in Russia during the First World War and the post-war period. Zarkovic (ibid.) lists at least ten statisticians who contributed to sampling theory but are not known in western countries. A.A. Tchuprov was an exception because he was also a well-known statistician in Western Europe and he published several papers in English and German on topics relating to sampling theory. An interesting detail is A.G. Kovalevsky's article titled *Basic Theory of Sampling Methods* (1924). According to Kish (Kish 1995), it was the first article written on survey sampling but it was not widely known in Western Europe because it was written in Russian. Unfortunately, nearly all copies of it disappeared in the throes of the Bolshevist Revolution. Kovalevsky confined himself to probability samples only because "then it is possible to develop an objective and scientific theory of sampling". Kovalevsky's inference model was based on the Laplace–Bayes method or paradigm, but otherwise the mathematical approach was modern. Kovalevsky also presented the theory of stratified sampling with optimal allocation.

10.2 Neyman's contributions on survey sampling

The starting point of modern statistical inference for finite populations was Neyman's paper on sampling, which he read for the Royal Statistical Society in 1934. Before this, he had already published significant papers on statistical inference, especially on hypothesis testing, together with Egon Pearson (see Neyman and Pearson 1933). In these papers, the authors established the so-called Neyman–Pearson test theory and introduced a new mode of inference they called **inductive behaviour**.

For a long time, it has been held that Neyman's scientific career started after he visited London at the end of 1920s. Statisticians have only recently rediscovered Neyman's early contributions on the design of experiments in agricultural research and on the analysis of sampling distributions (see Feinberg and Tanur 1966). There are obvious reasons why these papers were not discovered before: Neyman never actually referred to his early works. They were originally published in Polish journals in Polish under a different name. These publications

[118] At the end of the 19th century, probability theory and statistical methods were extensively studied in Russia. The contributions of mathematicians such as Chebyshev, Liabounov, Kolmogorov, Markov, and Bernstein were significant developments, which have proved to have lasting value. However, their works and contributions are not widely known in Western Europe, partly because their publications were often in Russian and partly because the papers were not published in the well-known western journals.

indicate that very early in his career, well before his first visit to England, Neyman had received comprehensive training in statistical methods and in their application in agricultural research. He was trained in Russia, so his training was based on the Russian tradition.

Neyman's first contribution to statistical inference for finite populations was his 1934 paper[119] (Neyman 1934). However, the mathematical treatment in this paper was not fully developed. His next paper (Neyman 1937) was more elaborated and rigorous. The later paper was published after the publication of Kolmogorov's monograph on the axiomatic system of probability (Kolmogorov 1933), and that seems to have had a strong influence on Neyman's thinking.

During his visit to the United States in the late 1930s, Neyman was asked about a special sampling problem to which he could not immediately give an answer. When he came home to England, he solved the problem and published it in 1938. In that paper, Neyman presented the idea of double sampling. This paper had an outstanding influence on the development of survey research because it provided a method for carrying out a large-scale survey in a large country. It was an acute problem in the U.S. at that time because the predecessor of the Current Population Survey was in the planning stage. Neyman's paper addressed acute needs.

10.2.1 Neyman's early papers

At the beginning of the 1920s, Neyman was already well aware of statistical methods. The two papers he wrote at that time deal with the design of experiments in agricultural research and sampling distributions. The writing style, experimental setup, and treatment of problems resembled that of Russian statisticians. In the first paper, Neyman conceptualises the assignment of treatments to units in an experiment as a drawing of balls from several urns without replacement; one urn for each treatment. He defined his model for analysing field experiments as follows:

> "Let us take v urns, as many as the number of varieties to be compared, so that each variety is associated with exactly one urn.
>
> In the i^{th} urn, let us put m balls (as many balls as plots of the field), with labels indicating the unknown potential yield of the i^{th} variety on the respective plot, along with the label of the plot. Thus on each ball we have one of the expressions
>
> $$(13) \qquad U_{i1}, U_{i2}, \ldots, U_{ik}, \ldots, U_{im}$$
>
> where i denotes the number of the urn (variety) and k denotes the plot number, while U_{ik} is the yield of the i^{th} variety on the k^{th} plot. The number
>
> $$a_i = \frac{\sum_{k=1}^{m} U_{ik}}{m}$$
>
> is the average of the numbers (13) and is the best estimate of the yield from the ith variety on the field.

[119] Actually, his first paper in this field was published one year earlier in Poland (Neyman 1933), but it is not well-known because it was in Polish with a short English résumé.

> Further suppose that our urns have the property that if one ball is taken from one of them, then balls having the same (plot) label disappear from all the other urns.
>
> We will use this scheme many times below and will call it the scheme with v urns.
>
> If we dealt with an experiment with one variety, we would have a scheme with one urn. In this case, expressions denoting yields will not have a variety index.
>
> The goal of a field experiment which consists of the comparison of v varieties will be regarded as equivalent to the problem of comparing the numbers $a_1, a_2, \dots ,a_v$ or their estimates by way of drawing several balls from an urn." (Spława-Neyman 1923, p. 467)

The probabilistic setup is an elaborated version of the classical urn model, except that there are several urns and the balls in urns have labels giving the values of the variable (yield of a variety on a specified plot). This setup is close to the one Lexis had developed, and it was common in Russian texts. For example, Tchuprov applied a similar setup in his analysis of distributions (see Tchuprov 1918, 1923a, and 1923b). A noteworthy feature is that Neyman separated the variables and their observed values, which was rare at that time. The article aims at finding the "best estimate" for the mean. In order to attain this, Neyman applied "Markov's results"[120], which in modern terminology is known as the Gauss-Markov theorem[121].

The urns in Neyman's setup have this special property: the removal of a ball (representing the outcome of an experimental unit) from one urn causes it to disappear from the other urns as well. The point is to analyse the influence of changing probabilities caused by drawing balls from finite urns. If the number of varieties and number of plots are small, the drawings of the balls from the urns are not independent. This paper also brings up a significant limitation that this thinking model brings about: subsequent drawings of balls from the urns are not independent, but in a real field experiment, the single observations are independent.

[120] **Andrei Markov** (1856–1922) was a Russian mathematician, a disciple of Chebyshev, and a collaborator of Tchuprov. In addition, he was a highly appreciated teacher. At the age of 30, Markov became a professor at St. Petersburg University and a member of the St. Petersburg Academy of Sciences. His works were well-known amongst the Russian statisticians. His textbook, "Calculus of Probabilities," was published four times in Russian (first edition in 1900) and was translated into German in 1912. Kovalevsky also applied "Markov's method" in deriving estimators for sample surveys (see Kovalevsky 1924).

[121] The Gauss–Markov theorem was first discovered by Gauss and later independently by Markov. It states that in a linear regression model in which errors have an expectation of zero and are uncorrelated and have equal variances, the **best linear unbiased estimator** (BLUE) of the coefficients is given by the ordinary least squares estimator, "best" meaning minimum variance among all linear unbiased estimators.

Assume that $y_i = \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$ and $E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2 < \infty, Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$. A linear estimator of β_j is a linear combination $\widehat{\beta}_j = c_{ij} y_1 + \dots + c_{nj} y_j$. The estimator is unbiased only if $E(\widehat{\beta}_j) = \beta_j$. If $\sum_{j=1}^{k} \lambda_j \beta_j$ is some linear combination of the coefficients, then the mean square error of the corresponding estimation is

$$E\left(\left(\sum_{j=1}^{k} \lambda_j (\widehat{\beta}_j - \beta_j)\right)^2\right)$$

The best linear unbiased estimator of β is the one with the smallest error for every linear combination λ. The fact that the errors need not be normal or independent and identically distributed, but merely uncorrelated, makes the theorem suitable for many situations.

The other early paper of Neyman deals with the "*Theory of small Samples drawn from a finite Population*" (Spława -Neyman 1925)[122]. He defined a small sample from a finite population to be one in which all individual in the sample are drawn in a single drawing, i.e., without replacement. The aim of the paper was to give formulas for the moments of sampling distribution.

Neyman started by noting that it is possible to draw N different samples of size n from a population of size m ($N=m!/n!(m-n)!$). The values of the research variables in each sample, S_i, were denoted by x_{i1}, x_{i2}, ... x_{in}, and their mean was denoted by x_i. Neyman analyzed four related problems in this article. The first problem was to calculate the four first moments about the mean for the distribution of the sample means x_i, $i = 1, 2, \ldots N$. The analysis was based on Pearson's frequency curve typology and parameterization.

The second problem was to define "the second moment of the squared deviation of a sample, the sample being taken at random from a finite population with a given distribution." He obtained a result which he claimed to be a generalization of formulas given by other statisticians and which he believed "to be novel and of considerable importance". He concludes this part by saying that the result agrees with the value given by "Student" when the population size, m, is indefinitely large. Here Neyman also referred to a paper by Tchuprov on the same topic, published in 1918 (see the footnote on this page).

The third problem was to analyze the "correlation between the square of the deviation of the mean of a sample from the mean of the sampled population with the square of the standard deviation, the sample being taken at random from a finite population with any given distribution."

The last problem was "the correlation between the mean of a sample and its squared standard deviation, the sample being taken at random from a finite population." He comes up with the result that the only case of independence is when "the original population is normal and indefinitely large". If the population is indefinitely large, but not normal, the standard deviation and mean are not independent. Neyman concludes that "this result seems to me important, because it shows that the normal curve is the only curve by which, knowing the frequency distribution $y= f(x)$ of the mean of the sample and the frequency

[122] There was an episode related to this paper. Neyman referred to results published by Tchuprov in Biometrika a few years earlier. In 1927, Greenwood and Isserlis published in JRSS a comment about this article saying "...Spława-Neyman ... had done considerably less than justice to the work of the late Professor A. A. Tschuprov..." (Greenwood and Isserlis 1927). The authors continue by showing that Tchuprov had published the most important results of this paper many years earlier and in a more general form (see also Fienberg and Tanur 1966). Isserlis knew Tchuprov's papers well because he had translated some of them from Russian into English. Greenwood and Isserlis comment as follows: "... The late professor Tschuprow did more than any other man of science to familiarize continental statisticians with valuable English work, and has spoken with generous appreciation of the English biometric school. If his friends and pupils find that his work is depreciated or ignored by younger men writing in English journals, and that such conduct passes in England without protest, they can hardly fail to infer that English biometricians read only their own papers or those published in a single English journal."

distribution $y = \psi(\sigma)$ of the standard deviation, we reach the frequency surface of means and standard deviations simply by multiplying $F(x,\sigma)=f(x)\times\psi(\sigma)$."[123]

Feinberg and Tanur (1996) claim that Neyman and Fisher had parallel ideas on experimental design in the 1920s. One example is that they both realized the importance of randomization. However, there were also major differences in their statistical analysis of experiments. The inference model that Neyman employed in his first papers indicates that his theory was based on the Laplace–Bayes paradigm, which Fisher had discarded many years earlier. In addition, the ideas about statistical estimation that Fisher and Neyman employed at that time were far from each other.

10.2.2 The 1934 paper of the representative method

The 1934 paper was the first of Neyman's works dealing explicitly with statistical inference for finite (human) populations, but before that he had given lectures on the topic at the University of Warsaw and at the University College in London.

The paper served two different aims: the mathematical development of estimators and the analysis of the applicability of purposive selection – and subsequently showing its severe problems. As a result, Neyman established the modern principle of the interval estimation (i.e., confidence intervals) and the so-called optimal stratification. The basic idea of interval estimation was not new, though. It was already present in Laplace's and Bowley's methods, and as Neyman noted, Bayes had already presented it (see Neyman 1934). The difference was that Neyman composed the confidence intervals on an inherently different probabilistic principle than Laplace and Bowley, by applying Fisher's fiducial argument and inference model presented by Fisher. A significant difference between Fisher's and Neyman's approaches is that in Neyman's approach, it was not assumed that observations had a probability distribution. Fisher's approach could only be applied for the inference from hypothetical populations, i.e., when samples were drawn from a specified distribution *f(x)* and not a real population.

An equally important issue was that Neyman was the first who showed that inference for finite populations was possible without the *a priori* probabilities and without any reference to a superpopulation. Fisher had earlier shown the same thing in reference to hypothetical populations. The new method that Neyman presented also proved to be a real alternative to Laplace's inverse inference for the inference in fixed populations. Compared to Bowley's method, Neyman's method could be applied with ease in many situations without problematic assumptions about the nature of the population and the two-phase approach of first solving the direct problem and then solving the indirect problem.

Neyman's theory was based on a new inference model of **drawing samples repeatedly from the same finite population**. Epistemologically, statistical inference in Neyman's approach was based on the principle of Inductive Behaviour, which

[123] When Neyman visited London for the first time, he presented his papers to Karl Pearson. Pearson bluntly denied the result concerning independence: "That may be true in Poland, Mr. Neyman, but it is not true here". Later, Neyman said that at that time, Karl Pearson did not understand the difference between independence and lack of correlation (see Lehman 2008).

he had first developed (together with Egon Pearson) for hypothesis testing[124]. It was profoundly different from Fisher's principle of Inductive Reasoning.

Neyman got his original motivation to write this paper from Bowley's memorandum for ISI. In the beginning, Neyman said this about the inducement to write the paper:

> "Owing to the work of the International Statistical Institute [reference to Jensen's report (Jensen 1926)] and perhaps still more to personal achievements of Professor A. L. Bowley, the theory and the possibility of practical applications of the representative method has attracted the attention of many statisticians in different countries. ... But I think that if practical statistics has acquired something valuable in the representative method, this is due primarily to Professor A. L. Bowley, who not only was one of the first to apply this method in practice [reference to Bowley's survey in Reading (Bowley 1913)], but also wrote a very fundamental memoir [reference to Bowley's report to ISI (Bowley 1926)] giving a theory of the method." (Neyman 1934)

The main point at which Neyman aimed his argument was that in the ISI report, the "two different aspects of the representative method" random selection and purposive selection were treated as if the selection could be done on equal terms, with both methods being equally recommended. His main purpose was to show that purposive selection was inferior to random selection.

In the beginning of the paper, before the description of the mathematical theory, Neyman acknowledges the other forerunner who had given rise to his method:

> "...However, since Bowley's book was written, an approach to the problem of this type has been suggested by professor R.A. Fisher which removes the difficulties involved in the lack of knowledge of the a priori probability law [reference to Fisher 1922, 1925a]. ... Avoiding the necessity of appeals to the somewhat vague statements based on probabilities *a posteriori*, Fisher's theory becomes, I think, the very basis of the theory of representative method.
>
> The possibility of solving the problems of statistical estimation independently from any knowledge of the a priori probability laws, discovered by R.A. Fisher, makes it superfluous to make any appeals to the Bayes' theorem.
>
> The whole procedure consists really in solving the problems which professor Bowley termed direct problems: given a hypothetical population, to find the distribution of certain characters in repeated samples. If this problem is solved, then the solution of the other problem, which takes the place of inverse probability, can be shown to follow." (Neyman 1934)

Neyman referred to the fiducial inference that Fisher had introduced a few years earlier (see Fisher, 1930). It is noteworthy that Neyman did not immediately rule out the need for *a priori* probabilities. He only said that it is not necessary to know *a priori* probabilities, and initially he formulated his theory following the Laplace-Bayes paradigm (see also Chang 1976).

Neyman defines the problem of the representative method to be a problem of estimation. This definition was new in this context and was essentially differ-

[124] The principle of inductive behavior in hypothesis testing means that if a scientist always rejects his or her null hypotheses at a given risk level, e.g., 5 percent, he or she knows that during his or her whole career, only 5% of the decisions have been wrong, though he or she does not know which ones.

ent from the problem that Bowley had treated. Neyman realized the importance of this new theory. He wrote, "... the present solution means not less than a revolution in *the theory* of statistics."

10.2.2.1 Theory of estimation

Neyman set two requirements for the "collective characters"[125] to be calculated from a sample. In Fisher's terminology, they were called statistics:

1. They must follow a frequency distribution, which is already tabled or may be easily calculated.
2. The resulting confidence intervals should be as narrow as possible.

It is noteworthy that Neyman did not separate parameter and statistic, as Fisher did, and he did not explicitly state that the collective characters in the population were constants. To obtain the required statistics Neyman first introduced three new concepts:

> The **mathematical expectation**, which is the mean value of the estimates, e.g., θ', in repeated samples from a population in which the true value is θ, that is, $E(\theta')=\theta$ [126].
>
> A **linear estimate**, which is linear with regard to the sample values, that is, $\theta' = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_p x_p$.
>
> An estimate of θ' is a **best linear estimate** of θ if it is linear in respect to sample values, x_i, and its standard error is less than that of any other linear estimates of θ.

Ten years earlier, Fisher had introduced the inference model that was based on repeatedly drawing samples from the same population, which in Fisher's theory meant samples from the same distribution (see Fisher 1922). The novelty of Neyman's approach was that it was aimed at finite populations. The idea of repeated samples did not exist in any form in Bowley's writings. The definition of expected value was not completely new because he already had a similar definition in his early papers.

Neyman pointed out that the best linear estimates as he defined them have some important advantages:

1. If n is large, their distribution practically always closely follows the normal distribution.

2. In most cases, they are easily found by applying "Markov's method".

[125] "Collective character" is a term that was coined and frequently used by Russian statisticians meaning a parameter calculated over all elements of the collective.

[126] Neyman did not require that estimates should be unbiased, but he said that only estimates where $E(\theta')=\theta$ would be considered.

3. The same method provides the estimate of their standard error.

4. If the estimate θ' of θ is a linear estimate, and if μ is the estimate of its standard error, then in cases where the sampled population is normally distributed, the ratio $t = \frac{\theta' - \theta}{\mu}$ follows the "Student's" distribution, which depends only on the size of the sample. Neyman continues by stating that this result, which was due to R.A. Fisher, leads directly to the construction of the confidence intervals. If t_ε is Fisher's fiducial coefficient, then the confidence interval with confidence coefficient $\varepsilon = .99$ for observed values θ' and μ will be given by inequality $\theta' - \mu t_\varepsilon \leq \theta \leq \theta' + \mu t_\varepsilon$.

5. Referring to Pearson's experiments, Neyman remarks that the result is "very approximately true for various linear estimates by fairly skew distributions", provided the sample size is more than 15.

Neyman referred to "Markov's method" (Markov 1912), which in modern statistical texts is better known as the Gauss-Markov Theorem. Probably it was not well-known in Western Europe at that time because it was published in Russia and originally written in Russian, and Markov's monograph was first translated into German and only later in English. Neyman referred to Markov's method already in the paper that was published in 1923. In the discussion following Neyman's presentation, Fisher noticed that "Markov's method" was almost the same as what Gauss had invented a century earlier. Hald (1998) noted that Neyman obviously was not aware of the Gauss' work. Neyman concludes that the properties of the linear estimates obtained with "Markov's method" make them exceedingly valuable for their use in applying the representative method.

Neyman continued showing that by Markov's method, estimates with desired features could be "in most cases easily found". Desired features of estimates were the following: they are asymptotically normally distributed, whose standard error is easily found, and for which the "Student's" distribution applies directly.

Mathematically, Neyman's approach is partly an application of Fisher's theory of estimation and fiducial argument in constructing the confidence limits, but there are also significant differences even in the very setup. The introduction to the proof starts as follows:

> "Suppose we are taking samples, Σ, from some population π. We are interested in a certain collective character of this population, say θ. Denote by x a collective character of the sample Σ and suppose that we have been able to deduce its frequency distribution, say $p(x|\theta)$, in repeated samples and that this is independent on the unknown collective character, θ, of the population π.
>
> The collective characters I am speaking are arbitrary. The position may be illustrated, for instance, by supposing that the collective character θ is the proportion of a certain type of individuals in the population π, and x the proportion of the same type of individuals in the sample. The distribution of x is then a binomial, depending upon the value of θ.
>
> Denote by $\phi(\theta)$ the unknown probability distribution *a priori of* θ. Suppose that the general conditions of sampling and the properties of the collective characters

> θ and x define certain values which these characters may possess. In the example mentioned above, θ, the proportion of individuals of the given type in the population may be any number between 0 and 1. On the other hand, x, the proportion of these individuals in the sample, say n, could have values of the form k/n, k being an integer $0 \leq k \leq n$." (Neyman 1934, p. 589)

Neyman initiated his proof by assuming that the parameter, θ, had an *a priori* distribution $\phi(\theta)$, which implies that he did not consider the parameter to be a constant. Actually, he did not take a stand on the nature of the parameter in this article. This is an important difference between Fisher's and Neyman's approaches: Fisher did not consider an *a priori* distribution of a population parameter to be feasible. Neyman shows later that the *a priori* probabilities are cancelled out, but he considers *a priori* distribution as relevant anyway. He did not explain why *a priori* probability was introduced. In any case, the reference to *a priori* probability distribution of the parameter indicates that in some sense the idea of the Laplace– Bayes paradigm had a role in Neyman's thinking.

> "The new form of the estimation of a collective character θ may be stated as follows: given any positive number $\varepsilon < 1$, to associate with any possible value of x an interval
>
> $$\theta_1(x) < \theta_2(x)$$
>
> such that if we accept the rule stating that the unknown value of the collective character θ is contained within the limits
>
> $$\theta_1(x') < \theta < \theta_2(x')$$
>
> every time the actual sampling provides us with value $x = x'$, the probability of our being wrong is less than at most equal to $1-\varepsilon$, and this whatever the probability law *a priori* $\phi(\theta)$.
>
> The value of ε, chosen in a quite arbitrary manner, I propose to call the 'confidence coefficient'. If I choose, for instance, $\varepsilon = .99$ and find for every possible x the interval $[\theta_1(x'),\theta_2(x')]$ having the properties defined, we could roughly describe the position by saying that we have 99 per cent. confidence in the fact that θ is contained between $\theta_1(x)$ and $\theta_2(x)$, the numbers $\theta_1(x)$ and $\theta_2(x)$ are what R. A. Fisher calls fiducial limits of θ. Since the word "fiducial" has been associated with the concept of "fiducial probability" which has caused the misunderstanding I have already referred, and which in reality cannot be distinguished from the ordinary concept of probability, I prefer to avoid the term and call the intervals $[\theta_1(x'),\theta_2(x')]$ the confidence intervals, corresponding to the confidence coefficient ε." (Neyman 1934)

Neyman showed that the *a priori* probabilities in the formulas would be cancelled out after integration (or summing). By this, he proved that in estimation, the *a priori* distribution of the parameter was obsolete. Interestingly enough, this is a similar derivation as in Bowley's analysis of the accuracy of estimates in which he shows that the accuracy is the same, whatever form the *a priori* distribution has (see Chapter 8.4.1).

10.2.2.2 Critic on purposive selection

Neyman said explicitly that the most important part of the paper was the analysis – and critique – of the method of purposive selection (Neyman 1934, p. 621).

Neyman showed, both mathematically and by example, that purposive selection would lead to inconsistent estimation. The other part of the analysis was based on an analysis of a paper by Gini and Galvani (Gini and Galvani 1929). They had drawn a purposive sample of 29 from the 214 districts of Italy to save about 13.5% of the 1921 population census data of Italy. They applied the method that Bowley described in 1926 (see Chapter 8).

Neyman defined the term "purposive selection" as a method or procedure divided in two parts: the method of obtaining the sample and the method of estimation of, e.g., an average of a variable, x.

The method of obtaining the sample assumes that the population Π of individuals is divided into several, M, districts which form the population π. The number of individuals in each district, v_i, is known, and for each district the value of one or more numerical characters y_i, which Bowley called controls, are known. For the i^{th}, the sum $u_i = \sum_j x_{ij}$ and the mean value $\bar{x}_i = u_i / v_i$. Neyman argued that the basic hypothesis of purposive selection is that the numbers $\bar{x}_i$ are correlated with the control y_i and that the regression of $\bar{x}_i$ on y_i is linear. Neyman referred to this as the hypothesis H.

If the hypothesis holds, forming the sample consists in purposive selection of such districts for which the weighted mean $Y' = \sum \mathrm{v} y / \sum \mathrm{v}$ has the same value, or at least as nearly the same as is possible, as it has for the whole population, say Y.

Neyman continues showing that this method of sampling is a special case of stratified random sampling by groups. As Neyman defines it, the method of purposive selection consists of (a) dividing the population of districts into a second-order strata according to the values of y and v, and (b) selecting randomly from each stratum a defined number of districts. Neyman emphasized that this interpretation of the method of purposive selection is necessary if it is supposed to "be treated from the point of view of the probability theory, … as there is no room for probabilities, for standard errors, etc., where there is no random variation of random sampling."

Obviously, the method that Neyman analyzed was not exactly the same design as Bowley had meant, but Neyman's modification was necessary in order to analyse the method using probability theory.

Neyman showed that in the purposive selection thus defined, there were severe inconsistencies even in the simplest cases where the hypothesis of linear dependence of control variable and target variable holds, and that the method could lead to biased samples. Neyman asks three questions: (1) Is it likely to find in practice instances where the hypothesis underlying the method of purposive selection are satisfied, that is, the hypothesis concerning the linearity of the regression and the hypothesis concerning the variation [of x_i] within the second order strata? (2) If instances where the hypotheses are not exactly satisfied are found, what would be the result of ignoring this fact and applying the purposive selection? (3) Is it possible to get any better method that purposive selection, i.e., a method that would not lose its property of being consistent when the hypothesis concerning the linearity of the regression is not satisfied?

As an answer to the first question, Neyman shows that such an instance may be found, but "it is difficult to judge how often we shall meet in practise considerable divergences from linearity." Neyman concludes that it is safer to assume that the hypothesis concerning the linearity does not hold. In addition, he suspected that the hypothesis concerning the variation within the second order strata is probably never satisfied.

As an answer to the second question, Neyman showed that estimates cease to be unbiased when an assumption on the shape of the regression cannot be made. The estimates could be kept consistent only by special adjustments.

The answer to the third question was as expected: "There is no essential difficulty to find the best unbiased estimates of the average X determined from a sample obtained by the method of stratified sampling by groups." Neyman added that stratification does not affect the method of obtaining the estimate. "In any case and whatever the variances of the x_i within the strata, the best linear estimate of X is always the same."

Neyman proved that the consistency of estimates obtained by applying Gauss-Markov theorem does not depend on any arbitrary hypotheses concerning the sampled population. "The only condition, which must be satisfied, is that the samples should contain districts from every stratum." The standard errors of the estimates, however, depend on the variability of variables within the strata.

In addition to the survey carried out by Gini and Galvani, Neyman referred to three other enquiries carried out a few years earlier. In Neyman's mind, the most important enquiry in which the representative method (or random sampling by groups) was used was the *New Survey of London Life and Labour*, carried out under Bowley's guidance. The Polish Institute for Social Problems, under Neyman's advisory conducted another enquiry, concerning the structure of Polish workers. In that enquiry, data was collected applying a random stratified sampling of groups. The third survey he referred to was an enquiry into the farming conditions in Bulgaria by Oscar Anderson[127], who applied stratified sampling by groups.

Neyman's conclusion and recommendation for a sample survey was as follows:

> "The final conclusion which both the theoretical considerations and the above examples suggest is that the only method which can be advised for general use is the method of stratified random sampling. If the conditions of the practical work allow, then the elements of the sampling should be individuals. Otherwise we may sample groups, which, however, should be as small as possible."

Neyman did not throw away purposive selection, however. As the last conclusion in his paper (ibid.), he said:

> "There are instances when we may select individuals purposely with great success. Such is, for instance, the case when we are interested in regression of some variety y on x, in which case the selection of individuals with values of x varying within broad limits would give us more precision."

[127] **Oscar Anderson** (1887–1960) was originally a Russian statistician, one of Tchuprov's students, who was widely known and recognized in Romania and central Europe. He belonged to the 'continental school' of statistics and worked in the tradition of Lexis and von Bortkewicz. Anderson was considered one of the main promoters of purposive selection. He was also one of the discussants of Neyman's paper.

10.2.2.3 Optimum allocation

For a long time, stratification based on optimum allocation has been regarded as one of the most significant results of Neyman's 1934 paper. However, for Neyman, it seemed to be a less important (side) result than the method of estimation and the critique on purposive selection. His main purpose seemed to be to show that with the help of the "Markov's method", unbiased estimates could also be obtained in the case of stratification and it led to the optimum allocation.

Bowley referred to strictly random sampling as unrestricted and stratified sampling as restricted sampling. In both cases, the principle was the same: all members of the population had an equal chance to be selected. That means that in every strata, the sampling fraction, f_i, was the same, $f_i = n_i/N_i$. Neyman showed that unbiased estimates did not require equal sampling fractions. In most cases, more efficient estimates would be obtained with an optimum allocation where sampling fractions varied between strata. In the simplest form, optimum allocation means that the sample size, n_i, in stratum i is determined by the formula $n_i = n\frac{N_i S_i}{\sum N_i S_i}$ where N_i is the size of the population and S_i is the standard deviation of the research variable in stratum i.

The implication of Neyman's optimum allocation is that every member of the population did not have an equal chance to be included in the sample. However, Neyman did not discuss it, and it did not raise any attention in the discussion after the presentation. It took many years before inclusion probabilities were explicitly included in estimators. In the next paper, he defined that a basic requirement is that each member of the population should have an equal probability to be selected.

The principle of optimum allocation had a longer bearing, however. In the paper of 1938, it was centrally involved, and it was applied in the development of the sampling design of the Current Population Survey. The implicit consequence of varying inclusion probabilities became the basis for the construction of estimators.

There have been doubts about the origin of the optimum allocation principle. Russian statisticians, A.A. Tchuprov (Tchuprov 1923a and 1923b) and A. Kovalevsky (Kovalevsky 1924) had presented virtually the same results ten years before Neyman. There has been some dispute whether or not Neyman was aware of these results. Fienberg and Tanur (1966) made a comprehensive account of it and came to the conclusion that Neyman must have at least known Tchuprov's works. Much later, Neyman admitted in public the priority of Tchuprov's paper (Neyman 1952), but he did not say that he knew those results at the beginning of the 1930s. Tchuprov referred to Markov several times, but he derived the formula for optimum allocation using different mathematics than Neyman. On the other hand, Kovalevsky applied "Markov's method" the same way as Neyman. However, Neyman never referred to Kovalesky's paper.

10.2.2.4 Neyman's own conclusions

Neyman admitted the difficulty of defining representative samples and instead proposes sampling methods that would yield a sample that can be regarded as representative. In closing the 1934 paper, Neyman said:

> "If there are difficulties in defining the "generally representative sample", I think it is possible to define what should be termed a *representative method of sampling* and a *consistent method of estimation*. These may be defined accurately as follows. I should use these words with regard to the method of sampling and to the method of estimation, if they make possible an estimate of the accuracy of the results obtained in the sense of the new form of the problem of estimation, *irrespective of the unknown properties of the population studied*. Thus, if we are interested in a collective character X of a population π and use methods of sampling and estimations, allowing us to ascribe to every possible sample, Σ, a confidence interval $X_1(\Sigma),X_2(\Sigma)$ such that the frequency of errors in the statements
>
> $$X_1(\Sigma) \leq X \leq X_2(\Sigma) \qquad (10.43)$$
>
> does not exceed the limit $1-\varepsilon$ prescribed in advance, *whatever the unknown properties of the population*, I should call the method of sampling representative and the method of estimation consistent. We have seen that the method of random sampling allows a consistent estimate of the average X what ever the properties of the population. Choosing properly the elements of sampling we may deal with large samples, for which the frequency distribution of the best linear estimates is practically normal, and there are no difficulties in calculating the confidence intervals. Thus the method of random stratified sampling may be called a representative method in the sense I am using. This, of course, does not mean that we shall always get correct results when using this method. On the contrary, erroneous judgements of the form (43) must happen, but it is known how often they will happen in the long run: their probability is equal to ε." (Neyman 1934)

The definition "representative method of sampling and a consistent method of estimation" is close to the concept of **sampling strategy** in modern sampling theory[128]. It is an important definition because it "freed" the sample from the requirement of being representative. It suffices that the sampling method is "representative", and this, in turn, means that the sampling method enables the calculation of estimates. This definition is inherently different from Fisher's definitions in that it does not involve a hypothesis about the population under study. Estimation can be applied in any population without information about its distributions. This makes Neyman's method readily applicable in most cases.

The last sentence of the citation indicates what Neyman's line of scientific induction was. Basically, it was similar to the long-run frequency interpretation of probability: in the long run, it is known how often a confidence interval includes a population parameter. The idea is the same as in the Neyman's and Pearson's test theory, which is called inductive behaviour[129]. In the next paper (Neyman 1937), he also gave the philosophical justification for inductive behaviour.

[128] In the modern literature, a sampling strategy H is defined as a pair $H = (p, t)$, where p is a specified sampling design and t is an estimator defined in it.

[129] In 1933, Neyman (and Egon Pearson) defined: "Without hoping to know whether each separate hypothesis is true or false, we may search for the rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong." (Neyman and Pearson 1933)

10.2.2.5 Discussion on Neyman's paper

Neyman read the paper for the Royal Statistical Society in 1934 and it was followed by a discussion, as usual. The discussants were Bowley, Egon Pearson, Isserlis, Fisher, and Oscar Andersson. Anderson was invited because he had conducted a very large survey in Romania using a method resembling purposive selection. Neyman read his paper one year before Fisher's first appearance in the Royal Statistical Society (see previous chapter). Already in the discussion on Neyman's paper, the different views of the younger statisticians and the "old guard" became apparent. Especially the old guard of statisticians was very sceptical about the method of confidence intervals and the mode of inductive reasoning, but comments were more courteous than the comments on Fisher's paper.

Discussion was opened by Bowley, the chairman of the session. In the beginning, he said that he was "very glad Professor Fisher is present, as it is his work that Dr. Neyman has accepted and incorporated. I am not certain whether to ask for an explanation or to cast a doubt." (Neyman 1934, p. 608). In the beginning, he also defended himself by saying that he also had had doubts about the purposive selection when he wrote the memorandum for the ISI[130]. He continued by defending the sampling methods that he had applied in the surveys he had carried out in the UK and explained the insoluble practical difficulties that an unrestricted random selection, i.e., simple random sampling, would bring up. Only Anderson explicitly defended purposive selection, claiming that in some cases, it would yield more accurate estimates than random sampling.

After Bowley had explained what he had done and why, he brought up his scepticism about the theory Neyman had presented:

> "... I am not certain whether to ask for explanation or cast a doubt. It is suggested in the paper that the work is difficult to follow and I may be one of those who have been misled by it. I can only say I have read it at the time it appeared and since, and I read Dr. Neyman's elucidation of it yesterday with great care. I am referring to Dr. Neyman's confidence limits. I am not sure that the "confidence" is not a "confidence trick." Put in a simple form I think the method is as follows: – Given that in a sample of 1000 taken random, there are 1 in 10 of defined quality, and given that the population from which the sample was drawn contained any proportion between 120 and 80 per thousand, then the chance of such occurrence is less than one in twenty (approx.). ... Does that really take us any further? Do we know more than was known to Todhunter? Does it take us beyond Karl Pearson or Edgeworth? Does it really lead us toward what we need – the chance that in the universe which we are sampling the proportion is within these certain limits? I think it does not. I think we are in the position of knowing that *either* an improbable event has occurred *or* the proportion in the population is within these limits. To balance these things we must make an estimate and form a judgement as to the likelihood of the proportion in the universe – the very thing that is supposed to be eliminated. I do not say that we are making crude judgements that everything is equal throughout the possible range, but I think we are making some assumptions or we have not got any further. ... The statement of the theory is not convincing, and until I am convinced I am doubtful of its validity." (Neyman 1934, p. 609)

[130] Obviously, Bowley never applied purposive selection in those surveys that he carried out.

This citation reveals Bowley's unwillingness to understand the new inference model which Neyman had put forward. Bowley stuck to the Laplace–Bayes inference model and did not concede that the proposed model would be any better than the existing ones. Isserlis, the next discussant, also argued that statistical inference requires *a priori* probabilities.

The last discussant was Fisher. He supported Neyman's ideas, but at the same time he was annoyed because Neyman had discarded the concept of fiducial probability and he suspected that Neyman had misunderstood the idea of fiducial inference.

Neyman begins the reply to the comments by saying:

> "The present discussion has shown, I think, (i) that my criticism against the method of purposive selection was sufficiently convincing, and (ii) that the section concerned with the confidence interval and the problems of estimations were not."

A noteworthy detail in the discussion was that none of the discussants mentioned the new stratification method that Neyman had presented or its implications in sampling. Neither did Neyman refer to it in his final comments. Although the discussants did not pay attention to it, P.V. Sukhatme, a colleague of Neyman from the University College, published soon after Neyman's presentation a paper where he compared optimum allocation and proportional allocation in stratification (Sukhatme 1935). Sukhatme proved that optimum allocation almost always yields more accurate estimates than proportional allocation and the difference is considerable when the population is extremely heterogeneous. The only case when Neyman's stratification may prove to be not recommended is when there are several different study variables that are not correlated. Interestingly enough, Sukhatme (ibid.) alluded to multi-stage sampling designs for solving this problem.

It should be emphasized that the ideas Neyman presented were new, although they were a combination of Bowley's and Fisher's ideas. The mathematical background seems to originate from the works of Russian statisticians who had analysed similar problems for a long time. Neyman partly applied the same mathematical methods he had applied while working in Russia and Poland.

The disagreement between the old guard statisticians defending the Laplace–Bayes paradigm and Neyman (and Fisher) was apparent. In this session of the Royal Statistical Society, the critic was polite, unlike a few months later in the discussion after Fisher's presentation concerning inductive inference.

10.2.3 The 1937 paper

The writing style in the 1934 paper was slightly obscure and the paper bore more emphasis on the critique of Purposive Selection than on the mathematical derivation of estimates and stratification. Three years later, Neyman published another paper (Neyman 1937) in which the mathematical treatment was more minute and stricter than in the first one. The conception of probability was especially treated in a stricter manner. Kolmogorov had published his monograph on the axiomatic system of probability in 1933 (Komogorov 1933), but Neyman obviously had access to it only after he had submitted his previous paper.

In Neyman's 1937 paper, the influence of Kolmogorov's theory is prominent, whereas in the 1934 paper, the treatment of probability was closer to that from the beginning of the century. The 1937 paper was an exposition of mathematics behind the confidence intervals and the associated philosophy of statistical inference. The paper started with an account of Kolmogorov's axiomatic system and its implications in statistics.

The sole topic of the paper was actually an application of the axiomatic probability theory to confidence intervals. Neyman dealt neither with Purposive Selection nor stratification. The paper did not only concern sampling from a finite population, but had more emphasis on inference from experiments. However, Neyman considered the setups and purposes of experimental design and sampling to be nearly the same, and he developed the theory that he thought would apply to both. Neyman had collaborated with researchers working on bacteriology where he had applied the method of confidence intervals in an experimental setup (see Matuszewski et. al. 1935).

Neyman also compared at length three different estimation methods: "Bayesian", maximum likelihood estimation, and the "method following Markov". The "Bayesian" method, which was actually Laplace's methods, was discarded by Neyman because it was not in accordance with the probability theory that he had adopted. He favoured the method of Markov (to the maximum likelihood), which led to the best (minimum variance) of the unbiased estimates which are linear functions of the observations.

Although Neyman had presented the idea of optimum allocation in the previous paper, implying unequal inclusion probabilities, he defined random sample in the "classical" way: "the probability of each individual of the population being included in the sample is the same" and separate drawings are mutually independent (except in drawing from a finite population without replacement).

A central argument in this paper was that Neyman explicitly discarded the whole concept of *a priori* probability. In addition, he did not consider the Bayes' or the Laplace's formula as feasible for inference for two reasons. In most cases, he explained, the parameters are constants and consequently they do not have a priori distributions. Secondly, if parameters had a stochastic nature, their distributions were generally unknown.

In this paper, Neyman explained in minute detail what he meant by confidence limits and how they should be interpreted:

> "...Returning to the inequalities $[\underline{\theta}(E') \leq \theta_1^0 \leq \bar{\theta}(E')]$, we notice that while the central part, θ_1^0, is a constant, the extreme parts $\underline{\theta}(E')$ and $\bar{\theta}(E')$ are particular values of random variables. In fact, the coordinates of the sample point E are the random variables ... and if $\underline{\theta}(E)$ and $\bar{\theta}(E)$ are single-valued functions of E, they must be random variables themselves.
>
> Therefore, whenever the functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ are defined in one way or another, but the sample point E is not yet fixed by observation, we may legitimately discuss the probability of $\underline{\theta}(E)$ and $\bar{\theta}(E)$ fulfilling any given inequality and in particular the inequalities analogous to (18), in which, however, we must drop the dashes specifying a particular fixed sample point E'. We may also try to select$\underline{\theta}(E)$ and $\bar{\theta}(E)$ so that the probability of $\underline{\theta}(E)$ falling short of θ_1^0 and at the same time of $\bar{\theta}(E)$ exceeding θ_1^0, is equal to any number α between zero and unity, fixed in advance. If θ_1^0 denotes the true value of θ_1 then of course this probability must

be calculated under the assumption that θ_1^0 *is* the true value of θ_1. Thus we can look for two function $\underline{\theta}(E)$ and $\bar{\theta}(E)$, such that

$$P\left\{\underline{\theta}(E) \le \theta_1^0 \le \bar{\theta}(E) \mid \theta_1^0\right\} = \alpha \quad . \; . \; . \; . \; . \; . \; . \qquad (20)$$

and require that the equation (20) holds good *whatever* the value θ_1^0 of θ_1 and *whatever* the values of the other parameters $\theta_2, \theta_3, \ldots \theta_l$ involved in the probability law of the *X*'s may be.

The functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ satisfying the above conditions will be called the lower and the upper confidence limits of θ_1. The value α of the probability (20) will be called the confidence coefficient, and the interval, say $\delta(E)$, from $\underline{\theta}(E)$ to $\theta(E)$, the confidence interval corresponding to the confidence coefficient α.

It is obvious that the form of the functions $\underline{\theta}(E)$ and $\theta(E)$ must depend upon the probability law $p(E \mid \theta_1, \theta_2, \ldots \theta_l)$.

It will be seen that the solution of the mathematical problem of determining the confidence limits $\underline{\theta}(E)$ and $\theta(E)$ provides the solution of the practical problem of estimation by interval. For suppose that the functions $\underline{\theta}(E)$ and $\theta(E)$are determined so that the equation (20) does hold good whatever the values of all the parameters $\theta_1, \theta_2, \ldots \theta_l$, may be, and α is some fraction close to unity, say $\alpha = 0.99$. We can then tell the practical statistician that whenever he is certain that the form of the probability law of the *X*'s is given by the function $p(E \mid \theta_1, \theta_2, \ldots \theta_l)$ which served to determine $\underline{\theta}(E)$ and $\theta(E)$, he may estimate θ_1 by making the following three steps: *(a)* he must perform the random experiment and observe the particular values x_1, $x_2, \ldots x_n$ of the *X*'s ; *(b)* he must use these values to calculate the corresponding values of $\underline{\theta}(E)$ and $\bar{\theta}(E)$; and *(c)* he must state that $\underline{\theta}(E) \le \theta_1^0 \le \bar{\theta}(E)$, where θ_1^0 denotes the true value of θ_1. How can this recommendation be justified?

The justification lies in the character of probabilities as used here, and in the law of great numbers. According to this empirical law, which has been confirmed by numerous experiments, whenever we frequently and independently repeat a random experiment with a constant probability, α, of a certain result, A, then the relative frequency of the occurrence of this result approaches α. Now the three steps *(a)*, *(b)*, and *(c)* recommended to the practical statistician represent a random experiment which may result in a correct statement concerning the value of θ_1. This result may be denoted by A, and if the calculations leading to the functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ are correct, the probability of A will be constantly equal to α. In fact, the statement *(c)* concerning the value of θ_1 is only correct when $\underline{\theta}(E')$ falls below θ_1^0 and $\theta(E')$, above θ_1^0, and the probability of this is equal to α whenever θ_1^0 is the true value of θ_1. It follows that if the practical statistician applies permanently the rules *(a)*, *(b)* and *(c)* for purposes of estimating the value of the parameter θ_1 in the long run he will be correct in about 99 per cent of all cases.

It is important to notice that for this conclusion to be true, it is not necessary that the problem of estimation should be the same in all the cases. For instance, during a period of time the statistician may deal with a thousand problems of estimation and in each the parameter θ_1 to be estimated and the probability law of the *X*'s may be different. As far as in each case the functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ are properly calculated and correspond to the same value of α, his steps *(a)*, *(b)*, and *(c)*, though different in details of sampling and arithmetic, will have this in common – the probability of their resulting in a correct statement will be the same, α. Hence the frequency of actually correct statements will approach α.

It will be noticed that in the above description the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results *will* tend to α." (Neyman 1937)

This is the most illustrative expression of inductive behaviour in statistical inference. Lastly, Neyman took an example to clarify the idea of this behaviouristic inference principle:

> "The theoretical statistician constructing the functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$, having the above property (20), may be compared with the organizer of a game of chance in which the gambler has a certain range of possibilities to choose from while, whatever he actually chooses, the probability of his winning and thus the probability of the bank losing has permanently the same value, $1 - \alpha$.
>
> The choice of the gambler on what to bet, which is beyond the control of the bank, corresponds to the uncontrolled possibilities of θ_1 having this or that value. The case in which the bank wins the game corresponds to the correct statement of the actual value θ_1. In both cases the frequency of "successes" in a long series of future "games" is approximately known. On the other hand, if the owner of the bank, say, in the case of roulette, knows that in a particular game ball has stopped at sector No. 1, this information does not help him in any way to guess how the gamblers have betted. Similarly, once the sample E' is drawn and the values of $\underline{\theta}(E')$ and $\theta(E')$ determined, the calculus of probability adopted here is helpless to provide answer to the question of what is the true value of θ_1."

Neyman discarded the use of *a priori* probabilities and also demonstrated why he did not accept Laplace's approach in constructing confidence limits (which he erroneously called Bayes' approach):

He considered n variables $X_1, X_2, \ldots X_n$ which were supposed to have "an elementary probability law" $p(x_1 \ldots x_n \mid \theta_1, \theta_2, \ldots \theta_l)$ that depends on l parameters. The values of parameters were assumed to be constants (that is, $\theta_1, \theta_2, \ldots \theta_l$ are not random variables) but their numerical values were not known. For estimating one parameter, say θ_1, Neyman defined two functions $\theta(E)$ and $\underline{\theta}(E) \leq \bar{\theta}(E)$. Functions had a single value at any point E of the sample space. If E' is the observed sample point, it is assumed to be possible to calculate values of the functions and state that the true value of θ_1, say $\theta_1{}^0$, is within the limits $\underline{\theta}(E') \leq \theta_1{}^0 \leq \bar{\theta}(E')$.

Neyman required that the probability of $\theta_1{}^0$ falling within these limits should be large, say $\alpha = 0.99$. According to the Laplace–Bayes method, this condition would be given by the formula

$$P\left\{\underline{\theta}(E') \leq \theta_1{}^0 \leq \bar{\theta}(E') \mid E'\right\} = \alpha$$

The probability is conditional to the observed sample point, E', which indicates that the parameter, $\theta_1{}^0$, is assumed to be a random variable. That is in contradiction with the assumption that the parameter would be constant. Under this assumption, the only values the probability can have are zero or unity, whatever the fixed point E' and the values $\underline{\theta}(E')$ and $\bar{\theta}(E')$ are.

Neyman concludes that this is inherently different than the idea in constructing confidence limits: for confidence limits, the probability should be calculated under the assumption that the true value of θ_1 is $\theta_1{}^0$

$$P\left\{\underline{\theta}(E) \leq \theta_1{}^0 \leq \bar{\theta}(E) \mid \theta_1{}^0\right\} = \alpha$$

and it is required that this equation holds whatever the value $\theta_1{}^0$ of θ_1 may be. The probability is conditional to the true value of the parameter.

10.2.4 The 1938 paper

In 1937, W. Edwards Deming invited Neyman to the Graduate School in the U.S. Department of Agriculture. After Neyman's presentation at a conference, Milton Friedman and Sidney Wilcox presented a problem to Neyman which he was not able to solve immediately. When Neyman got back to London, he started to work on it. The result was published in 1938 with the title, "*Contribution to the Theory of Sampling Human Populations*" (Neyman 1938). In this paper, Neyman defined a method that is now known as double sampling. The problem presented to Neyman was as follows:

A field survey is to be undertaken to determine the average value of some character of a population, π, for example the amount of money that families spend for food in a population of families residing in a certain district. The phenomenon is so complex that the collection of data requires long interviews by specially trained enumerators and, hence, the cost of data collection per family is relatively high. The total survey costs must remain at an acceptable level, which would mean that only a small sample could be selected. If the character under study has considerable variation, the sample might be too small to yield sufficiently accurate estimates.

The purpose is to estimate the average, $\bar{x}$, of a character, i.e., variable x. It is assumed to be correlated with another variable that is easier to obtain with lower cost per family. An accurate estimate of the second variable, y, can be obtained at a relatively small expense, and for any given value of it, the variability of the original variable will be smaller than in the whole population. Neyman concluded that "... a more accurate estimate of the original character may be obtained for the same total expenditure by arranging the sampling of the population in two steps." In the first step, only data for the second variable, y, is obtained from a large random sample in order to obtain an estimate of its distribution. In the second step, the sample is stratified according to the values of the second variable, y, to draw a smaller random sample from each stratum. Neyman states that "the question is to determine for a given expenditure, the sizes of the initial sample and the subsequent samples which yield the most accurate estimate for the original character".

Formally the problem is as follows: The first sample, e.g., is a simple random sample of size n'. Based on the first sample, the population is stratified into a number of classes according to the variable y. That means that the range of y is divided, e.g., in s intervals, each interval defining a stratum of the population. Let $W_n = N_h/N$ be the proportion of the population falling into stratum h, and let $w_h = n_h'/n'$ be the proportion of the first sample falling into stratum h. The second sample is a stratified random sample of size n (n_h units from stratum h), in which the second variable, x, is measured.

The unit costs of data collections are C_n and $C_{n'}$ and the combined total cost of data collection $C = nC_n + n'C_{n'}$. The problem is to choose n' and $n_{h\,,}$ so that the variance of the estimate with given costs is minimized.

The derivation of n_h and n' that leads to minimum variance is fairly complicated. In this, Neyman again applied the Gauss-Markov theorem. He suggested taking n_h to be proportional to $W_h S_h$, (where S_h is the standard deviation of x_i in stratum h) when n' and n are given. Hence the proposed optimal allocation is

$$n_h = n\left(\frac{W_h S_h}{\sum W_h S_h}\right)$$

Neyman examined also the condition when double sampling would be preferred to simple random sampling and showed that it was not always the case. It was possible that double sampling would be worse than simple random sampling.

What was probably more important was that Neyman's paper directly addressed existing needs in the United States and showed how complex survey design could be approached.

Double sampling is a special case of two-phase, or multi-phase, sampling designs which are frequent in large-scale surveys in many countries. If the information concerning the population is scarce, then two-phase sampling provides a method to obtain accurate samples at lower costs. Later it has been observed that the theory is also useful for estimation in the presence of non-response. In addition, the article of Neyman showed a way to use auxiliary information to increase the accuracy of estimation. In this sense, the last paper on sampling (Neyman 1938) may be regarded as important as the first one (Neyman 1934).

10.3 Conclusions

Neyman's contributions on statistical inference are numerous, but he wrote only a few papers on statistical inference for finite populations and sampling theory. Nevertheless, the three papers he wrote in the 1930s established the foundations of modern statistical sampling theory. It is a popular perception that the first paper was the most important, but that may not be the case. Neyman presented his central ideas in it, but they were still fairly obscure and incoherent. In the next paper, published in 1937, Neyman expressed in rigorous fashion the mathematical foundations of the theory for confidence intervals and the justification for his inference model (inductive behaviour) in finite population inference. The second paper was also more theoretically focused solely on the mathematics of confidence intervals. One of the main points was the application of Kolmogorov's axiomatic probability system in estimation theory.

The second paper did not provide any tools for the survey practice, however. The third paper addressed directly to a practical problem in undertaking a survey in large human population, and gave tools to design complex sample surveys. After Neyman's third paper, and after some other European statisticians (e.g., Cochran and Sukhatme) settled in the U.S., a period of rapid development of sampling methods took place. Important contributions to the modern sampling theory began to appear on the first half of 1940s. After that the development was very rapid and by the first half of 1950s the classical sampling theory was established.

Russian statisticians were active in developing statistical methods in the first quarter of the 20th century. Neyman must have been aware of the results that Russian statisticians had obtained in the early 1920s (see Fienberg and Tanur 1966). First of all, he was trained in Russia and the style writing in Neyman's early contributions is close to that of Russian statisticians, and furthermore, his results are in accordance with the Russian school. Therefore, it seems implausible that he had not been aware of its results.

An obvious conclusion is that Neyman was aware of the works of Russian statisticians such as Markov and Chebyshev. Neyman refers to Markov in many instances, including the paper he read before the Royal Statistical Society in 1934. Neyman held the Markov method (Gauss-Markov theorem) better than the principle of maximum likelihood. In the 1934 paper, he also refers to other Russian statisticians, such as Anderson, Bernstein, and Orzeki, but he does not refer to Tchuprov or Kovalevsky, who had contributed to sampling methods on finite populations already in the early 1920s. On the other hand, Neyman had earlier referred to Tchuprov's other results (other than optimum allocation), which were published in the paper in 1923.

One of Neyman's greatest inventions was the application of Fisher's estimation theory to samples from finite populations and subsequently the development of an estimation method based on confidence intervals. A central outcome of Neyman's contributions on statistical inference was that they "freed" the sampling theory from the practical and theoretical difficulties of Laplace's principle. Probably without this, the sampling theory could not be developed to what it is now. At least it is difficult to see how the modern survey methodology could have been established on Bowley's theory.

Another significant invention of Neyman was the Best Linear Unbiased Estimate (BLUE), which he obtained by using the (Gauss-)Markov theorem. Later, the BLUE estimators (or more generally linear estimators) have been essential for the development of estimation methods. The use of the Gauss-Markov theorem is not necessary in the derivation of estimators, but at that time, it revealed the linear nature of the estimators. Neyman's estimation method also had another important characteristic: it was possible to construct estimators without any reference to a probability distribution of variables. Thus, finite population parameters could be estimated with only very slight assumptions on the nature of the observed variables. In addition, estimators derived by Markov's method had some desired features: they are asymptotically normally distributed, their standard error is obtained with the same method, and the "students" distribution can be applied to construct confidence intervals.

The "distribution free" estimation made Neyman's method readily (and easily) applicable in most cases where surveys were needed. It opened a totally new terrain for the development of survey sampling. It is difficult to imagine how survey sampling as it is practiced today could be formulated based on maximum likelihood estimation in which the distribution of a variable is a starting point.

The critique that Neyman directed at purposive selection was so persuasive and convincing that purposive selection disappeared from the statistical writings for a long time. Obviously, attempts to develop it further did not continue, and purposive selection disappeared from the arsenal of national statistical institutes[131]. Interestingly enough, purposive selection close to the form Bowley presented it re-emerged in the 1970s in the form of balanced sampling.

Purposive selection did not disappear completely from the survey research armature. A bit more widely understood, the method remained a standard sampling method in marketing and opinion surveys in the form of quota sampling. One should also bear in mind that the idea of purposive selection has remained in the sampling theory in designing stratification.

Neyman's inference model, inductive behaviour, has been essential for the development of the survey method because it gives a quick and more objective interpretation to epistemological probability (partly concealing its problems, though). The principle of inductive behaviour resembles long-run frequency interpretation of probability. This interpretation was soon unequivocally accepted by survey statisticians, and discussion about the nature of inductive inference disappeared from the discussion on statistical inference and from sampling literature.

[131] It is difficult to find examples of surveys where purposive selection was used (except that of Gini and Galvani). Purposive selection was very difficult to apply in practice, and that probably set limitations on its use.

11 Fisher-Neyman paradigm of statistical inference for finite populations

The first papers which Neyman published indicate that he was trained within Laplace–Bayes paradigm and he was applying its principles. After moving to England, Neyman soon became aware of Fisher's philosophy. Erich Lehman, a Neyman's student at Berkeley (USA), colleague, and friend, wrote in 2008 that the year 1925–26 was difficult to Neyman and also to Egon Pearson. They began to realize that the work of Fisher required a rethinking of the current philosophy of inference. This was exceptionally difficult for Egon Pearson because his father "was not able or never saw the need to" make such a shift (Lehman 2008).

It is widely held that the contributions of Jerzy Neyman in 1930s established a new model of statistical inference for finite populations which is the basis of currently prevailing inference paradigm in greatest part of survey research. This paradigm can be called Fisher – Neyman paradigm of statistical inference for finite populations.

The justification to call the new paradigm Fisher-Neyman paradigm comes from the fact that Fisher's statistical theory made the basis of theoretical development. Fisher created alone the foundations of estimation theory and the principles statistical inference as they are currently understood. However, Fisher did not contribute directly to the inference for finite populations. Neyman's efforts were essential in adjusting Fisher's inference theory to finite population problems.

The ideas Neyman presented were innovative, even though they were partly based on Bowley's and Fisher's ideas. In addition, Neyman applied the mathematical methods he had learned while studying and working in Russia and Poland.

In text touching sampling techniques, practically never the linkage between Neyman's and Fisher's theories has been brought up. Neyman adopted from Fisher the inference model of drawing repeated samples and applied it in sampling from finite populations. This inference model is the corner stone of modern inference theory. When he adopted it in the mid 1920s Neyman changed profoundly his approach to inference.

There are also significant differences between Fisher's and Neyman's approaches. Instead of using Fisher's Maximum Likelihood estimators Neyman developed Best Linear Unbiased Estimators (BLUE) by using (Gauss-)Markov theory. Neyman's BLUE estimators could be applied independent of the distributions of variables and practically in any finite populations. This feature made Neyman's method very appealing for survey research. Neyman developed the idea of interval estimation or confidence intervals for estimators. Originally confidence intervals were based on Fisher's fiducial intervals. Later it appeared that Fisher's fiducial intervals and Neyman's confidence intervals are conceptually different.

Fisher's ideas on estimation and statistical inference quickly gained ground among young statisticians but the old guard obviously never accepted them (Lehman ibid.). The new generation started to foster and develop the new idea of inference both for hypothetical and finite populations. The inference model which Neyman created for finite populations replaced the method that Bowley had developed in 1920s, which was based on the Laplace–Bayes paradigm. After Neyman's 1937 paper, and after some other European statisticians (e.g., Cochran and Sukhatme) settled in the U.S., a period of rapid development of sampling methods took place. Important contributions to the modern sampling theory began to appear on the first half of 1940s. It took still a decade before sampling methods were mature enough to be postulated in the well-known textbooks and by the first half of 1950s, the classical sampling theory was established.

The paradigm shift can be regarded as intellectually violent, as can be concluded from the documented discussion after Neyman's presentation to the Royal Statistical Society in 1934 and especially from the discussion after Fisher's presentation few months later.

Typical features of Fisher-Neyman paradigm

The most important characteristic of the Fisher–Neyman paradigm is its inference model, which is based on the idea of repeatedly drawing samples from the same population, thus creating the sampling distribution. The Central Limit Theorem is the basis of statistical inference. Confidence intervals for estimates are formed using the Student's distribution. The estimation method involves practically no assumptions regarding the distributions of observed variables.

Another central characteristic of the Fisher–Neyman method is the fact that no assumption, neither implicit nor explicit, of superpopulation is needed. Instead, population parameters are assumed to be constants. Consequently, they do not have probability distributions, and therefore, *a priori* distribution has no role.

In most cases, statistical inference is based on Neyman's idea of inductive behaviour[132]: a scientist working by the same criteria will make correct decisions in the predefined proportion, or a predefined proportion of tolerance intervals will include a population parameter. It is not possible to say anything about the truth of a single decision or about whether a single tolerance interval includes the parameter.

[132] Fisher never accepted inductive behaviour as a feasible mode of inductive reasoning. This was a central point in the long-lasting dispute between Fisher and Neyman. Fisher's standpoint on this is clearly shown in the following text: "..., it would still be true that the Natural Sciences can only be successfully conducted by responsible and independent thinkers applying their minds and their imaginations. The idea that this responsibility can be delegated to a giant computer programmed with Decision Functions belongs to the phantasy of circles rather remote from scientific research. The view has, however, really been advanced (Neyman 1938) that Inductive reasoning does not exist, but only 'Inductive behaviour'!" (Fisher 1959, p. 60)

12 Emergence of modern sampling techniques

12.1 Introduction

The three papers which Neyman published in the 1930s established the foundations of the theory of statistical inference for finite population. In the first paper, which was partly a reaction to Bowley's memorandum to the ISI, he emphasized the importance of randomization for statistical inference, not only random selection of units in order to obtain a representative sample (Neyman 1934). That was a novel idea which was not presented earlier. The impact of the paper was not immediate, however. As Hansen and Madow put it, "there was still the need for communication, understanding, acceptance, and the adaptation and extension of the results he [Neyman] had presented." (Hansen and Madow 1976). Hansen and Madow (ibid.) also claim that at the end of 1930s, sampling as a new method was not accepted as trustworthy by the public or by the administration (in the U.S.).

The first two papers of Neyman were theoretical and did not help in solving the problems which statistical offices were struggling with. The last paper (of these three) was more influential than has often been recognized because it was motivated by a real and acute sampling problem in the U.S. administration. In this paper, Neyman showed a principle how sampling in a large-scale survey could be done rigorously and lead to reasonable data collection costs. Hansen and Madow confessed that this paper had two different but equally important messages for them: the advantages of the methods are rarely universal; and the rational decisions on what survey design to use are possible only if some previous knowledge of the population is available. These facts, which are obvious today, put the development of sampling designs in the U.S. Bureau of the Census on a right track (Hansen and Madow ibid.).

In the UK, statistical methods were also developed at Rothamsted Experimental Station but the main focus was on experimental (agricultural) research. Even research concerning sampling methods aimed at sampling for agricultural surveys. Nevertheless, the development done at Rothamsted had significant, though partly indirect, influence on survey sampling. Especially, the new theory for statistical inference, based on the theory (paradigm) Fisher had created, gained ground quickly[133]. In addition, several famous statisticians started their careers at Rothamsted, and Fisher's influence was obvious. Some of them later continued their careers at Iowa State College in the U.S. (for example William Cochran and Oscar Kempthorne). In addition, Rothamsted Experimental Station served as an example for the establishment of the (statistical) Experimental Station at Iowa State College in 1933 (see David, 1984).

[133] In addition, Neyman and Egon Pearson quickly embraced Fisher's theory and developed their own theory of statistical testing. That is not touched on in this context because it aims at statistical inference for infinite hypothetical populations.

One of the ground-breaking studies on the use of sampling methods was carried out by F. Yates and I. Zacopanay at Rothamsted Experimental Station (Yates and Zacopanay 1935). They analysed the estimation of yields of cereal crops by sampling methods. Their approach originated from analysis of variance, which was frequently applied in the analysis of field experiments. However, Yates and Zacopanay aimed at estimating totals and means, typical to a survey, not to reveal effects or causes. In addition, estimation was not based on an assumption about the distributions of variables, i.e., the estimators were "distribution free". Samples were selected at random, and the authors concluded that "bias cannot arise if proper methods of random selection are adhered to." They described how to determine the optimal percentage of sampling and showed that with their apparatus, approximately 9% would be optimal. They also analysed the gain obtained by subdivision, i.e., stratification, of plots for sampling and noted that subdivision was "advantageous". They found out that sampling for the ratio of grain to total produce could give much more precise results than the usual procedure.

The paper of Yates and Zacopanay (ibid) was an opening in using modern sampling methods in agricultural research and it was often cited. There were also some other notable characteristics in this paper: it was not related to Neyman's work or ideas; random selection was assumed as a self evident method (obviously due to Fisher); the main criteria in comparison was sampling error although it was not explicitly defined, implicitly it was based on Fisher's repeated sampling (from a finite population) "variation from sample to sample". The basic idea of statistical inference was that of Fisher's, although it was not expressed explicitly.

While the foundations of modern survey sampling methodology were laid in the UK, the current methodology was created in the United States. The methods were formed in a relatively short period at the end of 1930s and in the beginning of the 1940s. Very few traces can be found that any remarkable activity in this field existed in Europe in the 1940s. In the United States, the development mainly took place in two centres: at Iowa State College and the U.S. Bureau of the Census (see Hansen and Madow 1976, David 1984, Rao 2005). However, the idea soon spread, and after the very beginning, many other institutions started to develop sampling methods.

In the U.S. in the 1930s, there were several reasons that sparked the rapid development of sampling methods: in the background, there was the fact that in the U.S. there already existed a long and strong tradition of survey research in several areas; the socio-economic situation in the country required measures from the government, especially Franklin D. Roosevelt's New Deal political programs. In addition, there was rising criticism of the sampling methods that were applied to various surveys. According to Hansen and Madow (1976), in late the 1930s, the estimates of the number of unemployed persons, obtained by many different methods, varied from 3 to 11 million.

12.2 Early history of survey sampling in the United States

Stephan (1948) gives a thorough account of the situation and the desire for efficient sampling methods in the U.S. before the 1940s. In the beginning of the 1940s, there was already nearly a century long tradition of data collection in many areas. Stephan (ibid) gives examples of four general lines of statistical activities which had a long tradition, and in which better sampling methods were required: In agriculture, for example crop estimates; in economic statistics of prices, wages, employment, etc.; in statistical social surveys and health studies, and in public opinion polling.

Agricultural crop and livestock estimation: Already in the first half of the 19th century, a variety of statistics were collected in the United States concerning agricultural production, such as acreage planted in each principal or special crop, the estimated yield, actual yields, numbers of livestock, equipment, farm labour, etc. Agricultural research embraced several different methods which, strictly speaking, do not belong to survey methods, such as periodic censuses, voluntary reporting by selected respondents, and records produced in connection with taxation, marketing, and foreign trade.

A statistical division was established in United States Department of Agriculture in 1865. In 1866, regular reports on acreage, condition of crops, yield, and livestock were begun. The reports were based on sample data but, according to Stephan (ibid.), the sampling methods in these surveys were simple. The methods were a compromise between the cost and slow returns of enumeration on the one hand, and being without any current information on the other hand.

Economic statistics: The history of economic surveys in the U.S. is nearly as long as that of agricultural surveys. Some states of the U.S. established a statistical bureau in the second half of the 19th century, and the Federal Government set up a Labor Bureau in 1884. These bureaus developed staffs of field investigators to collect periodic data relying mainly on mailed questionnaires. The aim was, for example, to get wholesale prices "in representative markets" and to get labor data by sending agents into various districts with a list of employers from which they could choose what they believed to be a representative group.

Social surveys and health surveys: Sampling had been used extensively in studies of poverty and unemployment already at the end of 19th century. For example, during the depression of 1873-79, Carroll D. Wright used police in 19 cities and wrote to assessors in 375 towns throughout Massachusetts inquiring about the number of unemployed. Wright was a forerunner of the representative method and later he took frequent part in the ISI meetings. He was also in correspondence with Anders Kiaer.

Numerous health surveys were carried out in the U.S. from time to time. Stephan (1948) mentioned many examples in which the applied sampling methods aimed at obtaining a sample that represents the population.

Public opinion polling: The practice of surveying public opinion emerged in the U.S. from the simple beginnings of the "straw vote", conducted by newspapers to sense public reactions to candidates in elections and to obtain material for human interest stories by interviewing the "man in the street". According to George Gallup, the first "straw poll" was carried out in 1836 (Gallup 1976). Already before 1900, the New York Herald had collected pre-election reports and estimates from all over the United States and attempted to forecast the outcome of elections. The habit to conduct similar polls spread to most of the large newspapers.

According to Gallup (1976), the first modern and successful public opinion poll was conducted in 1936 in the context of presidential elections. By "modern", Gallup meant a survey in which the selection of respondents aimed explicitly and actively at obtaining a representative sample. Data collection was not based on random sampling. It was carried out using the "Gallup method", which in practice was quota sampling. Hansen remarks (Hansen 1987) that the first Gallup Poll in 1936 had the greatest impact on public acceptance of sampling because of its successful performance.

Although the early polls employed unsophisticated methods in the modern sense, they had an important role in paving the way for more important surveys. Opinion polls have a special role in survey research in the sense that their results are confirmed (or disconfirmed) in the elections. This practically never happens in most other surveys. Successful opinion polls raised reliance on the survey method amongst the general public. On the other hand, in the U.S., there were some ill-fated examples of public opinion polls that received considerable attention in newspapers in the 1920s and 1930s. Probably the unsuccessful examples also paved the way for more rigorous and methodically sound sampling designs because the nearly catastrophic failures proved the importance of a correct sampling method. In 1939, Stephan made a detailed analysis on why these polls had failed and gave a step-by-step guide how a large-scale survey should be carried out (Stephan 1940). He emphasized, referring to Bowley, that it was necessary to have means to access the selected population. In other words, a frame with adequate coverage and access information was required in drawing a sample.

In addition to public opinion polls, market research studies and consumer surveys conducted by business concerns, publications, and advertising agencies were popular in the U.S. before the 1940s. According to Stephan (1948), there was also in this area a growing interest in public opinion research among political scientists, sociologists, and others, which led to scientific interest in the improvement of the technique of opinion polling.

Stephan (ibid.) noted that while in the early instances, the sampling procedures were simple and usually employed uncritically with no great attention to accuracy and representativeness, the problems of observing and recording data were almost always far more serious than the problems of sampling. In modern survey literature, it is customary to distinguish sampling and non-sampling errors. Stephan refers to non-sampling errors and emphasized that their impact on results may be more serious than that of sampling errors.

12.3 Development of random sampling techniques

Systematic sampling has long been in use in many different areas. For example, in taking samples geological research and the practice of "cruising" a forest and sampling trees at uniform intervals has been well established in forestry for a long time. In addition, systematic sampling was frequently used in agricultural surveys. A comparable example in social research is the series of studies of workers in the unemployment insurance system which John Hilton made beginning in 1923 (see Chapter 8) in the UK. In those studies, the samples were selected systematically from the files of the Labor Exchanges. Hilton found a sample of only one per cent quite satisfactory to meet the practical administrative and policy-making purposes for which the studies were made.

In surveys on human population, also the selection of households in Kiaer's Representative Method and the sampling method which Bowley applied can be characterized as systematic sampling. In the first half of the 20th century, some kind of systematic selection of respondents was nearly the only method which in practice was possible on human populations because of the scarce information at the disposal of samplers. Another reason was the fact that enumeration in a systematic sample was easy to organize and data collection costs were predictable. Systematic sampling was not regarded as strictly random, though, and it lacked a theoretical basis. Only in 1944 William and Lillian Madow published a paper (Madow and Madow 1944) where they laid the mathematical foundations of systematic sampling as an exact sampling method. The most important message was that under a few assumptions, a systematic sample could be regarded as a simple random sample.

In the 1920s and 1930s in the U.S., serious attention was given to problems of sampling methodology but the use of random and even systematic sampling procedures in statistical work made slow progress (Stephan 1948). Margaret Hogg, who had worked together with Bowley in some of the surveys (see Chapter 8), came to the U.S. and made a critical study of employment and unemployment statistics. In an article published in the Journal of the American Statistical Association (Hogg 1930), she made a strong plea for rigorous methods of sampling and cast doubt on some surveys in which the sample was selected by judgment rather than random procedures. In the spring of 1931, Hogg had made a survey of unemployment, partly to test the practical difficulties of applying a random sampling method, and also to develop better questionnaires and statistical categories for unemployment surveys (see Hogg 1932).

The spring of 1937 became a turning point, after Neyman had delivered a series of lectures on mathematical statistics and probability at the Graduate School in the Department of Agriculture[134]. During the visit, he also gave a number of

[134] This visit also led to an offer for Neyman to join the University of California at Berkeley. In 1938, Neyman accepted a mathematics professorship and soon after that, he established the Statistical Laboratory at Berkeley. He spent the rest of his life as the director of the laboratory. In Berkeley, Neyman concentrated on other parts of statistical science and contributed very little on sampling theory (see Lehman 2008)

lectures on sampling. A question asked from Neyman after one of these lectures led to the famous paper on double sampling (Neyman 1938). Most survey statisticians already accepted random sampling as a method that would yield representative samples. However, no methods were available how random sampling could be carried out in a country like the United States (or actually in any country). Among other things, Neyman's paper showed that statistically rigorous theory of sampling is attainable at the same time with manageable fieldwork and with acceptable costs. Although Neyman did not have direct contacts with the U.S. Bureau of the Census, his papers and lectures where known and according to Hansen (1987), they strongly stimulated the development of sampling methods. Still at the end of 1930s the Bureau of the Census upheld the idea that it could not undertake sampling surveys because that would discredit its results on other areas; only complete coverage was accepted. (see Hansen 1987 and Olkin 1987).

In the 1930s, the estimates of the number of unemployed persons varied considerably and therefore a study called *Census of Unemployment* was undertaken. It was a nationwide voluntary registration of the unemployed and partially unemployed persons. A questionnaire was delivered to every household in the U.S. with a plea for cooperation, signed by the president. The fieldwork was carried through the U.S. Post Office. Some statisticians, however, foresaw lack of validity in the voluntary registration and they persuaded the administration to conduct the so-called *Enumerative Check Census* in a sample of areas. The Enumerative Check involved an enumeration of a two percent sample of the total population composed of all households within postal delivery routes. The routes were selected by random sampling from all delivery routes. Mail carriers did the practical interviewing in households. They also identified the voluntary mail returns and linked them to interviews. By this arrangement, it was possible to apply ratio estimation which was based on the voluntary registration returns.

The Enumerative Check achieved recognition in the U.S. Bureau of the Census as well as elsewhere in the administration. It showed that large-scale sample surveys could make substantial contributions, and under appropriate design and control, could produce timely information that was more accurate than complete censuses or national registrations. According to Hansen (1987), the success of the Enumerative Check together with the success of the first Gallup Poll were the major events that set strong precedents for the future use of sampling in official statistics.

The Enumerative Check Census has been considered as an immediate consequence of the lectures Neyman gave in Washington and as the step that gave the Bureau of the Census the confidence to use sampling in the 1940 Census (Hansen 1987). The Enumerative Check led to the *Sample Survey of Unemployment*, which was started in 1940 as a monthly activity of the Work Projects Administration (WPA) to measure unemployment (see Frankel and Stock 1941). In 1942, responsibility for the Sample Survey of Unemployment was transferred to the Bureau of the Census, and it was renamed the Current Population Survey.

Inspired by the success of the Enumerative Check, the Bureau of Census introduced sampling as an important part of the data collection method in the 1940 Census. Its purpose was to gather additional information that could not be included in the census schedule (see Stephan, et. al. 1940). This was another success that supported sampling efforts at the Census Bureau (see Hansen and Madow 1976).

12.4 Institutes developing survey methods

According to Hansen and Madow (1978), there was intense pressure in the 1930s to collect information about the population of the U.S. to design policies and to develop social programs. Stephan (1948) identified two important general developments affecting the use of sampling which took place in the United States in 1933. One was the organization of large-scale work projects for the unemployed under the national programs, and the second reason was the enlistment of many leading statisticians from the universities and business in the reorganization of government statistical work (see Stephan 1948 or Olkin 1987). The statistical needs of the government had increased greatly as it tried to solve the problems of the recession and undertook various New Deal programs (see Hansen 1987). At the same time, the advantages of probability sampling in terms of greater scope, reduced cost, greater speed, and model-free features were gradually recognized.

There were actually only two centres in the U.S. which actively developed sampling methods: Iowa State College and the Bureau of the Census. The third important centre was the Indian Statistical Institute.

Iowa State College

Iowa State College at Ames and its Statistical Laboratory had grown in the 1930s a centre for R.A. Fisher's new ideas in the U.S.. Snedecor, the founder of the laboratory, had persuaded the college that all experimental work should be properly treated statistically. This gave statistics at Iowa State College a status it had nowhere else in the world at that time (see David 1984). The emphasis in applied statistics was then on sample surveys and experimental design. In 1938, Cochran visited the college and agreed to return as professor the next year. He lectured on both topics there, and these lecture notes over the next ten years matured into well-known textbooks.

An important contribution to the development of sampling methods at Iowa State College was Raymond Jessen's Ph.D. thesis (Jessen 1942). From that originated the idea of area sampling for estimating farm facts and also rotating sampling. It was an experimental study undertaken to investigate questions "pertinent to the problem of collecting data by the sample survey method". The questions he sought to find answers were: (a) What is the amount and nature of error in data obtained by interview? (b) What is the best available sampling procedure? (c) What method of expanding sample data will provide the best estimate of state or subdivision totals?

Jessen's study was composed of two interrelated surveys on farms in Iowa, one in 1938 and the other in 1939. Approximately 50% of the 1938 sample was re-numerated in the 1939 survey. This design was later used as an example for developing rotating sampling designs, e.g., for the CPS. The sampling of farms in Iowa was based on a grid of approximately 1/4 square mile drawn on a map of Iowa. The sampling unit was a "quarter-section" grid and a county was set up as the stratum. The same proportions of townships were selected at random from each county. Quarter-sections were selected at random from each of the selected townships. The total number of agricultural quarter-sections in Iowa was at the time "about 219 176" and 0.4% of them were selected for the sample.

Enumerators were instructed to visit each farmstead situated on the selected grids to interview "either the operator or whomever might be familiar with the farm's business". Careful instructions were given to enumerators to substitute the farm in case no one was found at home or if the operator was not cooperating. The data collection method resembles that of Kiaer and even more closely that of Bowley. Jessen refers to both Bowley's works and Jensen's works but not to Kiaer's works.

Figure 12.1:
An enumerator interviewing a farmer at the end of the 1930s.
(Source: http://www.census.gov)

One purpose of Jessen's study was to compare the estimation of state totals with three different methods. The methods were based on the knowledge of either the total number of quarter-section grids, total land in farms, or total number of farms in the state. The count of quarter-section was obtained from maps. The total land in farms and the total number of farms was available from the Farm Census. All methods were found to be "not only relatively free from bias but also satisfactorily efficient". In addition, Jessen concluded that the quarter-section grid is an efficient unit "under widely varying circumstances". He also analysed the work of assessors and enumerators and found that there was considerable variation and inaccuracies and concluded that it was an important source of error.

It should be noted that Jessen was also an internationally acknowledged expert in sampling surveys. He conducted population and housing surveys in Peru and Argentina. He was assigned as the leader of a mission of American and British experts to Greece in 1946 to design and execute sampling methods to assess the completeness of the electoral lists for the post-war elections. Partly because of Jessen's international activities, the awareness of the new sampling theory spread all over the world.

The Iowa State College statistical laboratory and the Bureau of the Census were the two main centres in the U.S. where sampling methodology was developed systematically and there was a close connection between them. Later, Cochran was the chairman of an advisory committee to the Bureau of the Census, which was established to assist in designing samples for large-scale surveys (see Watson 1982).

U.S. Bureau of the Census

In the period starting roughly at the end of 1930s, survey statisticians at the Bureau of Census, especially Morris Hansen, William Hurwitz, William Madow, and Joseph Waksberg, made fundamental contributions to sample survey theory and practice. Many of the methods they developed are still used world-wide. These methods for large-scale surveys were designed for data collection from human population by enumerators or interviewers. That was the major difference compared to Cochran's approach in early the 1940s and to the activities at Iowa State College, which were mainly aimed at agricultural research. Hansen claims that the philosophy of Cochran's approach of using analysis of variance to finite population sampling differed from Neyman's original philosophy, which was based on random selection from a finite population (see Olkin 1987). Therefore Cochran's approach was not applicable in the surveys that were planned at the Bureau of the Census.

Indian Statistical Institute

India was the third place where sampling methods were developed significantly in the beginning of 1940s. Hansen claims that in the 1940s, sampling methods for large-scale surveys were actually developed only in two places: the Bureau of Census and the Indian Statistical Institute (Olkin 1987). Statistical Science has a long tradition in India. Already in 1927, Hubback had recognized the need for random sampling and its benefits in crop surveys: "The only way in which a satisfactory estimate can be found is by as close an approximation to random sampling as the circumstances permit, since that not only gets rid of the personal limitations of the experimenter but also makes it possible to say what is the probability with which the results of a given number of samples will be within a given range from the mean." (Hubback 1927). This citation reveals two important observations on random sampling: It avoids personal biases in sample selection, and sample size can be determined to satisfy a specified margin of error (see Rao, J. 2005 and Rao, B. 2006)

Prasanta Mahalanobis[135] established the Indian Statistical Institute in 1931. There, Mahalanobis made pioneering contributions to sampling by formulating cost and variance functions for the design of agricultural surveys. His 1944 paper (Mahalanobis 1944) provides theoretical results on the efficient design of sample surveys and their practical applications, in particular to crop acreage and yield surveys. The optimal allocation in stratified random sampling with cost per unit varying across strata is obtained as a special case of his general theory. As early as 1937, Mahalanobis used multi-stage designs for crop yield surveys villages, grids within villages, plots within grids, and cuts of different sizes and shapes as sampling units in the four stages of sampling (Murthy 1964). He also used a two-phase sampling design for estimating yield.

Mahalanobis developed a great variety of statistical methods for many purposes, not only for agricultural research. One of the most important has been the method of interpenetrating samples whose main purpose is to control and reduce non-sampling errors. One important consequence of the technique was

[135] **Prasanta C. Mahalanobis** (1893–1972) was an Indian scientist and applied statistician. He is best remembered for the *Mahalanobis distance,* a scale-free distance measure. He graduated in physics in 1912 from the Presidency College, Calcuta, and completed Tripos at King's College, Cambridge. He then returned to Calcutta. Inspired by the Biometrika he started his statistical work. In 1924, when he was working on the probable error of results of agricultural experiments, he met Ronald Fisher, with whom he established a life-long friendship. His most important contributions are related to large-scale sample surveys. He introduced the concept of pilot surveys and advocated the usefulness of sampling methods

its simplicity in the estimation of sampling variance regardless of the complexity of the form of the estimator.

P. V. Sukhatme, who studied in the 1930s under Neyman and later worked at the Iowa Statistical Laboratory, also made pioneering contributions to the design and analysis of large-scale agricultural surveys in India, using stratified multistage sampling. He developed efficient designs for the conduct of nationwide surveys on wheat and rice crops. Sukhatme's approach differed from that of Mahalanobis, who used very small plots for crop cutting employing an *ad hoc* staff of investigators (see Rao 2005).

12.5 Development of computing technology

Statistics production is very labour consuming and surveys are no exceptions. For example, it took seven years to complete the 1880 census in the U.S. Therefore, there was some concern that the 1890 census would not be completed before the 1900 Census. The problem was not in the collection of the data but in the processing of the data. The U.S. Bureau of Census rented tabulator machines invented by Herman Hollerith[136] for the 1890 Census. Their influence was dramatic: in a few months, an unofficial estimate of the United States population was obtained. The 1890 Census was completed in five years, although it was a far more extensive census than the previous one (see Bellhouse 2000). The advantages of Hollerith tabulation machines were soon noticed also outside the U.S. For example, Anders Kiaer used them for tabulation of his first survey.

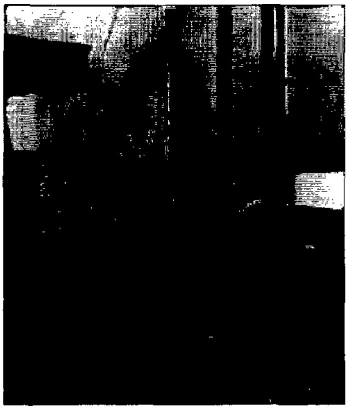


Figure 12.2:
Hollerith tabulating machine in use in 1902 at the U.S. Bureau of Census. (Source: http://www.census.gov)

The ISI conference in Bern in which Kiaer presented his Representative Method is historically interesting also in another respect: Heinrich Rauchberg presented a paper on the use the electric machine in the Austrian census (Rauchberg 1896). After that, Herman Hollerith took the floor to explain how he wanted to develop his machine further (Malaguerra 2000). The audience immediately recognised the utility of the machine. In particular, the French statistician Emile Cheyson predicted that the Hollerith machine would bring to statistics the same transformation as the introduction of mechanics did in industry (Malaguerra 2000).

The possibilities to carry out a large-scale survey increased essentially at the end of 1930s and in the 1940s because of the fast de-

[136] Herman Hollerith was an employee of the Census Bureau when he built the first mechanical tabulator. In 1896, he established a company, which was renamed International Business Machines (IBM) in the 1920s.

velopment of computing techniques. Due to fast punched card tabulators and calculators – and new algorithms – the results of large-scale surveys could be obtained in such a short time that they could be utilized effectively. During the 1930s, the Journal of the American Statistical Association included several articles about different algorithms for punched card calculators and, for example, about the frequency of punching errors. Grier[137] claims that without the advanced computing techniques, modern statistical methodology could easily have languished as an interesting theory, useful for small problems but otherwise impracticable. Especially in the data processing of a large-scale survey, computing technology is of central importance. Even fairly simple technical devises, in a modern sense, can take care of a great deal of handwork required in tabulating survey data. However, standard errors were seldom calculated in the 1930s due to the amount of work, which increased rapidly with the increasing sample size, and when they were calculated, the correct formulas were seldom applied (see Bellhouse, ibid.).

Several statistical laboratories developed computational methods in the 1930s, but Iowa State College was more productive than others were, partly because the Department of Agriculture financed their research (see David 1984 and Grier). In 1940, they managed to build a device which had the typical parts of a computer but it was never taken into use. A few years later, its central ideas were included in the first computer.

The first computer in the modern sense was the ENIAC, unveiled in 1946. Its capacity was very limited, but it already had the components which make a computer. ENIAC was built for military purposes, however. The U.S. Bureau of the Census received the first UNIVAC computer in 1951[138]. At that time, it was an efficient computer which used magnetic tape to store input/output rather than the punch tape or punch cards. UNIVAC I was first used for processing the 1950 census data. When the census was completed, the computer was used to process CPS data. The use of an efficient computer set new standards for the production of statistics, and its arithmetic capabilities enabled the use of significantly more complex estimators (and sampling designs) than before. Only after the computer was received, technology caught up to theory and it became possible to calculate standard errors for estimates, though still through approximations in the beginning (see Bellhouse 2000).

12.6 Formal development of sampling methods

Several people contributed to the development of modern sampling theory in some way. However, the most significant contributions came from a few men: Jerzy Neyman, William Cochran, Morris Hansen, William Hurwitz, and Wil-

[137] David Allan Grier's article "The Origins of Statistical Computing" is published on the Web site of ASA and has no other reference information than its address (see http://www.amstat.org/about/statisticians/index.cfm?fuseaction=papers).

[138] The initiative to design and build UNIVAC originated partly from the Census Bureau.

liam Madow (and the team at the U.S. Bureau of the Census), and in India, P. C. Mahalanobis and P. V. Sukhatme. One should note that the classical sampling theory to a great part was aimed at surveying human populations, either individuals or households.

12.6.1 Cochran's contributions to sampling theory

Cochran[139] is the author of probably the most frequently referred textbook on survey sampling (Cochran 1953) but he originally specialized in agricultural field experiments. He worked from 1934 to 1939 at the Rothamsted Experimental Station, which fostered Fisherian statistical methods on design of experiments (see Watson 1982). Before the publication of his book on sampling, Cochran wrote only three papers that dealt strictly with sampling methods (see Watson 1988). All the papers concerned sampling in agricultural research.

Cochran wrote his first paper on sampling (Cochran 1939) while he was still affiliated with Rothamsted Experimental Station. He read the paper at the 100th annual meeting of the American Statistical Association (ASA) in 1938. The paper contains several results which have later proved to be important: the use of analysis of variance model to estimate the gain in efficiency due to stratification, estimation of variance components in two-stage sampling for future studies on similar material, choice of sampling unit, regression estimation under two-phase sampling, and the effect of errors in strata sizes.

In this paper, Cochran referred to Neyman's 1934 paper only briefly mentioning that Neyman had shown that purposive selection would rarely give a representative sample. Next, Cochran stated, "A representative sample ... can clearly be obtained by giving every unit in the population an equal chance of being included in the sample". This is the only occasion when Cochran indirectly referred to randomization. He did not discuss the role of randomization in statistical inference at all. It seems that Cochran considered it as an established fact which did not require contemplation anymore.

In this paper, Cochran introduced the superpopulation concept in sampling theory (he did not use the word "superpopulation", though):

> "The finite population should itself be regarded as a random sample from some infinite population; thus the sample which is taken for enumeration is regarded as a subsample for a larger sample of the same infinite population." (Cochran 1939)

At that time, Cochran had critical views about the traditional finite population concept:

[139] **William Cochran** (1909–1980) studied at St. John's College in Cambridge. In 1934, he was hired as an assistant to Rothamsted, where he worked until 1939 writing papers on design of experiments or theoretical papers and one paper on sampling. In 1938, he visited Ames, Iowa (U.S.), and agreed to return in 1939 as professor. At Iowa, Cochran produced central ideas for survey sampling. In 1943–1944, Cochran joined the Princeton Statistical Research group. From 1949 to 1957, Cochran was at the Department of Biostatistics in the School of Hygiene and Public Health at John Hopkins. The Department of Statistics was established at Harvard University in 1957, and Cochran was appointed as a professor. He remained at Harvard the rest of his career.

> "Where the population consists of a single group, the results obtained by 'finite sampling theory' agree with those obtained by the analysis of variance. The former is, however, not easily extended to the case in which the population is subdivided into groups [Cochran's expression for stratification] at least so far as the situation arising in practice are concerned. Further, it is far removed from reality to regard the population as a fixed batch of known numbers. In economic and sociological studies the population is changing from day to day. The population at any time is often conventional, as for example with a population of farms or carpenters, owing to difficulty in defining a member of population. Errors in counting are bound to occur in any large-scale investigation and though they are not usually differentiated from the sampling errors, they will contribute to inaccuracy in any means which are calculated." (Cochran 1939)

Cochran's next paper on sampling, published in 1942, was written while he was already associated with Iowa State College (Cochran 1942). The paper was based on a presentation he gave in 1941 at the 103th annual meeting of the ASA. In the same meeting, several other well-known papers on sampling were also presented, such as "*Relative efficiency of various sampling units in population inquiries*" by Hansen and Hurwitz (Hansen and Hurwitz 1942) and a presentation by Frankel and Stock concerning the plans for "*the sample survey of unemployment*" by the WPA (Frankel and Stock 1941). Unfortunately, there are no records on discussions in this conference but probably sampling and survey methods were touched on.

The main question in Cochran's paper was: Should differences between the sizes of the sampling units be ignored or taken into account in selecting the sample and in making estimates from results of the sample? Cochran was thinking populations where sampling units differ in size, such as farms "which in the same county may vary in land acreage from few acres to over 1, 000 acres". In mathematical terms, he stated the problem of estimation in the following manner: Sampling units are drawn at random without regard to their sizes. How to estimate the population total of a quantity y, which can be measured on each sampling unit. Associated with each sampling unit is also a quantity x, which is called its area (Cochran used "area" to avoid confusion between "size of sample" and "size of sampling unit"). He assumed that some knowledge was available about the values of x in the sample, and possibly also in the population. In addition, Cochran assumed that the number of sampling units in the population may be considered infinite. In the footnote, Cochran noted, referring to Hansen's and Hurwiz' paper, that his model does not apply in sampling from a human population where the sampling unit is a household and a sub-unit is a person.

As a solution, Cochran developed a regression estimator which later gave the stimulus to develop general regression estimators (Särndal 2007). He showed that when the mean value of y is linearly related to the area of the sampling unit, with constant variance, i.e., linear regression $y = \alpha + \beta x + e$, where e has mean zero and constant variance, the linear regression estimate for population total, Y_l, of y is

$$Y_l = N\left\{\bar{y}_s + b\left(\bar{x}_p - \bar{x}_s\right)\right\} \tag{12.1}$$

where N is the number of sampling units in the population, b is the sample regression coefficient, and the suffixes p and s refer to population and sample, respectively. This estimator requires knowledge of the total number N of sampling units and the mean value of x in the population. In human population, these parameters were rarely known, which is a serious limitation.

Cochran derived the average bias under model deviations for simple random sampling as the sample size n increased, and also extended the results to weighted regression and derived the now well-known optimality result for the ratio estimator; namely, it is a "best unbiased linear estimate if the mean value and variance both change proportional to x ".

In this paper, Cochran introduced ratio estimation for sample surveys as a new method, although an early use of the ratio estimator dated back to Laplace (1774) and Graunt (1662). It is interesting that Cochran obviously was not aware of what Laplace had done earlier. Almost forty years later, Cochran published a paper where he appeared surprised at the existence of Laplace's survey and analysed Laplace's estimator from the current perspective (Cochran 1978).

The 1942 paper of Cochran is interesting from the perspective of sample selection, as well. The starting point was that sampling units were drawn at random without regard to their sizes, i.e., all sampling units have equal probabilities of inclusion. At the end of the paper, Cochran noted that the regression estimator remains unbiased under non-random sampling, provided the assumed linear regression model is correct. He concluded that "Thus the large sampling-units might be allotted a greater chance of inclusion in the sample, this procedure giving a more accurate estimate whenever the variance of y increases as x increases. On the other hand, if the method of selection discriminates in favour of certain sampling-units amongst those of the same area, bias may arise." Cochran did not further elaborate on the idea of drawing a sample with varying inclusion probabilities.

In the next paper on sampling, Cochran (1946) compared analytically the relative efficiency of alternative probability sampling strategies: systematic sample of every k^{th} element, a stratified random sample with one element per stratum, and a random sample. Cochran referred to the first investigation of the properties of systematic samples by W. and L. Madow (Madow and Madow 1944). Also, this study was carried out, in modern terms, under a superpopulation model. The object of the paper was to make comparisons for population in which the variance among the elements in any group of contiguous elements increases steadily as the size of the group increases. Cochran claims that this type of population has been regarded as applicable in field experimental work, where the variance among plots within blocks increase with the size of the block. This class of populations, according to Cochran, could be represented by a model in which the elements are serially correlated. He defined the population in the following manner:

> Elements of the population, x_i, $i=1, \dots, nk$, (e.g., n strata and k elements in each) are assumed to be drawn from a population in which
>
> $$\begin{aligned} &E(x_i) = \mu, \\ &E(x_i - \mu)^2 = \sigma^2, \\ &E(x_i - \mu)(x_{i+u} - \mu) = \rho_u\sigma^2, \rho_u \geq \sigma_v \geq 0 \end{aligned} \qquad (12.2)$$
>
> whenever $u < v$ and ρ_u is a serial correlation between x_i and x_{i+u}.

Cochran continued, specifying that "It is more reasonable to regard the finite population as being itself a sample from an infinite population in which the ρ's are monotone. ... Thus, comparison between the systematic and stratified random samples will be made not for a single finite population but for the average of finite populations drawn from an infinite population with monotone decreasing ρ." (Cochran 1946). As a result, he showed that the stratified random sample is always at least as accurate on the average as the (simple) random sample, and its relative efficiency is an increasing function of the size of the sample. However, he found no general results which were valid for the relative efficiency of the systematic sample. Rao (2001) argues that this paper has stimulated much subsequent research on the use of superpopulation models in the choice of probability sampling strategies.

12.6.2 Hansen's and Hurwitz' sampling design for the CPS

Before the development of sampling design for the Current Population Survey (CPS), Hansen and Hurwiz analysed the problems of stratification in sampling in the context of census (Hansen and Hurwitz 1942). The main problem was: which sampling units make the most efficient stratification in terms of cost and administrative limitations? The sampling units were different types of clusters, such as an individual person or household, or a small geographical area such as a city block, a segment of a block, a group of blocks, and a small rural area. The criterion was the relative efficiency in terms of relative magnitudes of sampling variances computed on the same unit basis. The analysis indicated that for most population and housing items, a large size sampling unit is considerably less efficient than a small one.

Another significant observation made by Hansen and Hurwitz (ibid.) was that there may exist a correlation between the elements within clusters, and this intra-cluster correlation influences the sampling error. They noted that usually in practice the intra-cluster correlation is positive and therefore the sampling of clusters is less efficient than the sampling of individuals, but many important exceptions could be found in which intra-cluster correlation was negative. Hansen and Hurwitz (ibid.) compared their results to the models proposed by Cochran, and Yates and Zacopanay and noted that these models do not permit a negative correlation and hence clusters can never be more efficient than a single individual. Hansen and Hurwitz (ibid.) showed that this limitation is not realistic in sampling from human populations where negative intra-cluster correlations are frequently observed.

After the Sample Survey of Unemployment was moved to Bureau of the Census Statistician, Hansen and Hurwitz started to develop a new sampling design for it, based completely on probability sampling (Olkin 1987). The result of this work is best documented in the paper entitled "On the Theory of Sampling from Finite Populations" by Hansen and Hurwitz (1943).

A significant condition for the development of the sampling design was presented by the costs of data collection. Also, the small amount of population-level information set limits to potential methods. Hansen and Hurwitz concluded that "these formulas have not practical utility unless there are also some considerations on differential costs." As one of the starting points, the authors explain what their motivation to develop the method was and what difficulties it was supposed to solve:

> "If, no matter how a sample be drawn, the costs were dependent entirely on the number of elements included in the sample, here would be no need for theory beyond the classical theories of Bernoulli and Poisson covering the independent random sampling of elements within strata, supplemented by the extension of the theory to finite populations, and the extension to optimum allocation of sampling units. Very often, however, in statistical investigations it is extremely costly, if not impossible, to carry out a plan of independent random sampling of elements in a population. Such sampling, in practice, requires that a listing identifying all the elements of the population be available, and frequently this listing does not exist or is too expensive to get. Even if such listing is available, the enumeration costs may be excessive if the sample is too widespread. Frequently also, there are other restrictions on the sample design, such as the requirement that enumerators work under close supervision of a limited number of supervisors, and as a consequence the field operations must be confined to a limiting number of administrative centers..." (Hansen and Hurwitz 1943)

Hansen and Hurwitz aspired to develop a method that would make the most effective use of available resources by organizing adequate field operations. In the background, there was also the problem that sampling was still a novel method and not generally accepted as trustworthy (see Hansen and Madow 1976).

The structure of the population which Hansen and Hurwitz (1943) wanted to study was the following: It is made up of L strata, with the i-th stratum containing M_i primary sampling units (PSU) of N_i elements each. X_{ijk} is the value of some characteristic of the element k of PSU j in the stratum i. The population average to be estimated is

$$\bar{X} = \sum_{i}^{L}\sum_{j}^{M_i}\sum_{k}^{N_i} X_{ijk} \Big/ \sum_{i}^{L} M_i N_i \qquad (12.3)$$

This can mean, for example, that the parameter to be estimated is the average income of households, $\bar{X}$, in a given city; X_{ijk} is the income of k-th household in the j-th city block in the i-th ward. If the sample consists of m_i PSUs from i-th stratum and n_i elements from each PSU, the "best linear unbiased estimate", in the sense Neyman defined it, is

$$\bar{X}' = \sum_{i}^{L} \frac{M_i N_i}{m_i n_i} \sum_{j}^{m_i}\sum_{k}^{n_i} X_{ijk} \Big/ \sum_{j}^{L} M_i N_i \qquad (12.4)$$

This estimate would be the most efficient if the number of sampling elements in each sampling unit within a stratum were the same. A problem was that the numbers of elements, M_i and N_j, are not always known. Hansen and Hurwitz (op. cit) concluded that if the numbers of elements differ between sampling units, a biased but consistent estimate can be found which has a smaller mean square error[140] than the best linear estimate. As an example, they showed that a ratio estimator $\bar{X}' = \sum_{i}^{m} X_i \Big/ {}_{N}\sum_{i}^{m} N_i$ has a smaller mean square error than $\bar{X}' = \left(M/_{m} \sum_{i}^{m} X_i \right) \Big/ N$ where $N = \sum_{i} N_i$. The ratio estimator is biased (with negligible bias) and nonlinear, but consistent. On the other hand, by using the ratio estimator, the authors could avoid the problem that unbiased estimates required the knowledge of N, which was rarely known.

After a thorough analysis of the circumstances and various comparisons between different approaches, Hansen and Hurwitz (1943) developed the basic theory of stratified two-stage sampling with one primary sampling unit (PSU) within each stratum drawn with a probability proportional to the size measure (PPS sampling) and then sub-sampled at a rate that ensures self-weighting within strata. PPS sampling design led to significant variance reduction by controlling the variability arising from unequal PSU sizes, without stratifying by size and thus allowing stratification on other variables to reduce the variance.

The design they ended up with was based on a selection of primary units with probabilities proportionate to the measure of their size. The actual scheme was to sample one primary unit per stratum without replacement. The current sampling design of the CPS described in Chapter 1 still follows the basic ideas of this original contribution, although it has been developed considerably.

Unequal inclusion probabilities were implicitly present already in Neyman's optimal allocation designs, although he did not pay attention to that. The prevailing paradigm at the end of the 1930s was that each element of the population should have equal inclusion probabilities. In Hansen's and Hurwitz' (ibid.) method, the possibility to have varying probabilities in selecting sampling units was explicitly articulated for the first time.

The complex sampling designs were planned to solve the practical problems of data collection in such a manner that estimators with acceptable (not necessarily maximal) accuracy could be produced. Aside from the theoretical breakthrough, one great difficulty could be avoided by the approach presented by Hansen and Hurwitz: it provided approximately equal interviewer workloads, which is important in designing the field work. It was central that a manageable field organisation could be established for data collection and a reasonable amount of field work was required to provide estimates.

12.7 Theory of statistical inference in the 1940s

It is obvious that the classical sampling theory was outlined in the 1940s in the United States at Iowa State University and even more at the Bureau of the Census. The one single factor behind the development was the need for an ac-

[140] This is probably the first time when mean square error is taken as an equal criterion with sampling variance in developing a sampling design.

curate and cost-efficient sampling design for the Current Population Survey. The characteristic feature of the development was that it aimed at the sampling of human populations. The main part of the classical sampling theory deals with sampling from finite human populations with a design-based approach.

The main aim in Hansen's and Hurwitz' works in early the 1940s was to develop a data collection apparatus that provided statistical data for the administration with acceptable costs and acceptable accuracy. The theory they developed allowed for the designing of complex multistage sample surveys, which have become the basis of large-scale social and economic surveys all over the world.

In the 1940s, Cochran published three frequently cited papers on sampling. All of them were related to agricultural research, not sampling in human populations. The approach in these articles was based on the idea of a superpopulation from which the observable finite population is a sample. Cochran's original approach can be seen as an offspring of Fisher's ideas applied to finite populations. Fisher defined a population to be hypothetical and infinite, "the resultant of the conditions we are studying", and he regarded a sample as drawn from a distribution *f(x)*. However, Cochran's approach in sampling was distribution-free. In addition, the methods were elaborated from Fisher's methods for experimental research, such as analysis of variance and regression analysis. In modern terminology, Cochran's approach in his papers in the 1940s was model-based. Only later did Cochran develop methods for finite and fixed populations, and in his famous textbook (Cochran 1953), he deals with sampling from finite (human) populations and his approach is design-based.

Another noteworthy feature in Cochran's early papers was that they were not related to Neyman's ideas on sampling. In the three papers referred to here, he mentions Neyman only once. The adoption of randomization is due to the work at Rothamsted and more of Fisher's influence.

An important breakthrough in the classical theory of survey sampling theory was when Horvitz and Thompson published a paper in 1952 (while affiliated with Iowa State Statistical Laboratory) on a general theory for constructing unbiased estimates (Horvitz and Thompson 1952)[141]. Hansen and Hurwitz (1943) obtained results on sampling with probability proportional to size and with replacement. Horvitz and Thompson extended this idea to sampling without replacement.

Horvitz and Thompson obtained the impetus for their work from the method Hansen and Hurwitz presented in the 1943 paper. One of their examples was "...entirely analogous to that specified by Hansen and Hurwitz ... when a single primary unit is drawn with probability proportionate to its estimated size". One problem they aim to solve was the limitation in the Hansen and Hurwitz design that an unbiased estimate of sampling variance of their estimator could not be obtained from the sample. Horvitz and Thompson (ibid.) provided a general method for sampling without replacement from a finite population with variable selection probabilities. They also gave an unbiased linear estimator for a population total and the sampling variance of the estimator.

[141] Narain derived the same formulas (Narain 1951) nearly at the same time as Horvitz and Thompson, but Narain's contribution became known many years later.

Horvitz and Thompson (ibid.) treated a population, U, consisting of N elements u_1, u_2, ... u_N. A sample of size n is drawn from the population without replacement, using arbitrary probabilities of selection. A probability of selection, $P(u_i)$, is predefined for each element u_i in the population, and it plays a central role in the further development in HT estimation. This was first time selection probabilities were explicitly included in estimation. A fundamental element in their method was also the probability distribution that the sampling design induces over all potential samples, s_i, of size n. This setup became a central element in the development of statistical inference for finite populations.

The scope of the current thesis is the early history of survey sampling up to the 1950s. In a way, it can be stated that Horvitz and Thompson completed the classical theory of survey sampling. The random sampling approach was almost unanimously accepted and HT estimators opened new ground for development of survey sampling theory for different applications. In Kuhn's terminology, the theory of statistical inference developed by Fisher and Neyman had reached the state of normal science. It was a paradigm that was accepted by most survey statisticians; it was an inherent part of basic training in statistics.

There has been a lot of development in the classical theory after the paper of Horvitz and Thompson. After the mid-1950s, discussion started on the basics of statistical inference and challenges to the random sampling approach appeared, but that discussion will be excluded from this thesis.

The classical books about statistical sampling theory were also published roughly at the same time (Cochran 1953, Hansen, Hurwitz and Madow 1953). The book by Hansen, Huwitz and Maddow was an offspring of the work the authors did at the Bureau of the Census (see Olkin 1987). Hansen's and Hurwitz' contribution, especially the paper published in 1943, has to be regarded as a watershed in sampling theory. After that, the classical theory for finite populations started to attain its current form. The approach of Cochran in his famous textbook (Cochran 1953) is related to the ideas Hansen and Hurwitz presented, and it is essentially different from his articles in the 1940s.

In the 1940s, the principles for randomization inference which Neyman presented in the three papers in the 1930s were already accepted unanimously. It is remarkable that statistical inference was explicitly dealt with in none of the papers published in the U.S. in the 1940s. Cochran, Hansen and Hurwitz, and Madow mention only occasionally the inference principle of drawing repeated samples from the same population and very seldom mention confidence limits. It seems that the problems of statistical inference were regarded as solved and there was no need to return to that question anymore. Neither is there any discussion on the principles of induction, which had caused a bitter dispute between Fisher and Neyman.

13 Summary and discussion

Today, sampling techniques are essential tools for national statistical offices. Probably the administrations in all democratic societies use information that has been obtained by surveys. In addition, social scientists all over the world are deeply dependent on survey data, as well as researchers in many other areas, such as social medicine, political science and marketing. In the modern world, sample surveys have an irreplaceable influence on the increase in knowledge for science, management and marketing. The prevailing sampling techniques for human populations were created in a relatively short period during the 1940s and 1950s. In a way, the classical theory reached its culmination in Horvitz-Thompson estimators (Horvitz and Thompson 1952). This theory was documented in two well-known books (Cochran 1953, Hansen, Hurwitz and Madow 1953). However, this period was preceded by a much longer, diversified and even fumbling period of a search for methods that could be generally accepted.

Sampling techniques involve two different but strongly connected and equally important tasks: drawing a sample from a population and calculating estimates for population parameters from the sample. Drawing the sample includes two phases: (1) the selection of sampling units from the population or from a sampling frame representing the population, and (2) the enumeration or collection of data from the selected units. Estimation methods are mathematical, based on probability theory, but sampling (especially enumeration) deals with practical problems. Even though sampling and estimation methods are strongly interrelated in the current theory, which has not always been the case. The early development of estimation and sampling methods followed different paths with different paces (see also Smith 1976).

Due to the practical problems of data collection, the development of sampling techniques has been interplayed between what is realizable in practice and what is mathematically tractable (see also Rao 2005 and O'Muircheartaigh 2005). Especially the practicalities of data collection (including its costs) and the possibilities for data processing have set limits to what has been regarded as viable for estimation methods.

When the history of modern survey research started at the end of 19th century, the infrastructures of society were less developed than now. The only realistic mode to collect data from households was by sending enumerators to visit them. Statistical sampling theory did not exist; consequently, the sample sizes were decided intuitively - and they became very large. For example, Kiaer's first sample consisted of 120,000 respondents (Kiaer 1895). Another consequence was that the calculation of estimates was based on intuitive ideas without a theoretical (mathematical) basis. In order to obtain reliable estimates from a sample, it was necessary that the sample was a miniature of the population.

Random selection was known to be an advantageous sampling method because it would provide representative samples, but the only known method to apply random selection required that all population units have equal inclusion probabilities, i.e., simple random sampling. A face-to-face enumeration of a large random sample leads to a costly and hard-to-handle data collection. Therefore, the first sampling designs were carried out with some sort of cluster or area sampling; their design was based on the latest census data. Clusters were purposely selected to ensure the "representative nature of the sample". After the data collection, the representative nature was verified against the data from the census. The selection of households within clusters was haphazard. Systematic selection in some form was the most common method.

The idea of representative sampling was presented at the end of the 19th century, but the mathematical sampling theory for it was formulated only at the beginning of the 1940s. For the first time, the mathematical theory was adapted to an operational data collection scheme in the U.S. Bureau of the Census. In the 1930s in the U.S., survey sampling was a popular topic among statisticians, but a lot of the development aimed at agricultural research. Hansen and Hurwitz took the critical step when they developed the sampling design for the Current Population Survey. In that design, they merged Jerzy Neyman's sampling theory and a data collection plan in which they applied the ideas of the Representative Method (Hansen and Hurwitz 1942, 1943).

The mathematical theory got its greatest impact from Neyman's three papers in the second half of the 1930s. The ideas presented in the third paper (Neyman 1938) led to a theory that made survey data collection from a human population possible in a large and diversified country like the United States. The method was explicitly based on varying inclusion probabilities. Implicitly unequal inclusion probabilities were already present in Neyman's first paper on sampling (Neyman 1934), but it was not recognized. Up until the 1940s, the general conception was that in random sampling, all population units should have equal inclusion probabilities.

The collection of data has not been the only practical problem in survey research. The facilities available for data processing[142] have also significantly influenced the development, more than has been recognised. In the early days, data processing used to be the most laborious phase of a survey, and even still in the 1940s, it set limitations on the development of sampling theory (see Cochran 1942 and Bellhouse 2000). Only in the 1950s did computers become available for surveyors, enabling statistical methods to be developed without concern for computational constraints. For example, the calculation of Horvitz-Thompson estimators (in which all or most sampling units have different inclusion probabilities) is not possible without an electronic computer. Even with a relatively small sample, computations by other means would become too labour-consuming and would take too much time to complete.

In the 19th century, the calculation of the results of a census could take several years. For example, in the United States, it took seven years before the 1880 census results were published, the greatest problem being its data processing. Because of complex estimators and estimator variances, the data processing in survey research is more demanding than in a census, despite the fact that sur-

[142] Data processing involves all those tasks by which data from sampling units are handled to produce statistical tables, such as the logistics in handling of questionnaires, data entry, editing, and calculation of estimates and calculations of standard errors. All these phases have not been necessary or possible throughout the history of survey research.

veys include fewer observations. Survey research in its current form and extent would not be possible without efficient computers.

13.1 The emergence of representative sampling

Modern sampling techniques were first introduced in social research for the needs of statistical institutes. However, partial investigations in agricultural research have a different and longer history, and they partly served as examples of applications for human populations. The characteristics of human populations are inherently different from those in agricultural research. Therefore, it was not immediately clear how to transfer the methodology. It required new ideas and new thinking before partial investigations in human populations could be considered. It was especially necessary that the law-like regularity and stability in these populations were discovered so that inductive generalisations were justified.

13.1.1 Birth of statistical thinking

Statistics and statistical thinking appeared in the course of the 19th century, at the end of the era of industrialization and urbanisation. The first national statistical institutes were established at the very end of the 18th century and during the next century. Practically all European countries had established one. In the first half of the 19th century, informal activity in statistics also became frequent. For example, a number of statistical societies were established and many statistical journals were started. Westergaard (1932) called the middle of the 19th century an era of enthusiasm and concluded that "everybody seemed to have statistics in the brain".

In the 19th century, the decennial censuses became the central activity in national statistical institutes. Eventually, the efforts of the institutes and societies produced an avalanche of printed numbers, as Hacking (1990) called it. For the first time in history, there was a large amount of information on the population structures, so changes in these structures and in social phenomena could be followed. In the International Statistical Conferences, organised between 1853 and 1876, the representatives of national institutes agreed on the harmonisation of statistics, which eventually resulted in comparable statistics among the European countries.

Adolphe Quetelet utilized the outburst of statistics, and in 1835, he published his first book, *"Social Physics"* (*"Physique sociale"*), which included a large number of tables on vital data, moral and criminal statistics, and anthropometry (Quetelet 1835). In the tables, Quetelet described the distributions of variables, which without exception were similarly bell-shaped. Moreover, he showed that these distributions were similar in different countries and stable over time. In this way, he showed that there was stability within social phenomena and that there was a regularity, or invariance, which could be called social law[143].

Quetelet thought that the statistical regularities were evidence of determinism. He argued that basically human beings were aimed to be similar and the distribution of characteristics was essentially an error distribution. Quetelet's error distribution was the error law that Laplace had derived and which Galton later called the Normal Distribution. Quetelet was the first social scientist to apply Laplace's error law to social phenomena, but he had several followers who fostered the idea. Using this reasoning, Quetelet developed the famous concept of the average man, "*l'homme moyen*".

By his analyses, Quetelet showed that the seemingly chaotic mass of observations actually followed a manageable distribution. Quetelet argued that if observations are investigated from a distance, it is possible to develop a science of collective phenomena: by losing sight of individuals, one can discover, through the social phenomena that dominate the masses, a set of laws (Quetelet 1835, 1848).

Empirical social research has been said to begin with Quetelet's works. He called the science 'social physics', which is often considered the origin of modern empirical sociology[144]. Quetelet published several books touching on the same topic (e.g., Quetelet 1848 and 1869), which subsequently inspired many scientists to develop new theories, thus generating a tradition of statistical research in social phenomena.

Quetelet's works inspired many of his contemporaries to further analyze social phenomena. For example, Ernst Engel observed that the proportion of a consumer's budget spent on food tends to decline as the consumer's income goes up. This so-called 'Engel's law' is said to be the first established social invariance. Another follower of Quetelet, Wilhelm Lexis, has been important for the emergence of statistical science because of his pioneering work on dispersion. Lexis' main topic was the development of mathematical methods in research on the stability of statistical series. Lexis had a significant influence on Edgeworth's thinking, and his analysis of dispersion has also been claimed to foreshadow Fisher's analysis of variance.

Partial investigations have a different story, however. In 1802, Laplace carried out a survey to estimate the population of France. It took nearly a century before the next partial investigation on a human population was undertaken. The reason was the commonplace disbelief in the homogeneity and regularity of human populations. In the 1830s, Quetelet planned to carry out a similar estimation of the population in the Low Countries, as Laplace had done in France. Baron De Keverberg criticized his plans saying that the sample could not reach representativeness because of the fundamental heterogeneity of the population, and the attempt to overcome the problem would divide the country into nearly

[143] Behind Quetelet's idea was his aim to show that laws similar to the laws of nature also govern social life. However, many of Quetelet's contemporaries did not accept the idea that regularities could be interpreted as laws.

[144] Sociology is usually regarded as a creation of the French philosopher **August Comte** (1798–1857), but he did not accept the statistical approach. Therefore, empirical sociology is usually dedicated to Quetelet.

as many sampling units as there were people (see Stigler 1986). Quetelet did not undertake any partial investigations in his lifetime.

The central point in De Keverberg's critique was that birth and death rates were not constant and hence the stability of statistical ratios could not be assumed and the urn model could not be used as an inference model as Laplace had done. In the absence of homogeneous groups, there could be no reliable inferences or inductive generalizations from a part to the whole. Opposite views also existed, but views like De Keverberg's were predominant, hindering partial investigations of human populations (see Stigler 1986). In 1911, Yule was still hesitant about the feasibility of random selection from human populations because of their heterogeneity (Yule 1911).

Partial investigations became feasible only after reliable and comprehensive statistics about the population were available – and after the regularity and stability of social phenomena had been generally accepted. Hacking (1990) considered that the avalanche of printed numbers was the essential condition for the unveiling of statistical regularities. At the end of 1800s, the subsequent censuses had yielded results, which showed consistently stable population characteristics from census to census; or if changes took place, they were regular and predictable. Censuses also helped the survey research in another way: information on the population structure has to be available in order to be able to draw a representative sample from it. That happened only after censuses had become commonplace and their results available.

13.1.2 Kiaer's Representative Method

It is generally held that the Representative Method, which Anders Kiaer set before the International Statistical Institute (ISI) in 1895, is the starting point of the modern data collection methodology for the survey research. Kiaer brought the issue in the agenda of the ISI, and in his talk, he described how he had used the method in a social survey in Norway. Lie has analysed minutely the history of why the method abruptly appeared in Norway (Lie 2002). In this, the matter at issue is research on human populations. Partial investigations and representative methods have a much longer history in agricultural surveys (see Didier 2002).

However, Kiaer obviously was not the first to use the method, not even in Norway (see Lie 2002). A similar method on a smaller scale had been applied earlier in Denmark (Jensen 1926). In addition, Kruskal and Mosteller (1980) noted that Carroll D. Wright had used a similar method in the United States, and Mespoulet (2002) claims (referring to Kaufmann (1922)) that A. Kaufmann already carried out a sampling survey in Russia between 1887 and 1890. In that survey, sampling was based on random selection. Chang (1976) argues that Russia was the first centre of the modern mathematical theory of sampling at the end of the 19th century. Both Zarkovic (1956, 1962) and Seneta (1985) claim that sample surveys were an extensively studied branch within statistical science in Russia in the end of the 19th century.

Nevertheless, Kiaer was one of the first to use the Representative Method independently from a census and in an important – and extensive – investigation. Most importantly, Kiaer's initiative eventually led to the acceptance of the

Representative Method by the ISI as a valid data collection method for official statistics. This, in turn, stimulated further development and new applications in social research, and finally ending up as a new branch of statistical science. Kiaer's first public appearance and the discussion that followed it was a critical turning point in the history of sample surveys on human populations. In addition, as Kruskal and Mosteller (1980) noted, Kiaer was the first man ever to use *analytically* the term *'la Méthode Représentative'*. Carroll D. Wright had used the same expression earlier, but according to Kruskal and Mosteller (ibid.), the term they used was so shallow (and used in a less influential way) that it cannot be considered as the starting point.

The idea in Kiaer's Representative Method was to form a sample that is a miniature of the population. This was in contradiction with the Monograph method, which was frequently applied at that time. In a Monograph survey, only typical cases were studied, and all extreme cases were discarded. In the Representative Method, it was essential that the distributions in the sample were close to the distributions found in the population. In a Monograph survey, distributions were not important.

In Kiaer's method, representativeness was ensured by the selection of the sample so that it covered – geographically, demographically, and economically – all characteristics of the country. In modern terms, Kiaer's 1895 sampling design can be described as a multi-stage stratified area sample with systematic sampling of households at the final stage in the urban areas. In rural areas, the final stage data collection had to be organised differently. Enumerators were instructed to follow distinct routes and while doing so, to visit houses of different types in the same neighbourhood, and in particular check that not only typical middle-class houses were visited but also the more well-to-do and the poor-looking houses, both for families and single persons.

The definition of strata was purposive, or rational, as it still is in modern sampling practice. Kiaer formed the strata based on the location and the type of municipalities (urban or rural, type of industry). After data collection, he verified that the sample truly could be regarded as a miniature by comparing the demographic data of the sample with the census data.

The reasons why Kiaer used this sampling design were practical: an enumeration of a genuine random sample had been nearly impossible to carry out because of the huge sample size (80,000 + 40,000 households). Only modern sampling theory has shown that dramatically smaller samples can be sufficiently accurate, but only simple random sampling was known in Kiaer's time.

When Kiaer presented the idea of the Representative Method for the first time in 1895 in the ISI general assembly, it was harshly criticised by the foremost statisticians in Europe. Its further handling in the forthcoming ISI meetings was accepted only by a narrow margin in a vote. Ironically, the reason behind its acceptance was the fear that if the Representative Method had been discarded, it had also set suspicions on the Monograph Method (Kiaer 1897a).

Partly the critic derived its origin from the distrust of partial investigations because of the same reasons de Keverberg had distrusted Quetelet's idea of partial investigation. Decennial censuses and monograph surveys were sufficient for their interests. Some statisticians had doubts about the validity of the method and said that at some times, surveys might provide interesting information but incomplete surveys should not be granted equal status with the statistical ideal and with "*la statistique serieuse*".

The criticism did not die off, but Kiaer continued to develop the method and he gave presentations about it at the next ISI meetings. The ISI officially accepted the Representative method as a valid method at the Rome meeting in 1924, according to the recommendation of a subcommittee. Two years later, Bowley (1926) published a memorandum where for the first time he presented a sampling theory for survey research.

Kiaer's Representative Method did not include any mathematical method to assess the accuracy of estimates. Estimation and inference were intuitive because the sample was a miniature of the population, but Kiaer was aware of sampling variation. Later, he even suggested that the stability of samples could be assessed by a method based on the idea of sample re-use (Kiaer 1901, p. 68). He did not try it, though.

In addition, Kiaer pointed out that the results of a partial investigation could be controlled to a certain degree, even if general statistics were not available. For example, he suggested that the observed regularity of the phenomena was one kind of a control (Kiaer 1897a). In addition, control could be done by comparing the results of one partial investigation with results obtained by different representative designs. Kiaer concluded that if one obtains approximately the same results by various methods, greater reliability could be placed on the results.

Kiaer's Representative Method aimed at providing an instrument to collect timelier and more in-depth data about social conditions. Decennial censuses provided data that were too little, too general, and too obsolete. In the 19th century, the collection of data in a census required many efforts. In addition, data processing and editing constituted a major effort. Kiaer used Hollerith machines in his first survey for the first time in Norway. The Representative Method combined with data processing by Hollerith machines could provide results essentially faster than a census could. This opened completely new possibilities for social research.

13.1.3 Social surveys in England

In many respects, Arthur Bowley of the London School of Economics should be regarded as one of the key persons in the history of survey research. In 1901, he was accepted as a member of the ISI, and he took part in the discussions on the Representative Method. Obviously, he realised the potential of the method and started to apply it in practice and to elaborate its statistical foundations. Bowley did not refer to Kiaer in any of his early papers, but it seems obvious that Kiaer's work had an influence on him. In the 1903 ISI meeting, Bowley played a decisive role in persuading the ISI to endorse Kiaer's ideas in a resolution. And in 1924, Bowley was nominated to a commission to study the applications of the Representative Method. He wrote an appendix to the memorandum that was the first English treatment of statistical estimation theory in sample surveys (Bowley 1926).

Social surveys, especially those concerning poverty, had a long tradition in Britain. As early as 1837, the Manchester Statistical Society had carried out a survey composed of interviews of 4,102 "families of working men in Manchester". One of the first published reports of a social survey is Heywood's *Report of an Enquiry, conducted House to House, into the state of 176 Families in Miles Platting, within the borough of Manchester, in 1837* (Heywood 1838).

The British Statistical Movement was an unofficial formation of ordinary people in the Victorian Britain. It was very active and it established statistical societies in many cities. Its members carried out many surveys mainly with the aim of revealing the living conditions in working-class households. Probably the most notable accomplishments of the movement are Booth's (1889-1903) and Rowntree's (1901) books. Bellhouse (1988) argues that Bowley should be considered a descendent of this movement.

Bowley observed that the Representative Method was relatively easy to apply in social research to reveal poverty. This can already be seen in his presidential address to the British Economic Society in 1906. The address proves that Bowley was aware of the central questions of survey research although there was no theory for sampling and hardly any experiences were available.

In this address, Bowley also sought to give an empirical proof of the validity of the Central Limit Theorem in a context which today would be called simple random sampling. In addition, Bowley showed that the accuracy of estimates does not depend on the size of the population, but only on "its nature" and on the size of the sample, and that accuracy can be increased, and the probable error decreased, by increasing the size of the sample.

At the beginning of the 20th century, the idea of sampling was vague, but Bowley aspired to make it more distinct and manageable. Therefore, he introduced the concept of a frame and emphasized its importance and the problems that its absence might introduce in social research and in sample surveys. At the end of the address to the Royal Society, he expressed a plea to establish a household registry in the UK, to be used as a sampling frame in social research.

Two decades later, in the memorandum to the ISI, Bowley wrote about practical matters in survey data collection and about the errors that may arise if data collection is not done properly (Bowley 1926). The text could be part of any modern text on survey research. By that time, he already had experience in the practicalities of survey work because he had carried out several of them.

A significant milestone in the history of survey research is the survey that Bowley carried out in Reading in 1912. Its sampling method was close to the one that Kiaer had applied in sampling houses in cities, except that Bowley did not use stratification. In modern terms, Bowley's method was systematic sampling (every tenth house in an alphabetical list of streets). Bowley regarded the obtained sample as random because "it did not involve any purposive elements". Simple random sampling had not been possible because of practical reasons of enumeration.

A consequence of the survey in Reading was that it set out the designing of several similar surveys in the UK. In 1913, Bowley's associates carried out surveys in three other cities and in the next year in two new cities. The survey in Reading was followed by a systematic sampling of census schedules in 1915

(Bowley and Burnett-Hurst 1915). Later, in 1927, Ford conducted a similar survey in Southampton (Ford 1934), and in 1929, the London School of Economics, under Bowley's direction, carried out a survey in London (Bowley 1929). In that survey, the "house sample" involved a two-way stratification. A few years later, Caradog-Jones carried out two analogous surveys, one in Liverpool in 1930 and the other in Merseyside in 1931 (Caradog-Jones 1931 and 1934).

Bowley's activity also had an indirect impact on the development of sample surveys in the United States (see Jessen 1942, and Stephan 1948). Margaret Hogg was an apprentice of Bowley and worked under his direction in some of the British surveys. In 1924, Bowley and Hogg together carried out follow-up surveys for those five surveys, which were conducted in 1913-1914 (see Bowley and Hogg 1925). At the end of the 1920s, Hogg moved to the United States to work for the Russell Sage Foundation. While in the U.S., she made an appeal for rigorous methods of sampling and cast some doubt on the value of surveys that had been made in the U.S., in which the sample was selected by judgment rather than random procedures (Hogg 1930). In 1931, Hogg conducted a survey of unemployment, partly to test the practical difficulties of applying a random sampling method, and partly to develop better questionnaires (Hogg 1932). Hogg's contributions had a noticeable impact on the development of survey sampling methods in the U.S. (Stephan 1948).

13.1.4 Kiaer vs. Bowley

It is a common notion that Kiaer was the key figure behind the emergence of the Representative method, but Bowley's contributions have largely been overlooked. Nevertheless, both of them have strongly influenced the birth and proliferation of survey techniques. They elaborated on similar methods, but there were significant differences between their approaches. Kiaer did not have an academic background and he acted mostly within the circles of national statistical institutes, whereas Bowley had a long career as a teacher at universities and later as a professor of statistics, and he was an active social researcher.

The surveys in which Bowley was involved in the UK probably had a substantial influence on the increase in social surveys outside the national statistical institutes. Kiaer's efforts may have left the Representative Method in the arsenal of statistical offices only as a substitute for censuses. In addition, Bowley's efforts started the development that led survey sampling to become a branch in statistical science.

Since 1905, Kiaer did not contribute to the Representative Method anymore. Obviously, the reason was the heavy criticism the method received in Norway (see Lie 2002). The Representative Method also disappeared from the agenda of the ISI meetings for twenty years. This emphasises Bowley's importance in the development of survey methods. He continued to elaborate and promote the methodology and showed its value to social research by carrying out, or assisting in, several surveys. Bowley's influence is also shown in the minute reporting of these surveys. Some of these reports became classical examples of social surveys.

13.1.5 Development in the United States

The sampling techniques that are currently applied in surveys are based on the theory that was created in the U.S. in a relatively short period in late 1940s, even though the foundations of the techniques were laid in Europe in the 1920s and 1930s. Stephan (1948) thoroughly analysed the rationale that led to this development. At the beginning of the 1950s, the classical theory was manifested in three textbooks on sampling theory that were published within a few years. The foundations of randomization inference are postulated in these books, and most of the surveys undertaken today apply the theoretical setup offered in them. The more recent textbooks, such as Särndal et al. (1989) or Lehtonen and Pahkinen (2004), are based on the same basic philosophy, even though the sampling techniques are elaborated and extended from their origins. Since the beginning, the theories and methods of sampling have also been integrated in the corpus of statistical science.

The single most important impetus for the development of modern sampling techniques came from the designing of the Current Population Survey (CPS) at the U.S. Bureau of the Census. According to Hansen and Madow (1978), there was intense pressure in the 1930s to collect reliable information about the population in the country, especially on unemployment and social conditions, to design policies, and to develop social programs. Without a solid theory, the sampling methods became vague, and estimation was intuitive (see Stephan 1948). Many different partial enumerations were carried out in the 1930s, but their reliability was suspected in general, and with a reason: the estimates on the number of unemployed people varied between 7 and 20 million.

At the end of 1930s, the Census Bureau still upheld the idea that it could not undertake sampling surveys because that would discredit its results on other areas. Only a complete enumeration was accepted (see Hansen 1987 and Olkin 1987). In 1937, Neyman delivered a series of lectures in the U.S., and after one of these lectures, Neyman was asked a question. That led to the famous paper on double sampling (Neyman 1938), in which Neyman showed that a statistically rigorous theory of sampling would be attainable, at the same time satisfying the requirements for manageable fieldwork and with acceptable costs. According to Hansen, this paper and Neyman's lectures stimulated the development of sampling methods for the CPS (Hansen 1987).

Hansen and Hurwitz started to develop a completely new probability sampling design. A significant condition for the development was presented by the costs of data collection. Hansen and Hurwitz concluded "... formulas have not practical utility unless there are also some considerations on differential costs." They aimed at developing a method that would make the most effective use of available resources by organizing adequate field operations.

At the end, they managed to develop the basic theory of stratified two-stage sampling with one primary sampling unit (PSU) within each stratum drawn with probability proportional to size measure (PPS sampling) and then sub-sampled at a rate that ensures self-weighting within strata. The result of their work is best documented in the paper entitled "*On the Theory of Sampling from Finite Populations*" by Hansen and Hurwitz (ibid.). The development of the design-based theory for sampling techniques seems to have started from this article. This work also led to the publication of one of the famous books on sampling theory, "*Survey Sampling Methods and Theory, I and II*" (Hansen, Hurwitz and Madow 1953). Interestingly enough, unequal inclusion probabilities were already implicitly present in Neyman's optimal allocation designs (Neyman 1934), but no one paid attention to that. In Hansen's and Hurwitz' method, the possibility to have varying inclusion probabilities was explicitly articulated for the first time.

The complex sampling design for the CPS provided approximately equal interviewer workloads, which is essential in organising the fieldwork. It was important that a manageable field organisation could be established for data collection and a reasonable amount of fieldwork was required to provide estimates. Otherwise, the CPS probably had not been started with this design – and the history of sampling techniques would be different.

The Current Population Survey was carried out for the first time in 1943 more or less with the design presented in Chapter 1. The same design was soon taken into use in national statistical institutes in many countries, for example, in Canada in 1945, in Japan in 1950, in France in 1955, and so on. In three decades, most national statistical institutes adopted a similar design for their Labour Force Surveys. Social researchers quickly comprehended the value of the method. The emergence of electronic computers in the 1950s and statistical software packages in the 1960s essentially increased the attraction to apply sampling techniques.

A noticeable feature is that while the current sampling techniques were developed in the U.S., the fundamental questions of statistical inference were not discussed at all. Obviously, the randomization inference principle that Fisher and Neyman presented in the 1920s and the 1930s were already accepted unanimously. Cochran, Hansen and Hurwitz, and also Madow only occasionally mention the inference principle of drawing repeated samples from the same population and they seldom mention confidence intervals. It seems that the problems of statistical inference were regarded as solved. Neither was there any discussion on the principles of induction, which had caused the bitter dispute between Fisher and Neyman.

13.2 History of randomization

Randomization in sampling techniques has two different roles: (1) random selection of population units is believed to yield a representative sample; and (2) randomization makes statistical inference viable. This difference has been overlooked in the writings on the history of survey sampling. The two roles of random selection have different histories. Although the origin of random selection is difficult to trace, its merits were already acknowledged at the end of the 19th century. Randomization inference, on the other hand, was created in the first quarter of the 20th century.

A frequent argument in the writings has been that at the beginning of the 20th century, statisticians were free to select between purposive and random selection without any fear of criticism. That is an ambiguous argument which is difficult to substantiate. If random selection means only probability sampling and all other selection methods are classified as purposive, then the argument could be understandable. Probability sampling was taken into use only in the 1940s, but haphazard sampling has been used for decades before that. The meaning of purposive selection is vague in the writings. It is not indicated whether it is some general sampling method other than probability sampling or if it means the method that Bowley presented in 1926.

In the next paragraph, random sampling means a method in which a sample is selected from a labelled population by some random mechanism. Haphazard sampling means a method in which the sampling units in a population are not labelled, but the selection is made by objective, semi-random methods, for example, systematically (e.g., every tenth house on a street). Purposive selection means a method in which sampling units are selected so that their (mean) values coincide with the population (mean) values.

13.2.1 Random selection of sample

It is difficult to find out when random selection in the context of partial investigations was applied for the first time. It is possible that its implications were first noted at the end of the 19th century in agricultural research. In Russian texts at that time, there are several references to the use of mechanical sampling, which in modern terms would mean systematic (area) sampling (Kaufmann 1913, Tchuprov 1910a and 1910b). However, random selection in human populations is not as straightforward as in agricultural research.

In Kiaer's Representative Method, the selection of households in cities was close to modern systematic sampling. It was relatively easy to put into practice because of the structure of cities, but in rural areas, enumerators had to select houses according to Kiaer's instructions. In another sampling investigation from 1891, census records Kiaer had applied random selection. He selected the units by the initial letters of the first name of a person. Basically, this is close to a randomization method, called the closest birthday method, which is still in use when selecting one respondent in a household.

At the beginning of the 20th century, statisticians were broadly aware of random sampling and its merits, but they also had to consider the practicalities of enumeration. Already in 1906, Bowley had introduced the concept of 'sampling frame' to mean a collection of population units from which a random sample could be drawn in such a manner that equal inclusion probabilities appeared plausible. Inclusion probabilities were assumed to be equal and therefore not needed in estimation, and consequently, there was no need to attach the probabilities to population units. In 1934, Neyman wrote, "Random sampling means a method of including in the sample single elements of the population with equal chances for each element." (Neyman 1934)

In the 1930s, random sampling still only meant simple random sampling because no other sampling methods were known. The problem was that simple random sampling could not be applied because of practical reasons. Drawing of such samples was not possible or it was too difficult, and the enumeration of such a sample had been very laborious and expensive to carry out because sample sizes were usually very large. At the end of the 19th century, Kiaer's determination of sample size was intuitive, and a sample was selected that was large enough to ensure accuracy. Only the modern sampling theory has shown that dramatically smaller samples can be sufficiently accurate.

The use of systematic or cluster sampling schemes was frequent. The data collection including fieldwork was relatively easy to organise, and systematic selection appeared unbiased because the selection was haphazard and objective, and hence credible. In addition, it was generally accepted that such a sample would be unbiased if "... the chance of inclusion in the sample is independent of the value X ..." (see Yule 1911).

In the first quarter of the 20th century, Bowley carried out several surveys in the cities of the UK, and he was also an advisor in several surveys. In all those investigations, the sample was selected by some kind of random or haphazard method. Bowley presented the method of purposive selection in his report to the ISI (Bowley 1926), but there is no evidence that he would have ever used it in practice. In fact, there are only a few instances where purposive selection was used (see Jensen 1926). In 1934, Neyman concluded that ". . . As the theory of purposive selection seems to have been extensively presented only in the two papers mentioned [i.e., Bowley 1926, Gini and Galvani 1929] while that of random sampling has been discussed probably by more than a hundred authors . . ." (Neyman 1934).

13.2.2 Randomization in statistical inference

In the current probability sampling theory given in the textbooks, the randomization inference is defined in the following way: It is possible to define a set of distinct samples, $\mathbf{S} = \{s_1, s_2, s_3, \dots s_M\}$, which can be obtained with the sampling procedure if applied to a specific population. A known probability of selection $p(s)$ is associated with each possible sample s_i. The assumed procedure gives every element in the population a probability of selection or the inclusion probability of the element. One sample is selected randomly so that each possible sample s_i receives exactly the probability $p(s)$. This method is called probability sampling. The function $p(.)$ defines a probability distribution on $\mathbf{S}$. It is called a sampling design. A finite population is the target of inferences and the stochastic structure is induced by the sampling design. This inference model induced by probability sampling is called design-based, or randomisation inference. It is assumed that for any sampling procedure that satisfies these properties, it would be possible to calculate a frequency distribution of the estimates generated by repeated drawing samples from the same population. This distribution has been shown to tend to normality as the sample size increases.

The importance of randomization for statistical inference has been acknowledged for a long time. Already in 1906, Bowley emphasized the importance of random selection for validating estimation. Moreover, Kovalevsky wrote in 1924 that valid estimation in partial investigations requires a random selection of units. However, the exact formulation of randomization inference emerged only in the textbooks written in early 1950s. For example, Hansen and Hurwitz

(1943) do not explicitly touch on inferential aspects at all. In statistical texts up to the 1940s, the role of randomization in the context of statistical inference was very vague, although its significance was acknowledged.

The idea that the sampling distribution was produced by drawing repeatedly random samples "in identical conditions" was already present in Edgeworth's and Yule's texts, but its implications for estimation were not paid attention to. It was Fisher who included the notion of randomisation as a central method in the design of experiments. In this context, he also introduced the principle of replication as the fundamental principle of experimental design. Replication is the main source of the estimate of error, while randomization ensures that the estimate will be unbiased. From that origin, the new meaning of randomization also spread to other areas of statistics.

The importance of randomization in statistical inference was in a concealed form in the design of Hansen and Hurwitz (1943). The importance of randomization became emphasized only after Horvitz and Thompson introduced unequal inclusion probabilities in the explicit form. If all inclusion probabilities are equal, they may be disregarded from the formulas and hence randomization will not be explicitly involved in estimation. Consequently, the calculation of estimates is possible even if the selection of sample is not explicitly random. It suffices that it is haphazard.

13.3 Milestones in the history of statistical inference up to the 1950s

It is unquestionable that the current methodology of statistical inference for finite populations stems from the contributions of Jerzy Neyman in the 1930s. However, the history of the methodology is considerably more diverse. Neyman's theory draws from both Fisher's theory of statistical inference and Bowley's sampling theory. Although Fisher dealt only with statistical inference when the population is infinite or hypothetical, the impact of his innovations was crucial for the inference for finite populations. In addition, in Neyman's thinking, it is possible to identify strains from the statistical ideas of the Russian school. Neyman studied statistics and started his career as a statistician in Russia and later in Poland without western influence until he was 27 years old (Fienberg et. al. 1966).

In the textbooks on statistics, it is customary to give the impression that the general theory of statistical inference started from Fisher's contributions in the 1920s. Yet there already existed a method for statistical inference before Fisher, but it was based on a different probabilistic setup than the current (Fisher's) theory, and it is commonly overlooked.

13.3.1 Inverse probability

In 1802, P.S. Laplace conducted the first scientifically ambitious partial investigation in which he applied an estimation method that was based on his Principle of Inverse Probability. Laplace's survey aimed to estimate the size of the population in France with a design that involved methods typical for a modern sample survey. In fact, he had already published the theoretical part for the survey twenty years before the survey was undertaken. In 1784, Laplace published a mémoire in which he derived the method to estimate the size of a population by a ratio estimator. In addition, he derived formulas to calculate the accuracy of the estimate, and therefore he was able to calculate a probabilistic interval estimate. This was because he showed that the error in the estimate due to sampling followed an error distribution that was later called the Normal Distribution.

Laplace calculated that the "erreur à craindre", which was conceptually close to the standard error, given the data, was 107,550 persons, and he concluded that it makes "the odds about 300,000 to 1 against an error of more than half a million". Eventually, the estimation of population took place on 22 September 1802. Laplace reported that the population of France on the specific day was 28,352,845 inhabitants and that the probability that the error in the estimate was more than 500,000 was 1:300,000, meaning that the probability that the true number of inhabitants would be less than 27,852,845 or more than 28,352,845 is 0.0000033.

Laplace's interval estimate was very close to the modern interval estimate, or confidence interval of the modern sampling theory. In addition, a noteworthy feature was that before the survey, Laplace had calculated the sample size needed to attain the required accuracy.

Laplace's estimation of the population of France was not the first time in history when the ratio estimator was used, however. A century earlier, an English merchant, John Graunt, estimated the size of the population of London by a similar method. The difference was that Graunt's method was completely intuitive and he did not calculate the accuracy of the ratio estimate. Interestingly enough, Laplace's estimation method is used even today in estimating the size of wildlife populations using the so-called mark-recapture techniques (see Pollock 1981).

Characteristic of Laplace's method was that the inference model[145] was based on Bernoulli trials, which induced the Binominal distributions. The factorials in binominal coefficients were expanded with Stirling's formula and Euler's series, and then terms that in large samples would become negligible were discarded. This is the way Laplace derived the Law of Error.

In modern texts on statistical science, Laplace's contributions to statistical theory have been largely disregarded, although Gauss' and Bayes' contributions are well brought up. A possible explanation is that Laplace's theory is erroneously regarded as "Bayesian" inference and it contradicts the predominant Fisherian inference. If this is the case, it is partly based on confusion. Laplace's method is his own elaboration, and Bayes' probably had only a minor influence on its later development (see Hald 1998 and Stigler 1986). Laplacian inference does not contradict Fisherian inference, but they are based on a different approach and a different inference model and hence possibly on a different paradigm. Partly, the two different approaches

[145] The inference model is an intermediate model – a thought experiment – which links an abstract probability model to real-world phenomena. A thought experiment is typically composed of a setup that could be tested experimentally if necessary.

stem from two different worldviews. Laplace's worldview has been depicted as Newtonian. In the same sense, Fisher's worldview can be depicted as Einsteinian, even though the analogy is not strict. Consequently, rather than Fisher's, Laplace's method should be regarded as the first instance of statistical inference, even though it was based on a different approach to solve the problem.

Thomas Bayes and Thomas Simpson were the first to deal with topics related to the inverse method. Bayes' method, published in his Essay, has been regarded as the first expression aimed at introducing a method for inverse probability. In the Essay, Bayes derived his principle by the geometrical method of Newton, and the famous Bayes' formula was extracted much later. According to Stigler (1986), Bayes' Essay remained nearly unknown for many years and its influence in the development of probability in the 19th century was minimal.

In the 19th century, Bayes is only briefly mentioned in a few statistical texts, whereas references to Laplace's works are frequent. As a typical example, Todhunter (1865), in the first comprehensive account on the history of probability, only briefly and superficially mentions Bayes (5 pages), whereas the part that deals with Laplace's contributions takes up nearly one-quarter of the book (about 150 pages). Todhunter (ibid.) concludes "...on the whole the Theory of probability is more indebted to him than to any other mathematician." Laplace's influence on the 19th century scientific world was momentous, while Bayes' Essay was almost unknown. Bayes appeared in the statistical literature only in the latter half of the 20th century (see Fienberg 2006). Laplace also articulated, more clearly than Bayes, his argument for the choice of a uniform prior distribution, arguing that the posterior distribution of the parameter should be proportional to what in modern terms would be called the likelihood of the data.

Apart from the method of inverse inference, Laplace developed many of the central ideas in probability theory and statistical science. One of his most important discoveries was the Central Limit Theorem. In addition, Laplace presented the idea of statistical testing and introduced the idea of maximum likelihood in a rudimentary form[146], and according to Stigler (1986), he was also close to discovering the idea of sufficiency, which Fisher derived in the 1920s.

It is obvious that Laplace's works had a strong influence on scientific activity in the 19th century and their impact was wide-ranging (see Todhunter 1865, Stigler 1986, or Hald 1998, 2007). Therefore, Laplace's contributions and ideas should be considered as the first formal presentation of statistical inference.

13.3.2 Bowley's sampling theory

Bowley already calculated simple confidence intervals for the survey in Reading in 1912. He attached calculations to the results to assess the accuracy of estimates: If in a "sample of 622 working-class households we find respectively 5, 10, 20, 40, 50 per cent of cases, we may expect that the percentage in the whole are within 5±1, 10±1, 20±1½, 40±2, 50±2 and may be nearly certain that they are within 5±3,10±4, 20±5, 40±6, 50±6." Bowley concluded that the probabil-

[146] The first appearance of the idea of maximum likelihood has been dated back to the writings of Daniel Bernoulli and Heinrich Lambert (see Hald 1998).

ity "is about 2 to 1 in favour of the true being within the limits for the first set, and 1 to 250 for the second set." (Bowley 1913) Calculations were based on the Normal Distribution and standard deviations.

In the 1920s, Bowley began to elaborate a mathematical approach to sampling theory based mainly on Edgeworth's contributions who, in turn, had elaborated Laplace's methods. Since Laplace, Bowley seems to have been the first who applied probability and the Laws of Errors on a (randomly selected) sample from a finite population. Bowley's method is summarized in a memorandum to the ISI, published in 1926 (see also Hald 1998). Three years before the memorandum, Bowley had written an article entitled *The precision of measurements estimated from a sample* (Bowley 1923). The article treated the "inverse problems in statistics", applying the method that Edgeworth had already presented in 1908. Two years before Bowley's memorandum to the ISI, Kovalevsky had written a mathematical treatment on the same topic (Kovalevsky 1924). Obviously, it was not known in other parts of Europe because it was in Russian and most of its copies disappeared in the throes of the revolution in Russia.

Bowley separately analysed random sampling and purposive selection; and in random sampling, he analysed simple random sampling and stratified sampling. Under random sampling, Bowley derived formulas for the estimation of the accuracy of "prevalence of one attribute", for "distribution of attributes", and for estimates of the average. The approach to the problem under random sampling follows the lines of thought that were already present in Laplace's works, but Bowley applied them to a wider range of problems. In deriving estimates, Bowley applied the same mathematical methods that Laplace had used, i.e. applying Stirling's formula to solve factorials, Taylor's expansion, and ignoring terms that became negligible in large samples. He also applied the Method of Moments that Edgeworth and Karl Pearson had applied frequently.

The inverse probability principle of Laplace involves the *a priori* probability distribution of a population parameter *P*. Bowley also included *a priori* probability distribution because he believed that some information about the population, or "the universe", would be needed to be able to estimate population parameters. The distribution of the parameter was unknown, and therefore an assumed *a priori* distribution was introduced. Unlike Laplace, Bowley did not agree with the assumption of a uniform distribution of priors. Instead, he assumed that the *a priori* probability distribution, *F(P)*, is continuous and derivable in the neighbourhood of $P = p$, where p is the proportion observed in the sample. However, Bowley showed that the *a priori* distribution vanishes and therefore it does not exist in the final formulas anymore.

Bowley observed that the probability that the population proportion, *P*, is within the limits $p \pm z$ where $z = x/n$ (p being the observed proportion and n the sample size), is approximately

$$P(p-z \le P \le p+z) = \int_{-z}^{z} \frac{1}{\sqrt{\left\{2\pi pq\left(\frac{1}{n}-\frac{1}{N}\right)\right\}}} e^{-\frac{z^2}{2pq\left(\frac{1}{n}-\frac{1}{N}\right)}} dz$$

The corresponding formula for an estimate of average, $\overline{x}$, is approximately

$$P(\overline{x}-x \leq \overline{u} \leq \overline{x}+x)=\int_{-x}^{x} \frac{1}{s\sqrt{(2\pi)}} e^{-\frac{x^2}{2s^2}} dx,$$

if $\overline{u}$ is the population mean and s the sample standard deviation, and terms of order $1/\sqrt{n}$ are disregarded.

Bowley also derived both of the formulas for stratified sampling. By stratification, he meant a method where an equal proportion of units are selected at random from each stratum, i.e. proportional stratification. Bowley observed that in every case, the accuracy of estimation increases by stratification, and in some cases, the improvement was considerable.

A chapter in Bowley's memorandum dealt with purposive selection. In Bowley's purposive selection method, the population under investigation was assumed to consist of *N* districts, or clusters in modern language. The aim of a survey is to estimate *P*, the proportion of the units in the aggregate of clusters having the attribute of interest (or the average of a variable).

Bowley assumed that there are one or more associated variables, "controls", whose values are known in every district and the partial regression equation between the study variable, *x*, with control variables, y_i, $i=1,\ldots, t$, is linear. This was used to calculate an adjustment factor *K*. All the terms in *K* can either be calculated exactly from the population data or from the sample. The standard deviation of the error term is calculated from the standard deviation of the study variable and partial correlations between the study variable and controls.

In Bowley's method, the districts were selected (purposively) in such a way that the average for each control variable 'is the same in the aggregate of them as it is in the universe'. This requirement was essential. It involves the assumption that if the averages match, then the selected districts compose a representative sample from the "universe", i.e., the population. The method was laborious, and obviously only Gini and Galvani (1929) carried out a survey using it.

Purposive sampling in this sense fell into oblivion until Royall and Herson (1973) defined a balanced sample as a sample. They defined a balanced sample of order *T* as one for which the sample mean $\bar{x}_s^{(t)}$ of variable $x_j{}'$ was equal to its population mean $\bar{x}^{(t)}$, for $t = 1, 2, \ldots, T$. The basic idea in Bowley's purposive selection and the modern balanced sampling is the same: a sample is made representative of the population by purposive selection of sampling units by matching control variables. A balanced sample differs from Bowley's purposive selection in what are regarded as sampling units. In balanced sampling, they are single observations or measurements, but in Bowley's purposive selection, they are aggregate values of 'districts' or clusters.

Lately, balance sampling has again emerged in the statistical literature, probably because modern computers have facilitated the sampling process (see Deville and Tillé 2004). Chauvet (2009) has also introduced stratified balanced sampling designs.

13.3.3 Neyman's inference model for finite populations

Neyman's paper in 1934 created a new approach to deal with the problems of inference within a finite population framework. He was inspired by Bowley's report to the ISI, but Neyman drew his key ideas from the concepts and ideas that Fisher had put forward in the 1920s. Statistical inference was a familiar topic to Neyman. In 1933, he had published, together with Egon Pearson, a paper in which they established the so-called Neyman-Pearson theory for testing statistical hypotheses (Neyman and Pearson 1933).

Fisher established a new theory of statistical estimation in two papers that were published in 1922 and in 1925 (Fisher 1922 and 1925a) while working as a statistician at the Rothamsted Experimental Station. The theory of estimation in the modern sense did not exist before these contributions. Fisher's paper in 1922 included a great number of completely new ideas and it revolutionized statistical theory. Stigler (2005) regarded it as an astonishing piece of work because "it announces and sketches out a new science of statistics, with new definitions, a new conceptual framework and enough hard mathematical analysis to confirm the potential and richness of this new structure." After all these years, it is easy to see that this article was a watershed in the development of statistical science, and Fisher can be called the founder of modern statistics (Rao, R. 1992).

In 1925, Fisher published his first book, *Statistical Methods for Research Workers*, in which he presented the analysis of variance and statistical significance testing. In this context, he introduced his idea of statistical inference, which he called inductive reasoning. In 1930, Fisher presented his Fiducial Argument to replace the inverse probability principle.

Although Fisher did not contribute to the inference for finite populations, his statistical theory made the basis for it. Three single issues that also have a significant bearing on the finite population inference were the new concept of population, fiducial inference, and the new inference model. In the theory building before Fisher, a population was not assumed to be stable but was viewed as continually changing. Therefore, its parameters were regarded as being stochastic, and this was indicated by *a priori* probability. The observable population was a realisation of a superpopulation. In Fisher's theory, population was assumed to be invariable, and consequently, population parameters were constants. Hence, *a priori* probabilities became irrelevant. In addition, Fisher had strong philosophical reasons to discard *a priori* probabilities.

The central idea in the fiducial argument was revising the "Student's" (1908) formula. From a normal population with a mean value μ, a sample of size n has been drawn. From the sample, two statistics, the mean $\bar{x}$ and the variance s^2, are calculated. The quantity t, defined by equation

$$t = \frac{(\bar{x} - \mu)\sqrt{n}}{s}$$

is distributed in different samples in a distribution dependent only from the sample size, n. With the help of the Student's t-distribution, it is possible to calculate, for each value of n, what value of t will be exceeded with any assigned frequency. Statistics t is a continuous function of the unknown parameter, the mean, and with observable values $\bar{x}$, s, and n. Consequently, Fisher argued, the inequality $t > t_l$ is equivalent to the inequality $\mu < \bar{x} - st_1 / \sqrt{n}$. This inequality became the basis of Neyman's confidence intervals (Neyman 1934). In 1937, he introduced a solid mathematical theory for estimation using confidence intervals (Neyman 1937).

Fisher introduced the idea of repeatedly drawing samples from the same population to replace Bernoulli trials as the inference model. This procedure yields a continuous sampling distribution for the estimate, which became the basis of inference. Neyman adopted Fisher's inference model, and repeatedly drawing samples from the same finite population became the basis of Neyman's inference model.

Instead of using Fisher's Maximum Likelihood estimators, Neyman developed Best Linear Unbiased Estimators (BLUE) by using a method that Markov[147] had presented in 1900 ("best" meaning minimum variance). Markov's method is based on the ordinary least squares estimation. Unlike Maximum Likelihood estimators, BLUE estimators could be applied independently of the distributions of variables and practically in any finite populations. This feature made Neyman's method very appealing for survey research.

Using Markov's method, Neyman also derived his formula for optimal allocation in stratification (Neyman 1934). Later, this appeared to be an important result because it yielded estimators for double sampling (Neyman 1938). This in turn directly addressed the needs existing in the U. S. by showing how a complex survey design should be approached. The three papers that Neyman published in the 1930s set up the foundations for statistical inference for finite populations.

The first papers, which Neyman published in the early 1920s, indicate that he was trained within the Laplace-Bayes paradigm, and he was applying its principles. To a certain degree, his ideas stem from the Russian school of statistics, and the inference model for finite populations partly reflects his early ideas. Tchuprov and Kovalevsky had already discovered optimal stratification in the early 1920s. Kovalevsky's derivation is similar to Neyman's basing on the method of Markov. For a long time, there has been controversy over whether Neyman was aware of Tchuprov's or Kovalevsky's results in 1934, but a final conclusion has not been reached.

Later, there appeared a significant disagreement with Fisher and Neyman. Neyman did not concur with Fisher in the nature of statistical inference. Neyman said that statistical inference provides rules of behaviour, not rules of reasoning, as Fisher thought. Neyman explained that a scientist applying inductive behaviour might be wrong in 5% of his decisions, but he is not able to say which decisions are right and which ones are wrong.

[147] In modern statistical literature, Markov's method is known as the the Gauss-Markov theorem. It was first discovered by Gauss and later independently by Markov.

13.4 Paradigms

In 1962, Thomas Kuhn introduced the concept of paradigms to explain how scientific ideas develop (Kuhn 1962). Kuhn argued that scientific research and thoughts are defined by paradigms, or conceptual worldviews. The thesis of Kuhn was that scientific disciplines, once they have emerged from the pre-paradigmatic stage, undergo periods of so-called 'normal science', which allow them to obtain rapidly a high degree of precision and progress. During the period of normal science, research develops as a steady and cumulative acquisition of knowledge where new findings and results of experiments are added to previous knowledge to form more accurate or extensive theories.

For Kuhn, normal science meant research based on past achievements that a scientific community acknowledges for a time as supplying the foundation for its further practice. Such achievements are described both in elementary and advanced textbooks. These textbooks explain the body of accepted theory, illustrate many of its applications, and compare these applications with exemplary observations and experiments.

Normal science is dependent on the adoption of a universally accepted paradigm that defines research problems for the scientist, tells him or her what to expect, and provides the methods that he or she will use in solving them. Kuhn used the term 'paradigm' in two different senses. On the one hand, it stands for an entire collection of beliefs, values, techniques, and so on, shared by the members of a scientific community. On the other, it denotes one sort of element in that constellation, the concrete solutions which, employed as models or examples, can replace explicit rules as a basis for the solution of the remaining problems of normal science. The first sense Kuhn called sociological: a paradigm is what the members of a scientific community share, and, conversely, a scientific community consists of men and women who share the paradigm.

From the very beginning, new scientists are indoctrinated into the prevailing paradigm. Consequently, only young scientists who are not yet so deeply indoctrinated into accepted theories (such as Newton, Lavoisier, or Einstein) can manage to sweep away an old paradigm.

Kuhn called the replacement of the old paradigm with the new one a 'scientific revolution', or 'paradigm shift'. At first, the scientific community resists the replacement, but with time, the success of the new paradigm gains enough support to win out. According to Kuhn, paradigm shifts have often been intellectually violent.

The scientists within the new discipline see the world in a different way than it "was" under the old paradigm. Kuhn argued that a scientific revolution is a non-cumulative developmental episode in which an older paradigm is replaced in whole or in part by an incompatible new one. The new paradigm cannot build on the preceding one – it can only supplant it.

The history of statistical inference has not been analysed much in respect to paradigms and paradigm shifts. One reason may be the fact that paradigms in a methodological science cannot be similar as in natural sciences which Kuhn (ibid.) analysed. In statistical inference, rather than a collection of beliefs and values, paradigms are composed of points of view from which to approach problems, and models to look for solutions.

Probably another reason has been the widespread conception that the theory for statistical inference was created by R. A. Fisher in the 1920s. The texts dealing with the history of probability cover either the development up to the beginning of the 20th century or the development since Fisher. Recently however, Anders Hald published a book (Hald 2007) on the history on parametric statistical inference from the 18th century to the beginning of modern times. He recognized three revolutions in the history of statistical inference. Although a revolution is not a paradigm, paradigms, according to Kuhn (ibid.), are often started from scientific revolutions.

Paradigms have been a controversial topic ever since Thomas Kuhn brought them up in 1962. An important question is whether paradigms would be of any importance in the context of this thesis. The author's opinion is that they are. If paradigms can be identified, Kuhn's theory explains one part of the historical development of statistical science. Especially, the history of statistical inference can be seen from a different perspective: as a flow of ideas starting from the 18th century up to modern times.

13.4.1 Paradigms in the Representative Method

Bellhouse (1988) argues that the initial paradigm in survey sampling is that of the desire to collect a representative sample as presented by Anders Kiaer in the 1890s. Bellhouse also says that there are earlier examples of partial investigations but that they illustrate the randomness in research as is typical for the pre-paradigmatic times. Bellhouse (ibid.) thinks that Kiaer's initiative of the Representative Method led to a new paradigm of statistical data collection.

The reports about the ISI meetings around the turn of the century indicate that the new paradigm was not immediately accepted. Some comments about the Representative Method appear hostile, and some famous statisticians urged that the method should not be discussed anymore in the ISI meetings, but after a voting, it was decided to keep it on the agenda of the next meetings[148]. Interestingly enough, at the beginning of the 1940s, sample surveys were still not accepted at the U.S. Bureau of the Census (see Hansen and Madow 1976). To some degree, the acceptance of the Representative Method as a new paradigm seemed to be an intellectually violent revolution in the sense Kuhn (1962) defined it.

The Representative Method did not supplant full enumerations. Instead, sample surveys and full enumerations have been carried out side-by-side for more than a century. It was also Kiaer's intention that the new method would become a supplementary method for statistical institutes to be used along with full enumerations. Therefore, it was not a paradigm shift but rather the beginning of a new paradigm as also Bellhouse (1988) has argued.

[148] Obviously, the discussion in the ISI meeting was even more heated than the official ISI report reveals. According to an unofficial version of the protocol done by the Swiss Statistical Society, the assembly decided that the question would not be studied further (Malaguerra 2000). The official protocol of the ISI tells otherwise.

13.4.2 Paradigms in Statistical inference

Brewer (1999) divided the history of survey sampling into three parts. He argues that in the first part, from the end of the 19th century up to around 1945, survey designers could select between randomization sampling and purposive sampling "...on an arbitrary basis, apparently without serious fear of criticism". During the next 25 years, Brewer claims, the random selection of samples went virtually unchallenged. Then during the 1970s, the choice re-emerged in the form of balanced sampling. Brewer (ibid.) calls the first period 'pre-paradigmatic' in the sense Kuhn defined it, and the second period was dominated by the randomization paradigm.

Bellhouse (1988) identifies a paradigm that started from Neyman's paper in 1934. In Bellhouse's mind, the reasons are twofold: the first is that by that paper, randomization was pointed to as the recommended solution in sample selection and the problems of purposive selection were shown indisputably; the second reason was that it provides a theory of point and interval estimation under randomization.

Neither Brewer (1999) nor Bellhouse (1988) explicitly analyze statistical inference although its existence in the background is obvious. Neither of them recognizes the history in the 19th century.

13.4.2.1 The Inverse Probability of Laplace

In 1774, Laplace presented his Principle of Inverse Probability in a memoir to the Royal Academy of Paris entitled *Mémoire sur la probabilités des causes par les évènemens* (Laplace 1774). The Principle was aimed at providing the probability of causes of events. Basically, the problem is the same as in statistical inference: to infer from a part the nature of the whole, or to infer from a sample the value of a parameter of the population.

In 1783, Laplace published another memoir in which he presented the idea for estimating the population of France. In the memoir, Laplace analyzed how "the size of sample should be determined to obtain a large probability that the error in the predicted population size is small". At the end, he was able determine the sample size needed to reach the predefined accuracy of estimation. He also defined the "error to fear", which is a measure similar to the modern standard error. Laplace observed that the (sampling) error was distributed according to a distribution which later was named the Normal Distribution. Hald (1998) concludes that the inference theory that Laplace presented is essentially correct (in the light of modern theory) for simple random sampling, although his model did not actually correspond to this mode of sampling.

Hald (1998) deems Laplace's 1774 memoir as one of the revolutionary papers in the history of statistical inference. Stigler (1986) points out that even after more than two centuries, it seems almost like a contemporary work. Stigler (ibid.) also claims that the influence of this piece of work was immense. It was from this memoir that the ideas, which are now called "Bayesian", first spread through the mathematical world. It is actually misleading to call these ideas "Bayesian". For many reasons, these ideas should rather be called "Laplacian".

Thomas Bayes' Essay was published in 1764, a decade before Laplace published his Principle. Laplace's Principle and Bayes' method are similar. However, according to Stigler (1978), Bayes' Essay was ignored until after 1780 and played no important role in the scientific debate until the 20th century. An apparent conclusion is that Laplace was not aware of Bayes' Essay. Both Stigler (1978) and Hald (1998) argue that Bayes' and Laplace's theories are conceptually and mathematically so different that they cannot be related.

Weatherford (1982) claims that the classical theory of probability reached its zenith in the work of Laplace. In addition, Laplace's mathematical treatment of the statistical problems provided new mathematical tools for the later development of the probability theory. His contributions were so influential that they dominated statistical thinking for nearly a century. He had several influential followers, such as Poisson, Quetelet, and Lexis, who fostered Laplacian science and imprinted it on the scientific thinking. Laplace's textbooks were still being reprinted at the end of the 19th century.

Typical features of Laplace's method for statistical inference are described in Chapter 5. It should be noted that the concept of *a priori* probability should not be confused with concepts like "credibility" or "degree of confirmation," or "strength of expectation," etc., as is often done in modern Bayesian theory. In Laplace's and Bayes' theory, *a priori* probability is an objective probability, but its value is not known and its value cannot be found experimentally.

Hald (1998) called Laplace's inference method the Bayes-Laplace model; Stigler (1978) called it the Bayes-Laplace method; and Jeffreys (1983) called it the Bayes–Laplace theory. Markedly, it also fulfils the characteristics of a paradigm in the sense Kuhn described it. In the 19th century, Laplace's theories were dominant in universities, and there did not seem to be any rival theories. In the writings on probability theory, Laplace's patterns of thought are prevalent throughout the 19th century. For example, Quetelet's and Lexis' theories are based on these ideas as well as on the theory building in Russia. Also, all English statisticians before Fisher worked from this paradigm.

Rietz (1924) found out that for a period of more than fifty years following the publication of Laplace's work in 1812, little of importance was contributed to the subject. He described the period starting after Laplace's and Gauss' most productive times as one of clarification and consolidation of the works of Laplace and Gauss. During that period, probability theory was extended to applications from the natural sciences to the social and biological sciences (see Hald 1998).

Anders Hald begins his book about the history of parametric statistical inference (Hald 2007) by saying: "The three revolutions in parametric statistical inference are due to Laplace, Laplace and Gauss (1809-1811), and Fisher." Hald argued that the first revolution in statistical inference, due to Laplace, took place between 1774 and 1786 when Laplace turned his attention from direct probability and derived his Principle of Inverse Probability. The second revolution that Hald (ibid.) identified took place in 1809-1828 when Gauss and Laplace, with the help of the Principle of Inverse Probability, discovered the Central Limit Theorem and the method of least squares. Hald (ibid.) argues that the second revolution was concluded by the contributions of Edgeworth (Edgeworth 1908 and 1909) in which he completed the large-sample theory of statistical inference by inverse probability initiated by Laplace and generalises Laplace's central limit theorem.

The inference method is a creation of Laplace, and obviously Bayes' influence was marginal – or none. In the 19th century, neither Bayes' inference model nor Bayes' thought patterns were referred to in writings on probability. It appears that Bayes' ideas were not the basis in the development of statistical methods. Therefore, Bayes name could even be dropped off and the model could only be called the Laplace paradigm. On the other hand, the term 'Bayesian' is strongly attached to modern statistical language while 'Laplacian' is not. Mentioning both gives a more illustrative expression to the nature of the paradigm. Therefore, calling the model the Laplace–Bayes paradigm seems warranted.

In 1926, Arthur Bowley published a well-thought theory for statistical inference for finite populations (see Chapter 13.3.1). Bowley leaned on Edgeworth's contribution at the beginning of the 20th century, and the theoretical framework was Laplacian. Possibly, it is the last important contribution that was an offspring of the Laplace–Bayes paradigm. The adherence to this paradigm may also be the reason why Bowley's paper fell into oblivion.

13.4.2.2 Fisher's inference theory

The third revolution in statistical inference that Hald (2007) recognises was initiated by R. A. Fisher in the 1920s. Nearly from the very beginning of his career, R.A. Fisher had attacked the principle of inverse probability and said that it was the greatest flaw in modern science. In 1922, he published a general theory of estimation in a paper titled, *On the Mathematical Foundations of Theoretical Statistics* (Fisher 1922). Hald (2007) claims that for the first time in the history of statistical science, a framework for frequency-based general theory of parametric statistical inference was clearly formulated. In this paper, Fisher defines the three criteria of estimation: consistency, efficiency and sufficiency (Fished 1922, pp. 309–310). From then on, these criteria became the standard properties in the discussion of estimates. In the Laplacian theory, estimation was treated intuitively, without a theoretical framework.

In this paper (Fisher 1922), Fisher created a new technical vocabulary for mathematical statistics to which he still added new concepts in later papers. Today it is not possible to discuss statistical theory without making use of Fisherian terminology. Behind the new words, there was a deeper meaning. For example, he made a clear distinction between sample and population values, conceptually, verbally and notationally. He introduced the term 'parameter', and he coined the term 'statistic' for a function of the sample, designated to estimate the value of a parameter. In this context, Fisher introduced the sampling distribution of a statistic. Other well-known terms that Fisher coined are, for example, null hypothesis, test of significance and level of significance. In the context of experimental design, he introduced randomization. Fisherian terminology gained recognition fairly quickly when the new generation of statisticians adopted it in the late 1920s. In statistical science, it has been dominant since the 1930s. Jeffreys fostered and upheld the Bayes–Laplace theory – as he called it – still in the late 1930s and even later (see Jeffreys 1983), but obviously, he did not gain wider acceptance anymore. In the 1960s, the neo-Bayesian inference theory started to develop to challenge the Fisherian theory (see Fienberg 2006).

Hald (1999) argues that Fisher "single-handedly" created the modern version of the method of maximum likelihood and introduced the likelihood function. Hald (2007) regarded the likelihood function as Fisher's greatest achievement in statistical inference. Another great achievement was the derivation of sampling distributions. Related to this, Fisher started to call the 'universe' 'population', and gave a precise meaning to population. More importantly, he defined population parameters to be constants and therefore, *a priori* distribution became obsolete and unnecessary in estimation. In this context, Fisher introduced a new inference model: repeated sampling from the same distribution. All this made a completely new approach and thinking model for inferential problems.

Moreover, due to Fisher's influence, the research problems of statistical science changed essentially. Before Fisher, the typical problems of statistical science dealt with distributions, fitting distributions, correlation and regression. Yule's textbook (Yule 1911) is an illustrative example of the typical problems of this era. Bowley's report to the ISI was one of the few papers on statistical inference. Fisher defined completely new areas for research, such as design of experiments, analysis of variance, statistical significance testing, estimation theory and inference theory.

At first, Fisher's ideas were largely ignored by the elite of statistical science in Britain (see Stigler 1978). Eventually, however, Fisher's contributions revolutionized almost every part of statistical science, especially the theory of estimation and statistical inference. At first sight, Fisher's revolution put statistical inference into a totally new form. However, Hald (2007) argues: ". . . many of Fisher's asymptotic results are identical to those of Laplace from a mathematical point of view, only a new interpretation is required." Fisher's contributions created an opening also for the development of modern survey sampling, even though Fisher did not directly contribute to it.

In the 1930s, Fisher presented his famous fiducial argument to replace the inverse probability principle, together with a new mode of statistical inference, which he called inductive reasoning. These contributions dealt with statistical inference for hypothetical populations, but fiducial argument proved to be instrumental also in the development of statistical inference for finite populations.

13.4.2.3 Neyman's method of statistical inference for finite populations

Jerzy Neyman obtained his education first in Russia, in the city of Kharkov, and after that in Poland. In 1924, Neyman obtained his doctor's degree from the University of Warsaw, using the work done at the National Agricultural Institute in Bydgoszcz as his thesis. From his first contributions, it can be concluded that he was trained within the Laplace–Bayes paradigm (see Chapter 10.2.1 or Spława-Neyman 1923 and 1925). In the early papers, his thinking model was based on Bernoulli trials. Only in the late 1920s, while he visited the UK, he began to realize that the work of Fisher required a rethinking of the current philosophy of inference (Lehman 2008). As a result, Neyman adopted a new approach to statistical inference.

Neyman gave up the superpopulation approach that was the essence of the Laplace-Bayes paradigm and defined population parameters as constants. Thus, *a* *priori* probabilities were not needed in estimation anymore. In addition, Neyman adopted Fisher's inference model of drawing repeated samples and applied it in sampling from finite populations. Currently, Fisher's and Neyman's inference model of drawing samples repeatedly from the same population is the cornerstone of modern inference theory for both hypothetical and finite populations.

Instead of using Fisher's Maximum Likelihood estimators (ML), Neyman developed Best Linear Unbiased Estimators (BLUE) by using (Gauss-)Markov theory. The BLUE estimators do not entail any assumptions on the distributions of variables as ML estimators did. Therefore, BLUE estimators could be applied practically in any finite populations. Another central contribution of Neyman was the idea of interval estimation or confidence intervals for estimators. Originally, confidence intervals were based on Fisher's fiducial intervals (see Chapter 13.3.3). Later, it appeared that Fisher's fiducial intervals and Neyman's confidence intervals are conceptually different. That was one reason why there appeared a long-lasting controversy between Fisher and Neyman. Another reason for the controversy was Fisher's and Neyman's different views of inductive inference: Fisher's mode was inductive reasoning, and Neyman's was inductive behaviour.

At the beginning of the 1930s, Neyman, together with Egon Pearson, developed the method for hypothesis testing (Neyman and Pearson 1933). Although it differs from Fisher's significance testing, the foundations of Neyman-Pearson test theory lie on Fisher's fundamental ideas about statistical inference.

Fisher's contributions in the 1920s initiated a totally new approach to the development of statistical theory, including inference. In roughly a single decade, it became the dominant approach, and since the 1930s, most of the books on mathematical statistics were based on it. It seems to be warranted to call it a new paradigm for statistical inference. Neyman developed the inference method further to be applied in finite population inference or survey sampling.

The three papers that Neyman published in the 1930s established the foundations of the theory of statistical inference for finite populations (Neyman 1934, 1937, and 1938). Bellhouse (1988) argues that a new paradigm started from Neyman's paper in 1934. The impact of Neyman's papers was not immediate, however. As Hansen and Madow put it, "there was still the need for communication, understanding, acceptance, and the adaptation and extension of the results he [Neyman] had presented." (Hansen and Madow 1976). The prevailing sampling techniques were created in a relatively short period during the 1940s and 1950s. Brewer (1999) argued that the period that started around 1945 was dominated by the randomization paradigm. This coincides with the discovery that inclusion probabilities in sampling need not be equal. At the beginning of the 1950s, this theory was documented in two well-known books (Cochran 1953, Hansen, Hurwitz and Madow 1953), which soon became the standard textbooks in the universities around the world. These books were used in training a new generation of survey statisticians.

Since the beginning of the 1950s, several textbooks have been written with the same approach and even the modern books of sampling techniques (e.g., Särndal et al. 1989 or Lehtonen and Pahkinen 2004) are based on the same basic philosophy, regardless of the fact that the sampling techniques have been elaborated and extended from their origins.

The new inference model replaced the approach that was based on Laplace's inverse probability principle, or the Laplace–Bayes paradigm. In this case, the paradigm shift may be regarded as Kuhnian: in all areas of statistical science, the Fisher–Neyman paradigm replaced the methods that were based on the Laplace–Bayes paradigm. Obviously, after the 1920s, most of the textbooks and most of the training in universities were based on the Fisher–Neyman paradigm. The rapid development of new statistical methods started and the entire field of statistical research changed. The characteristic features Fisher–Neyman paradigm are described in Chapter 11.

The shift of the paradigms also appeared intellectually violent, as can be concluded from the documented discussion after Neyman's presentation to the Royal Statistical Society in 1934 and especially from the discussion after Fisher's presentation one year later. Erich Lehmann, a friend and colleague of Neyman's at the University of California at Berkeley (U.S.A.), wrote in 2008 that the years 1925-1926 were difficult for Neyman and Egon Pearson. They began to realize that Fisher's work required rethinking the current philosophy of inference. According to Lehmann, this was exceptionally difficult for Egon Pearson because his father "was not able or never saw the need to" make such a shift (Lehmann 2008).

Calling the method the Fisher-Neyman paradigm is justified because it was formed by merging two methods: Fisher's ideas for estimation and statistical inference with those of Neyman. Both contributions are vital, but Fisher's revolutionary new idea of statistical inference is focal – or fiducial.

References

Aldrich, J. (2000), "Fisher's 'Inverse Probability' of 1930," *International Statistical Review*, 68, 155–172.

Aldrich, J. (1997), "R. A. Fisher and the Making of Maximum Likelihood 1912–1922," *Statistical Science*, 12, 162–176.

Arbuthnot, J. (1712), "An argument for divine providence, taken from the constant regularity observed in the births of both sexes," *Philosophical Transactions of Royal Society*, 27,186–190. Reprinted in Kendall, M. and Plackett, R. L. eds. (1977), *Studies in the History of Statistics and Probability*, Vol. 2, London: Griffin.

Armatte, M. (2001), "Developments in Statistical reasoning and their links with Mathematics," in *Changing Images in Mathematics*, eds. U. Bottazzini and A. Dahan-Dalmedico, London: Harwood Academic Publish.

Ashton, T. S. (1934), *Economic and Social Investigations in Manchester, 1833 – 1933*, London: P.S. King & Son.

Bayes, T. (1763), "An essay towards solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society*, 53, 370–418.

Bayes, T. (1958), "An Essay Towards Solving a Problem in the Doctrine of Chances (with biographical note by G.A. Barnard)," *Biometrika*, 45, 293–315.

Bellhouse, D. R. (2000), "Survey Sampling Theory over Twentieth Century and its Relation to Computing Technology," *Survey Methodology*, 26, 11–20.

Bellhouse, D. R. (1988), "A Brief History of Random Sampling Methods," in *Handbook of Statistics*, eds. P. Krishnaiah, R. and C. R. Rao: Elsevier Science Publishers, pp. 1–14.

Bellhouse, D. R. (2004), "The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentary of His Birth," *Statistical Science*, 19, 3–43.

Bernoulli, J. (1713), *Ars Conjectandi*, Basilea: Thurnisius.

Block, M. (1886), *Traité théoretique et pratique de Statistique*, Paris: Librairie Guillaumin et C.

Booth, C. (1889), *Life and Labour of the People in London*, London: Macmillan and Co.

Bortkewitch, L. v. (1898), *Das Gesetz der kleinen Zahlen*, Leipzig: Teubner.

Bowley, A. L. (1897), "Relation between the Accuracy of an Average and that of its Constituent Parts," *Journal of the Royal Statistical Society*, 60, 855–866.

Bowley, A. L. (1901), *Elements of Statistics*, London: P. S. King and Son.

Bowley, A. L. (1906), "Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science," *Journal of the Royal Statistical Society*, 69, 548–557.

Bowley, A. L. (1910), *An Elementary Manual of Statistics*, London: Macdonald and Evans.

Bowley, A. L. (1913), "Working-class Households in Reading," *Journal of the Royal Statistical Society*, 76, 672–701.

Bowley, A. L. (1923), "The precision of measurements estimated from samples," *Metron*, 2, 494–500.

Bowley, A. L. (1926), "Measurement of the precision attained in sampling," *Bulletin of the International Statistical Institute*, 22, 6–62.

Bowley, A. L. (1929), "New London Survey," *Journal of the Royal Statistical Society*, 92, 530–547.

Bowley, A. L. (1936), "The Application of Sampling to Economic and Sociological Problems," *Journal of the American Statistical Association*, 31, 474.

Bowley, A. L., and Allen, R. G. D. (1935), *Family Expenditure: a study of its variations*, Westminster: P.S. King & Son.

Bowley, A. L., and Burnett-Hurst, A. R. (1915), *Livelihood and Poverty*, London: Bell & Sons.

Bowley, A. L., and Hogg, M. H. (1925), *Has Poverty Diminished?*, London: P.S. King & Son.

Box, J. (1978), *R. A. Fisher, The Life of a Scientist*, New York: John Wiley & Sons.

Box, J. (1980), "R. A. Fisher and the Design of Experiments, 1922 – 1926," *The American Statistician*, 34, 1–7.

Brewer, K. R. W. (1963), "Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process," *Australian Journal of Statistics*, 5, 93–105.

Brewer, K. (1999), "Design-based or Prediction-based Inference? Stratified Random vs. Stratified Balanced Sampling," *International Statistical Review*, 67, 35–47.

Brewer, K. (2002), *Combining Survey Sampling Inference*, London: Arnold.

Bru, B. (1988), "Estimations Laplaciennes. Un exemple: la recherche de la population d'un grand Empire, 1785–1812 ," in *Estimation et sondages. Cinq contributions à l'histoire de la statistique*, ed. Mairesse, J., Paris: Economica, pp. 7–46.

Bru, B. (2001), "Siméon-Denis Poisson," in *Statisticians of the Centuries*, eds. C. C. Heyde and E. Seneta: Springer, pp. 123–126.

Caradog-Jones, D. (1931), "The Social Survey of Merseyside," *Journal of the Royal Statistical Society*, 94, 218–250.

Caradog-Jones, D. (1934), *The Social Survey of Merseyside, (3 vols)*, London: Hodder and Stoughton.

Caradog-Jones, D. (1941), "Evolution of the Social Survey in England since Booth," *Journal of the Royal Statistical Society*, 104, 818–826.

Chang, W. (1976), "Statistical Theories and Sampling Practice," in *On the History of Statistics and Probability*, ed. D. B. Owen, New York: Marcel Dekker, Inc., pp. 299–316.

Chatterjee, S. K. (2003), *Statistical Thought: A perspective and History*, Oxford: Oxford University Press.

Chauvet, G. (2009), "Stratified balanced sampling," *Survey Methodology*, 35, 115–119.

Cochran, W. G. (1939), "The Use of the Analysis of Variance in Enumeration by Sampling," *Journal of the American Statistical Association*, 492–510.

Cochran, W. G. (1942), "Sampling Theory when the Sampling Units are of Unequal Sizes," *Journal of the American Statistical Association*, 37, 199–212.

Cochran, W. G. (1946), "Relative accuracy of systematic and stratified random samples for a certain class of populations," *Annals of Mathematical Statistics*, 17, 164–177.

Cochran, W. (1953), *Sampling Techniques*, New York: John Wiley & Sons.

Cochran, W. G. (1978), "Laplace's Ratio Estimator," in *Contributions to Survey Sampling and Applied Statistics*, ed. H. A. David, New York: Academic Press, pp. 3–10.

Cramér, H. (1970), *Random Variables and Probability Distributions (3rd edition)*, Cambridge: Cambridge University Press. (1st edition in 1937).

Dale, A. I. (1999), *The History of Inverse Probability (2nd ed.)*, New York: Springer-Verlag.

David, H. A. (1984), "The Iowa State Statistical Laboratory: Antecedents and Early Years, 1914–47," in *Proceedings 50th Anniversary Conference Iowa State Statistical Laboratory*, eds. H. A. David and H. T. David: Iowa State University Press.

Deming, W. E. (1950), *Some Theory of Sampling*, New York: John Wiley & Sons.

Deming, W. E. (1953), "On the Distinction between enumerative and analytical Surveys," *Journal of the American Statistical Association*, 48, 244–255.

Desrosières, A. (1998), *The Politics of Large Numbers; A History of Statistical Reasoning*, Cambridge: Harvard University Press.

Deville, J.-C., and Tillé, Y. (2004), "Efficient balanced sampling: The cube method," *Biometrika*, 91, 893–912.

Didier, E. (2002), "Sampling and Democracy: Representativeness in the First United States Surveys," *Science in Context*, 15, 427–445.

Edwards, A. W. F. (1974), "The history of likelihood," *International Statistical Review*, 42, 9–15.

Edwards, A. W. F. (1997), "What Did Fisher Mean by "Inverse Probability" in 1912 – 1922?," *Statistical Science*, 12, 177–184.

Egdeworth, F. Y. (1906), "The Generalized Law of Error, or Law of Great Numbers," *Journal of the Royal Statistical Society*, 69, 497–539.

Egdeworth, F. Y. (1907), "On the Representation of Statistical Frequency by a Series," *Journal of the Royal Statistical Society*, 70, 102–109.

Egdeworth, F. Y. (1908), "On the Probable Errors of Frequency-Constants," *Journal of the Royal Statistical Society*, 71, 381–397.

Egdeworth, F. Y. (1909), "Addendum on "Probable Errors of Frequency-Constants"," *Journal of the Royal Statistical Society*, 72, 81–90.

Engel, E. (1857), "Die Productions- und Consumtionsverhältnisse des Königreichs Sachsens," *Zeitschrift des statistischen Bureaus des Königlich Sächsischen Ministerium des Inneren*. No. 8, 9.

Engel, E. (1861), *Die Methoden der Volkszählung*, Berlin: Verlagsbuchhandlung des Königlich Preussischen Statistischen Bureaus.

Engel, E. (1863), *Der Internationale Statistische Kongreß in Berlin*, Berlin: Königlichen Geheimen Ober-Hofbuchdruckerei.

Engel, E. (1864), "Die Beschlüsse des Internationalen Statistischen Kongresses in Seiner fünften Sitzungsperiode," *Verlagsbuchhandlung des Königlich Preußischen Statistischen Bureaus*.

Engel, E. (1866), *Der Preis der Arbeit. Zwei Vorlesungen*, Berlin: Habel.

Engel, E. (1871), "Systeme der Demologie," in *Das Statistische Seminar und das Studium der Statistik überhaupt*, Berlin: Königlich Preußische Statistische Bureau, pp. 198–210.

Engel, E. (1883), *Der Werth Des Menschen, Part 1:* Der Kostenwerth des Menschen, Berlin: Verlag von Leonhard Simion.

Fienberg, S. E. (2006), "When Did Bayesian Inference Become "Bayesian"?," *Bayesian Analysis*, 2006, 1–40.

Fienberg, S. E., and Tanur, J. M. (1966), "Reconsidering the Fundamental Contributions of Fisher and Neyman in Experimentation and Sampling," *International Statistical Review*, 64, 237–253.

Fischer, H. (2001), "Pierre-Simon Marquis de Laplace," in *Statisticians of the Centuries*, eds. C. Heyde and E. Seneta, New York: Springer-Verlag.

Fischer, H. (2010), *A History of the Central Limit Theorem; From Laplace to Donsker*, Heidelberg: Springer.

Fisher, R. A. (1912), "On an absolute criterion for fitting frequency curves," *Messenger of Mathematics*, 41, 155–160.

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society, A*, 222, 309–368.

Fisher, R. A. (1925a), "Theory of Statistical Estimation," *Proceedings of Cambridge Philosophical Society*, 22, 700–725.

Fisher, R. A. (1925b), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Fisher, R. A. (1930), "Inverse Inference," *Proceedings of Cambridge Philosophical Society*, 26, 528–535.

Fisher, R. A. (1935a), "The Logic of Inductive Inference," *Journal of the Royal Statistical Society*, 98, 39–82.

Fisher, R. A. (1935b), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Fisher, R. A. (1936), "Uncertain Inference," *Proceedings of the American Academy of Arts and Science*, 71, 245–258.

Fisher, R. A. (1955), "Statistical Methods and Scientific Induction," *Journal of the Royal Statistical Society, Ser. B*, 17, 69–78.

Ford, P. (1934), *Work and Wealth in a Modern Port: an Economic Survey of Southampton*, London: George Allen and Unwin.

Frankel, L. R., and Stock, J. S. (1941), "On the Sample Survey of Unemployment," *Journal of the American Statistical Association*, 36, 77–80.

Gallup, G. (1976), *The Sophisticated Poll Watcher's Guide*, Ephrata: Science Press.

Galton, F. (1869), *Hereditary Genius*, London: Macmillan and Co.

Galton, F. (1889), *Natural Inheritance*, London: Macmillan and Co.

George, B. F. (1936), "A Sample Investigation of the 1931 Population Census," *Journal of the Royal Statistical Society*, 99, 147.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L. (1989), *The Empire of Chance*, Cambridge: Cambridge University Press.

Gini, C., and Galvani, L. (1929), "Di una applicazione del metodo rappresentativo all'ultimo censimento Italiano della popolazione," *Annali di Statistica*, VI, 4, 1–107.

Gouraud, C. (1848), *Histoire de calcul des probabilités depuis ses origines jusqu'à nos jours*, Paris: Auguste Durand.

Graunt, J. (1662), *Natural and Political Observations upon the Bills of Mortality*, London: John Martyn.

Greenwood, M., and Isserlis, L. (1927), "A historical note on the problem of small samples," *Journal of the Royal Statistical Society*, 90, 347–352.

Grier D., "The Origins of Statistical Computing, " *http://www.amstat.org/about/statisticians/index.cfm?fuseaction=papers*

Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge: Cambridge University Press.

Hacking, I. (1975), *The Emergence of Probability*, Cambridge: Cambridge University Press.

Hacking, I. (1990), *The Taming of Chance*, Cambridge: Cambridge University Press.

Hald, A. (1990), *History of Probability and Statistics and Their Applications before 1750*, New York: John Wiley & Sons.

Hald, A. (1998), *A History of Mathematical Statistics from 1750 to 1930*: John Wiley & Sons.

Hald, A. (1999), "On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares," *Statistical Science*, 14, 214-222.

Hald, A. (2007), *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713–1935*, New York: Springer.

Hansen, M. H. (1987), "Some History and Reminiscences on Survey Sampling," *Statistical Science*, 2, 180-190.

Hansen, M. H., and Hurwitz, W. N. (1942), "Relative Efficiencies of Various Sampling Units in Population Inquiries," *Journal of the American Statistical Association*, 37, 89-94.

Hansen, M. H., and Hurwitz, W. N. (1943), "On the theory of sampling from a finite population," *Annals of Mathematical Statistics*, 14, 333-362.

Hansen, M. H., Hurwitz, W. N., Madow, W. G. (1953), *Survey sampling methods and theory, Vols. I and II*, New York: John Wiley & Sons.

Hansen, M. H., Dalenius, T., and Tepping, B. J. (1985), "The Development of Sample Surveys of Finite Populations," in *A Celebration of Statistics; The ISI Centenary Volume*, eds. A. C. Atkinson and S. E. Fienberg, New York: Springer-Verlag, pp. 327–354.

Hansen, M. H., and Madow, W. G. (1978), "Estimation and Inferences from Sample Surveys: Some Comments on Recent Developments," in *Survey Sampling and Measurement*, ed. N. K. Namboodiri: Academic Press.

Hansen, M. H., and Madow, W. G. (1976), "Some important Events in the Historical Development of Sample Surveys," in *On the History of Statistics and Probability*, ed. D. B. Oven: Marcel Dekker, Inc.

Hertz, S. (2001), "Georg von Mayr," in *Statisticians of the Centuries*, eds. C. C. Heyde and E. Seneta, New York: Springer, pp. 219-222.

Heyde, C. C., and Seneta, E. (eds.) (2001), *Statisticians of the History*, New York: Springer.

Heywood, J. (1838), "Report of an Inquiry, conducted from House to House, into the State of 176 families in Miles Platting, within the borough of Manchester, in 1837," *Journal of the Statistical Society of London*, 1, 34-36.

Hill, I. D. (1984), "Statistical Society of London – Royal Statistical Society. The first 100 years: 1834-1934," *Journal of the Royal Statistical Society, Ser. A*, 147, 130-139.

Hilton, J. (1924), "Enquiry by Sample, an Experiment and its Results," *Journal of the Royal Statistical Society*, 87.

Hilton, J. (1928), "Some Further Enquiries by Sample," *Journal of the Royal Statistical Society*, 91, 519–530.

Hogg, M. H. (1930), "Sources of Incomparability and Error in Employment-Unemployment Surveys," *Journal of the American Statistical Association*, 25, 284–294.

Hogg, M. H. (1932), *The Incidence of Work Shortage*, New York: Russell Sage Foundation.

Horvitz, D. G., Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685.

Hubback, J. A. (1927), "Sampling for rice yield in Bihar and Orissa", *Imperial Agricultural Research Institute Bulletin*, 166. (Reprinted in Sankya, 1946, vol. 7).

Hume, D. (1739), *A Treatise of Human Nature: Being an Attempt to introduce the experimental Method of Reasoning into Moral Subjects*, London. John Noon. Reprinted by Clarendon Press, Oxford by comments Selby-Bigge, L. A. (1888).

Hume, D. (1748), *An Enquiry Concerning Human Understanding*, London. Reprinted by Clarendon Press, Oxford by comments Selby-Bigge, L. A. (2nd edition 1902). Published in 1748 under the title *Philosophical Essay Concerning Human Understanding.*

Jeffreys, H. (1983), *Theory of Probability*, 3rd edition (1st edition 1939), London: Clarendon Press.

Jensen, A. (1926), "The Representative Method in Practice," *Bulletin of the International Statistical Institute*, 22, 386–420.

Jensen, A. (1928), "Purposive Selection," *Journal of the Royal Statistical Society*, 91, 541–547.

Jessen, R. J. (1942), "Statistical investigations of a sample survey for obtaining farm facts," *Iowa Agricultural Experimental Station Research Bulletin*, 303.

Kaufman, A. (1913), *Theorie und Methoden der Statistik*, Tübingen: Verlag von J. C. Mohr. (Originally Kaufman, A. (1912), Teoria i metody statistiki (in Russian), Moscow, izd. I.D. Sytina.

Kaufman, A. (1918), "Russia", in *The History of Statistics; their Development and Progress in many Countries*, Koren J. (ed.). New York: The MacMillan Company, 467–532.

Kaufman, A. (1922), *Statistical Science in Russia: Theory and Methods,1806–1917, (in Russian,"Statisticheskaia nauka v Rossii: teoriia I metodologiia, 1806–1917")*, Moscow, TsSU

Kendall, M. (1960), "Where shall the history of statistics begin?," *Biometrika*, 47, 447–449.

Kendal, M., and Plackett, R. L., eds. (1977), *Studies in the History of Statistics and Probability, Vol. 2*. London: Griffin.

Keverberg, Baron de (1827), "Notes sur Quetelet," *Nouveaux Mémoires de l'Academie royal des sciences et belles-lettres de Bruxelles*, 4, 175–192

Kiaer, A. N. (1895), "Observations et expériences concernant des dénombrements représentatives," *Bulletin of the International Statistical Institute*, 9, 176–183.

Kiaer, A. N. (1897a), "The Representative Method of Statistical Surveys," *Reprint of Kiaer's paper from the Norwegian Academy of Science and Letters, 1997. Oslo: Statistics Norway.*

Kiaer, A. N. (1897b), "Sur les méthodes représentatives ou typologiques appliquées à la statistique," *Bulletin of the International Statistical Institute*, 11, 180–185.

Kiaer, A. N. (1897c), "Nye undersökelse angaende indtägts- og formuesforhold i Norge," Kristiania: *Statökonomisk tidsskift*, 1-26.

Kiaer, A. N. (1899), "Die Repräsentative Untersuchungsmethode," *Allgemeines Statistisches Archiv*, 5, 1–22.

Kiaer, A. N. (1901), "Sur les méthodes représentatives ou typologiques," *Bulletin of the International Statistical Institute*, 3, 66–78.

Kiaer, A. N. (1903), "La méthode représentative," *Bulletin of the International Statistical Institute*, 14, 127–133.

Kiaer, A. N. (1905), "Untitled speech with discussion," *Bulletin of the International Statistical Institute*, 14, 119–134.

Kish, L. (1995), "The Hundred Years' War of Survey Sampling," *Statistics in Transition*, 2(5), 813-830. Reprinted in Galton, G., and Heeriga, s., eds. (2003). *Leslie Kish Selected Papers*, New Jersey: John Wiley & Sons,

Kish, L. (2002), "New Paradigms (Models) for Probability Sampling," *Survey Methodology*, 28, 31-34.

Kohn, S. S. (1922), "Die allrussischen Landwirtschaftszählungen von 1916 und 1917," *Nordisk Statistisk Tidskrift*, B.1, 125–134.

Kolmogorov, A. N. (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin: Ergeb. Math. No. 3. (English translation by N. Morrison, 1950, New York: Chelsea Publishing Co.)

Kovalevsky, A. G. (1924), "Basic Theory of Sampling Methods " (in Russian "Osnovy teorii vyborochnogo metoda"), Saratov: *Utchenie Zapiski Gosudarstvennogo Saratovskogo universiteta,II,vyp 4, Fakultet Khozaistavai i Prava, 60–138.* (Reproduced in a digital form in HELDA – The Digital Repository of University of Helsinki, http://hdl.handle.net/10138/25695)

Kruger, L., Daston, L., and Heidelberger, M. (eds.) (1987), *The probabilistic revolution, Vol 1: Ideas in history*, Cambridge, USA: MIT Press.

Kruger, L., Gingerenzer, G., and Morgan, M. (eds.) (1989), *The probabilistic revolution, Vol 2: Ideas in the sciences*, Cambridge, USA: MIT Press.

Kruskal, W., and Mosteller, F. (1980), "Representative Sampling, IV: The History of the Concept in Statistics, 1895–1939," *International Statistical Review*, 48, 169–195.

Kuhn, T. (1962), *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Lambert, J. (1760), *Photometria sive de Mensura et Gradibus Luminis, Colorum et Umrae*, Ausburg : Detleffsen.

Laplace, P. S. (1774), "Mémoire sur la probabilité des causes par les événements," *Mémoires de l'Académie des Sciences de Paris*, 6, 621–656.

Laplace, P. S. (1778), "Mémoire sur les probabilités," in *Mémoires de l'Académie Royale des Sciences de Paris*, Paris, pp. 1–98.

Laplace, P. S. (1783), "Sur les naissances, les mariages et les morts," *Mémoires de l'Académie Royale des Sciences de Paris*, 693–702.

Laplace, P. S. (1810), "Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur application aux probabilités," *Mémoires de l'Académie Royale des Sciences de Paris*, 301–345.

Laplace, P. S. (1812a), *Théorie analytique des probabilités*, Paris: Gauthier-Villar.

Laplace, P. S. (1812b), *Essai philosophique sur les probabilités*, Paris: Bougois.

Lehmann, E. L. (2008), *Reminiscences of a Statistician*, New York: Springer.

Lehtonen, R., and Pahkinen, E. (2004), *Practical Methods for Design and Analysis of Complex Surveys (2nd ed.)*, Chichester: John Wiley & Sons.

Lexis, W. (1875), *Einleitung in die Theorie der Bevölkerungsstatistik*, Strassburg:Trübner.

Lexis, W. (1877), *Zur Theorie der Massenerscheinungen in der meschlichen Geselschaft*, Freiburg: Fr. Wagner'sche Buchhandlung.

Lexis, W. (1879), "Über die Theorie der Stabilität statistischer Reihen," in *Jahrbücher für Nationalökonomie und Statistik* (Vol. 32), pp. 60–98.

Lexis, W. (1903), *Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik*, Jena: Fischer.

Lie, E. (2002), "The Rise and Fall of Sampling in Norway, 1875 – 1906," *Science in Context*, 15, 385–409.

Llewellyn-Smith, G. C. B. (1929), "The New Survey of London Life and Labour," *Journal of the Royal Statistical Society*, 92, 530–546.

Madow, W. G., and Madow, L. H. (1944), "On the theory of systematic sampling," *Annals of Mathematical Statistics*, 15, 1–24.

Mahalanobis, P. (1939), "Professor Ronald Alymer Fisher," *Sankhya*, 4.

Mahalanobis, P. (1944), "On Large-scale Sample Surveys," *Philosophical Transactions of the Royal Society*, 231, 329–451.

Malaguerra, C. (2000), "A Perspective in Time; Official Statistics in the 20th Century: Landmarks and Challenges," in *Statistics, a challenge for the future; Proceedings of the 85th DGINS conference, The Hague, May 1999*, European Communities, Luxembourg

Markov, A. A. (1912), *Wahrscheinlichkeitsrechnung.* (Translation of 2nd Russian edition, 1st ed. 1900, 2nd ed. 1908), Leipzig: Teubner.

Matuszewski, T., Neyman, J., and Supiska, J. (1935), "Statistical Studies in Questions of Bacteriology," *Journal of the Royal Statistical Society, Ser. B*, 2, 63–82.

Mayr, G. v. (1895), *Statistik und Gesellschaftslehre, vol I*, Freiburg: Mohr.

Mayr, G. v. (1914), "Strafrechtspflege, Kriminalstatistik, Kriminalpolitik," *Mitteilungen der Internationalen Kriminalistischer Vereinigung*, 21, 404–414.

Mespoulet, M. (2002), "From Typical Areas to Random Sampling: Sampling Methods in Russia from 1875 to 1930," *Science in Context*, 15, 411–425.

Missiakoulis, S. (2010), "Cecrops, King of Athens: the First (?) Recorded Population Census in History," *International Statistical Review*, 78, 413–418

Moivre, A. de (1718), *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play.* London: Pearson.

Mouat, F. (1885), "History of the Statistical Society," *Journal of the Royal Statistical Society*, 14–71.

Murthy, M. N. (1964), "On Mahalanobis' contributions to the development of sample survey theory and methods," in *Contributions to Statistics*, Calcutta: Statistical Publishing Society.

Narain, R. D. (1951), "On sampling without replacement with varying probabilities," *Journal of the Indian Society of Agricultural Statistics*, 3, 169–174.

Neyman, J. (1933), *Zarys teorii i praktyki badania struktury ludno ci metoda reprezentacyjna,*Warzawa: Institytut Spraw Społecznych. *(In Polish with an English summary: An Outline of the Theory and Practice of Representative Method, Applied in Social Research,* Warsaw: Institute for Social Problems.)

Neyman, J., and Pearson, E. (1933), "On the Problem of the most Efficient Test of Statistical Hypotheses," *Philosophical Transactions of the Royal Society, A*, 289–337.

Neyman, J. (1934), "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion)," *Journal of the Royal Statistical Society*, 97, 558–606.

Neyman, J. (1937), "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Philosophical Transactions of the Royal Society*, 333–380.

Neyman, J. (1938), "Contributions to the Theory of Sampling Human Populations," *Journal of the American Statistical Association*, 33, 101–116.

Neyman, J. (1952), "Recognition of priority," *Journal of the Royal Statistical Society*, 115, 602.

Neyman, J. (1971), "Foundations of the Behavioristic Statistics," in *Foundations of Statistical Inference*, eds. V. P. Godambe and D. A. Sprott: Holt, Rinehart and Winston of Canada.

O'Muircheartaigh, C. (2005), "Balancing Statistical Theory, Sampling Concepts, and Practicality in the Teaching of Survey Sampling," *Bulletin of the International Statistical Institute*, 55.

O'Muircheartaigh, C., and Wong, S. (1981), "The Impact of Sampling Theory on Survey Sampling Practice: A Review," *Bulletin of the International Statistical Institute.*

Olkin, I. (1987), "A Conversion with Morris Hansen," *Statistical Science*, 2, 162–179.

Pearson, E. S., and Kendall, M. (eds.) (1970), *Studies in the History of Statistics and Probability*, London: Griffin.

Pearson, K. (1892), *The Grammar of Science*, London: W. Scott.

Pearson, K. (1896), "Mathematical Contributions to the Theory of Evolution, III. Regression, Heredity, and Panmixia," *Philosophical Transactions of the Royal Society, A*, 1987, 253–318.

Pearson, K. (1902), "On the systematic fitting of curves to observations and measurements," *Biometrika*, 1, 265–303.

Pearson, K. (1906), "On the Curves which are the most suitable for describing the Frequency of Random Samples of a Population," *Biometrika*, v, 172.

Pearson, K. (1920), "The Fundamental Problem of Practical Statistics," *Biometrika*, 13.

Pearson, K. (1928), "On a Method Ascertaining Limits to the Actual Number of Marked Members in a Population of Given Size from a Sample," *Biometrika*, 20A, 149–174.

Pearson, K. (1978), "The history of statistics in the 17th and 18th centuries against the changing background of intellectual, scientific and religious though," in *Lectures by Karl Pearson given at the University College London during the academic sessions 1921–1933*, ed. E. Pearson, London: Charles Griffin & Co.

Perrot, J.-C. (1984), "The Golden Age of Regional Statistics (Year IV – 1804)," in *State and Statistics in France 1789 – 1815*, eds. J.-C. Perrot and S. J. Woolf, Chur: Harwood Academic Publishers, pp. 1–77.

Perrot, J.-C., and Woolf, S. J. (1984), *State and Statistics in France 1789 – 1815* (Vol. 2), eds. Revel, J. and M. Augé, Chur: Harwood Academic Publishers.

Plackett, R. L. (1958), "The principle of the arithmetic mean," *Biometrika*, 45, 130–135. (Reprinted in Kendal, M., and Plackett, R. L., eds. (1977), *Studies in the History of Statistics and Probability, Vol. 2*. London: Griffin.)

Poisson, S. D. (1829), "Sur la proportion des naissances des filles et des garçons," *Mémoires de l'Académie des Sciences de Paris*, 9, 239.

Poisson, S. D. (1837), *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile, précédées des Règles Générales du Calcul des Probabilités*, Paris: Bachelier.

Pollock, K. H. (1981), "Capture-recapture models: A review of current methods, assumptions and experimental design," in *Estimating Numbers of Terrestrial Birds* (Vol. 6), eds. J. Scott and C. Ralph.

Porter, T. M. (1986), *The Rise of Statistical Thinking 1820–1900*, Princeton: Princeton University Press.

Porter, T. M. (1995), *Trust in Numbers*, Princeton: Princeton University Press.

Porter, T. M. (2004), *Karl Pearson: The Scientific Life in a Statistical Age*, Princeton: Princeton University Press.

Pólya, G. (1920), " Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem," *Mathematische Zeitschrift*, 8, 171–181.

Quetelet, A. (1835), *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*, Paris: Bachelier.

Quetelet, A. (1869), *Physique sociale et essai sur le développement des facultés de l'homme (Tome 1 / 2)*, Bruxelles: C. Muquart.

Quetelet, A. (1848), *Du système social et des lois qui régissent*, Paris: Guillaum.

Quetelet, A. (1849), *Lettres à S.A.R. le duc régnant de Saxe-Coburg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politiques s*, Bruxelles: Reprinted by Arno Company *(Letters addressed to H.R.H. the Grand Duke of Saxe Coburg and Gotha, On the theory of probabilities, as applied to the moral and political science)*, New York 1981.

Rauchberg, H. (1896), La machine électrique à recensement, Expériences et améliorations, *Bulletin de l'Institut International de Statistique, Tome IX, Deuxième et dernière livraison*, 249–257.

Rao, B. L. S. P. (2006), "About statistics as a discipline in India," *Electronic Journal for History of Probability and Statistics*, 2.

Rao, J. N. K. (2005), "Interplay Between Sample Survey Theory and Practice: An Appraisal," *Survey Methodology*, 31, 117–138.

Rao, R. R. (1992), "R. A. Fisher: The Founder of Modern Statistics," *Statistical Science*, 7, 34–48.

Rietz, H. L. (1924), "On Certain Topics in the Mathematical Theory of Statistics," *Bulletin of the American Mathematical Society*, 417–453.

Rowntree, S. (1901), *Poverty, A study of Town Life*, London: Macmillan and Co.

Royall, R. (1970), "On finite population sampling theory under certain linear regression models," *Biometrika*, 57, 377–387.

Royall, R., and Herson. (1973), "Robust Estimation in finite populations I," *Journal of the American Statistical Association*, 68, 880–889.

Salmon, W. (1967), *The Foundations of Scientific Inference*, Pittsburg: University of Pittsburgh Press.

Salsburg, D. (2001), *The Lady Tasting Tea*, New York: Henry Holt and Company.

Schneider, I. (2006), "Direct and indirect influences of Jakob Bernoulli's *Ars conjectandi* in 18th century Great Brittain," Electronic *Journal for History of Probability and Statistics*, 2/1a. (http://www.jehps.net/juin2006.html)

Seidenfeld, T. (1992), "R. A. Fisher's Fiducial Argument and Bayes' Theorem," *Statistical Science*, 7, 358–368.

Seneta, E. (1985), "A sketch of the History of Survey Sampling in Russia," *Journal of the Royal Statistical Society, Ser. A*, 148, 118–125.

Seng, Y. (1951), "Historical Survey of the Development of Sampling Theories and Practice," *Journal of the Royal Statistical Society, Ser. A*, 114, 440–457.

Shafer, G. (1982), "Bayes's two arguments for the rule of conditioning," *Annals of Mathematical Statistics*, 10, 1075–1089.

Simpson, T. (1755), "On the Advantage of Taking the Mean of a Number of Observations in practical Astronomy," *Philosophical Transactions of the Royal Society*, 49, 82–93.

Smith, T. F. M. (1976), "The Foundations of Sampling: a Review (with discussion)," *Journal of the Royal Statistical Society, Ser. A*, 139, 183–204.

Spława-Neyman, J. (1925), "Contributions of the theory of small samples drawn from a finite population," *Biometrika*, 17, 472–479.

Spława-Neyman, J. (1923), "On the application of probability theory to agricultural experiments: Essay on principles. (Translated and published in English 1990 in Statistical Science)," *Statistical Science*, 5, 465–472.

Stephan, F. F. (1936), "Practical Problems of Sampling Procedure," *American Sociological Review*, 1, 569–580.

Stephan, F. F. (1940), "Representative Sampling in Large-Scale Surveys," *Journal of the American Statistical Association*, 343–351.

Stephan, F. F. (1941), "Stratification in Representative Sampling," *Journal of Marketing*, 6, 38–46.

Stephan, F. F. (1948), "History of the Uses of Modern Sampling Procedures," *Journal of the American Statistical Association*, 43, 12–39.

Stephan, F. F., Deming, W. E., and Hansen, M. H. (1940), "The Sampling Procedure of the 1940 Population Census," *Journal of the American Statistical Association*, 35, 615–630.

Stigler, S. (1973), "Studies in the History of Probability and Statistics. XXXII; Laplace, Fisher, and the discovery of the concept of sufficiency," *Biometrika*, 60, 439–445.

Stigler, S. (1978), "Mathematical statistics in the early States," *Annals of Mathematical Statistics*, 6, 239–265.

Stigler, S. (1983), "Who discovered Bayes's Theorem?," *The American Statistician*, 37, 290–296.

Stigler, S. (1982), "Thomas Bayes's Bayesian Inference," *Journal of the Royal Statistical Society, Ser. A*, 145, 250–258.

Stigler, S. (1986), *The History of statistics; the Measurement of Uncertainty before 1900*, Cambridge: The Belknap Press of Harvard University Press.

Stigler, S. (1986), "Laplace's 1774 memoir on inverse probability," *Statistical Science*, 1, 359–378.

Stigler, S. (1999), *Statistics on the Table*, Cambridge: Harvard University Press.

Stigler, S. (2005), "Fisher in 1921," *Statistical Science*, 20, 32–49.

Student. (1908a), "The Probable Error of a Mean," *Biometrika*, 6, 1-25.

Student. (1908b), "Probable Error of a Correlation Coefficient," *Biometrika*, 6, 302–310.

Sukhatme, P. V. (1935), "Contributions to the Theory of the Representative Method," *Journal of the Royal Statistical Society, Supp. 2*, 253–268.

Sukhatme, P. V. (1966), "Major Developments in Sampling Theory and Practice," in *Research Papers in Statistics: Festschrift for J. Neyman*, ed. F. David, New York: John Wiley & Sons, pp. 367–409.

Särndal, C.-E. (2007), Personal communication.

Särndal, C.-E., Swenson, B., and Wretman, J. (1989), *Model Assisted Survey Sampling*: Springer-Verlag.

Süssmilch, J. P. (1741), *Die göttliche Ordnung in den Verändrungen des menschlichen Geschlechts: aus der Geburt, dem Tode und der Fortpflanzung desselben erwiesen.*

Tchuprov, A. A. (1910a), *Essays in the Theory of Statistics (in Russian, "Otcherki po Teorii Statistiki")*, St. Petersburg: M. i S. Sabasnikovyh.

Tchuprov, A. A. (1910b), "Die repräsentative Untersuchung," in *Vortrag auf der 12 Jahresversamlung russischer Naturforscher und Arzte.*

Tchuprov, A. A. (1918), "On the Mathematical Expectation of the Moments of Frequency Distributions, I," *Biometrika*, 12, 140–169.

Tchuprov, A. A. (1919), "Zur Theorie der Stabilität statistischer Reihen," *Skandinavisk Aktuarietidskrift*, 199–263.

Tchuprov, A. A. (1920), "On the Mathematical Expectation of the Moments of Frequency Distributions, II," *Biometrika*, 13, 283–295.

Tchuprov, A. A. (1922), "Das Gesetz der grossen Zahlen und der stochastisch-statistiche Standpunkt in der modernen Wissenschaft," *Nordisk Statistisk Tidskrift*, 1, 39–67.

Tchuprov, A. A. (1923a), "On the mathematical expectation of the moments of frequency distributions in the case of correlated observations, (I – III)," *Metron*, 2, 461-493.

Tchuprov, A. A. (1923b), "On the mathematical expectation of the moments of frequency distribution in the case of correlated observations (IV–VI)," *Metron*, 2, 636–680.

Tchuprov, A. A. (1926), "Theory of Stability of Statistical Series (in Swedish "Teorien för statistiska räckors stabilitet")," *Nordisk Statistisk Tidskrift*, 5, 195–212.

Thatcher, A. R. (1964), "Relationship between Bayesian and confidence limits for prediction," *Journal of the Royal Statistical Society, Ser. B*, 26, 176–192.

Todhunter, I. (1865), *A History of the Mathematical Theory of Probability; From the time of Pascal to that of Laplace*, Cambridge and London: Macmillan and Co. (Reprinted by Chelsea, New York, 1965.)

Tönnies, F. (1925), "Moralstatistik," in *Soziologische Studien und Kritiken, 3 vols*, ed. G. Fischer, Jena, pp. 117–132.

Watson, G. S. (1982), "William Gemmel Cochran 1909–1980," *The Annals of Statistics*, 10, 1–10.

Weatherford, T. (1982), *Philosophical Foundations of Probability Theory*, London: Routledge & Kegan Paul.

Westergaard, H. (1932), *Contributions to the History of Statistics*, London: P.S. King & Sons, Ltd.

Wilcox, W. F. (1934), "Note on the chronology of statistical societies," *Journal of the American Statistical Association*, 29, 418–420.

Woolf, S. J. (1984), "Towards the History of the Origins of Statistics: France, 1789 – 1815," in *State and Statistics in France 1789 – 1815*, eds. J.-C. Perrot and S. J. Woolf, Chur: Harwood Academic Publishers, pp. 79–194.

Yates, F. (1946), "A Review of Recent Statistical Developments in Sampling and Sampling Surveys (with discussion)," *Journal of the Royal Statistical Society*, 109, 12–43.

Yates, F. (1949), *Sampling Methods for Censuses and Surveys*: Charles Griffin & Co.

Yates, F. (1951), "The Influence of Statistical Methods for Research Workers on the Development of the Science of Statistics," *Journal of the American Statistical Association*, 46, 19–34.

Yates, F., and Zacopanay, I. (1935), "The estimation of the efficiency of sampling with special reference to sampling for yield in cereal experiments," *Journal of Agricultural Science*, 25, 543–577.

Yule, G. (1896), "Notes on the History of Pauperism in England and Wales from 1850, Treated by the Method of Frequency-Curves; with an Introduction to the Method," *Journal of the Royal Statistical Society*, 59, 318–347.

Yule, G. (1897), "On the Theory of Correlations," *Journal of the Royal Statistical Society*, 60, 812–851.

Yule, G. (1911), *An Introduction to theory of Statistics*, London: Charles Griffin & Co.

Zabell, S. L. (1989), "R. A. Fisher on the history of inverse probability," *Statistical Science*, 4, 247–263.

Zabell, S. L. (1992), "R. A. Fisher and the Fiducial Argument," *Statistical Science*, 7, 369–387.

Zabell, S. L. (2008), "On Student's 1908 Article "The Probable Error of a Mean"," *The American Statistician*, 103, 1–8.

Zarkovic, B. (1956), "Note on the History of Sampling in Russia," *Journal of the Royal Statistical Society, Ser. A*, 119, 336–338.

Zarkovic, B. (1962), "A supplement to 'Note on the History of Sampling in Russia'," *Journal of the Royal Statistical Society, Ser. A*, 125, 580–582.

TUTKIMUKSIA-SARJA
RESEARCH REPORTS SERIES

Tilastokeskus on julkaissut Tutkimuksia v. 1966 alkaen, v. 1990 lähtien ovat ilmestyneet seuraavat:

164. **Henry Takala,** Kunnat ja kuntainliitot kansantalouden tilinpidossa. Tammikuu 1990. 60 s.

165. **Jarmo Hyrkkö,** Palkansaajien ansiotasoindeksi 1985=100. Tammikuu 1990. 66 s.

166. **Pekka Rytkönen,** Siivouspalvelu, ympäristöhuolto ja pesulapalvelu 1980-luvulla. Tammikuu 1990. 70 s.

167. **Jukka Muukkonen,** Luonnonvaratilinpito kestävän kehityksen kuvaajana. 1990. 119 s.

168. **Juha-Pekka Ollila,** Tieliikenteen tavarankuljetus 1980-luvulla. Helmikuu 1990. 45 s.

169. **Tuovi Allén – Seppo Laaksonen – Päivi Keinänen – Seija Ilmakunnas,** Palkkaa työstä ja sukupuolesta. Huhtikuu 1990. 90 s.

170. **Ari Tyrkkö,** Asuinolotiedot väestölaskennassa ja kotitaloustiedustelussa. Huhtikuu 1990. 63 s.

171. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Nuorten koulutus ja kotitausta. Toukokuu 1990. 115 s.

171b.**Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Education and the family background of the young in Finland. 1990. 115 pp.

172.. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Eriarvoisuus kuoleman edessä. Sosioekonomiset kuolleisuuserot Suomessa 1971–85. Kesäkuu 1990. 145 s.

173. **Jukka Muukkonen,** Sustainable development and natural resource accounting. August 1990. 96 pp.

174. **Iiris Niemi – Hannu Pääkkönen,** Time use changes in Finland in the 1980s. August 1990. 118 pp.

175. **Väinö Kannisto,** Mortality of the elderly in late 19th and early 20th century Finland. August 1990. 50 pp.

176. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Socioeconomic mortality differences in Finland 1971–85. December 1990. 108 pp.

177. **Jaana Lähteenmaa – Lasse Siurala,** Nuoret ja muutos. Tammikuu 1991. 211 s.

178. **Tuomo Martikainen – Risto Yrjönen,** Vaalit, puolueet ja yhteiskunnan muutos. Maaliskuu 1991. 120 s.

179. **Seppo Laaksonen,** Comparative Adjustments for Missingness in Short-term Panels. April 1991. 74 pp.

180. **Ágnes Babarczy – István Harcsa – Hannu Pääkkönen,** Time use trends in Finland and in Hungary. April 1991. 72 pp.

181. **Timo Matala,** Asumisen tuki 1988. Kesäkuu 1991. 64 s.

182. **Iiris Niemi – Parsla Eglite – Algimantas Mitrikas – V.D. Patrushev – Hannu Pääkkönen,** Time Use in Finland, Latvia, Lithuania and Russia. July 1991. 80 pp.

183. **Iiris Niemi – Hannu Pääkkönen,** Vuotuinen ajankäyttö. Joulukuu 1992. 83 s.

183b. **Iiris Niemi – Hannu Pääkkönen – Veli Rajaniemi – Seppo Laaksonen – Jarmo Lauri,** Vuotuinen ajankäyttö. Ajankäyttötutkimuksen 1987–88 taulukot. Elokuu 1991. 116 s.

184. **Ari Leppälahti – Mikael Åkerblom,** Industrial Innovation in Finland. August 1991. 82 pp.

185. **Maarit Säynevirta,** Indeksiteoria ja ansiotasoindeksi. Lokakuu 1991. 95 s.

186. **Ari Tyrkkö,** Ahtaasti asuvat. Syyskuu 1991. 134 s.

187. **Tuomo Martikainen – Risto Yrjönen,** Voting, parties and social change in Finland. October 1991. 108 pp.

188. **Timo Kolu,** Työelämän laatu 1977–1990. Työn ja hyvinvoinnin koettuja muutoksia. Tammikuu 1992. 194 s.

189. **Anna-Maija Lehto,** Työelämän laatu ja tasa-arvo. Tammikuu 1992. 196 s.

190. **Tuovi Allén – Päivi Keinänen – Seppo Laaksonen – Seija Ilmakunnas,** Wage from Work and Gender. A Study on Wage Differentials in Finland in 1985. 88 pp.

191. **Kirsti Ahlqvist,** Kodinomistajaksi velalla. Maaliskuu 1992. 98 s.

192. **Matti Simpanen – Irja Blomqvist,** Aikuiskoulutukseen osallistuminen. Aikuiskoulutustutkimus 1990. Toukokuu 1992. 135 s.

193. **Leena M. Kirjavainen – Bistra Anachkova – Seppo Laaksonen – Iiris Niemi – Hannu Pääkkönen** – Zahari Staikov, Housework Time in Bulgaria and Finland. June 1992. 131 pp.

194. **Pekka Haapala – Seppo Kouvonen,** Kuntasektorin työvoimakustannukset. Kesäkuu 1992. 70 s.

195. **Pirkko Aulin-Ahmavaara,** The Productivity of a Nation. November 1992. 72 pp.

196. **Tuula Melkas,** Valtion ja markkinoiden tuolla puolen. Kanssaihmisten apu Suomessa 1980-luvun lopulla. Joulukuu 1992. 150 s.

197. **Fjalar Finnäs,** Formation of unions and families in Finnish cohorts born 1938–67. April 1993. 58 pp.

198. **Antti Siikanen – Ari Tyrkkö,** Koti – Talous – Asuntomarkkinat. Kesäkuu 1993. 167 s.

199. **Timo Matala,** Asumisen tuki ja aravavuokralaiset. Kesäkuu 1993. 84 s.

200. **Arja Kinnunen,** Kuluttajahintaindeksi 1990=100. Menetelmät ja käytäntö. Elokuu 1993. 89 s.

201. **Matti Simpanen,** Aikuiskoulutus ja työelämä. Aikuiskoulutustutkimus 1990. Syyskuu 1993. 150 s.

202. **Martti Puohiniemi,** Suomalaisten arvot ja tulevaisuus. Lokakuu 1993. 100 s.

203. **Juha Kivinen – Ari Mäkinen,** Suomen elintarvike- ja metallituoteteollisuuden rakenteen, kannattavuuden ja suhdannevaihteluiden yhteys; ekonometrinen analyysi vuosilta 1974 – 1990. Marraskuu 1993. 92 s.

204. **Juha Nurmela,** Kotitalouksien energian kokonaiskulutus 1990. Marraskuu 1993. 108 s.

205a. **Georg Luther,** Suomen tilastotoimen historia vuoteen 1970. Joulukuu 1993. 382 s.

205b. **Georg Luther,** Statistikens historia i Finland till 1970. December 1993. 380 s.

206. **Riitta Harala – Eva Hänninen-Salmelin – Kaisa Kauppinen-Toropainen – Päivi Keinänen – Tuulikki Petäjäniemi – Sinikka Vanhala,** Naiset huipulla. Huhtikuu 1994. 64 s.

207. **Wangqiu Song,** Hedoninen regressioanalyysi kuluttajahintaindeksissä. Huhtikuu 1994. 100 s.

208. **Anne Koponen,** Työolot ja ammattillinen aikuiskoulutus 1990. Toukokuu 1994. 118 s.

209. **Fjalar Finnäs,** Language Shifts and Migration. May 1994. 37 pp.

210. **Erkki Pahkinen – Veijo Ritola,** Suhdannekäänne ja taloudelliset aikasarjat. Kesäkuu 1994. 200 s.

211. **Riitta Harala – Eva Hänninen-Salmelin – Kaisa Kauppinen-Toropainen – Päivi Keinänen – Tuulikki Petäjäniemi – Sinikka Vanhala,** Women at the Top. July 1994. 66 pp.

212. **Olavi Lehtoranta,** Teollisuuden tuottavuuskehityksen mittaminen toimialatasolla. Tammikuu 1995. 73 s.

213. **Kristiina Manderbacka,** Terveydentilan mittarit. Syyskuu 1995. 121 s.

214. **Andres Vikat,** Perheellistyminen Virossa ja Suomessa. Joulukuu 1995. 52 s.

215. **Mika Maliranta,** Suomen tehdasteollisuuden tuottavuus. Helmikuu 1996. 189 s.

216. **Juha Nurmela,** Kotitaloudet ja energia vuonna 2015. Huhtikuu 1996. 285 s.

217. **Rauno Sairinen,** Suomalaiset ja ympäristöpolitiikka. Elokuu 1996. 179 s.

218. **Johanna Moisander,** Attitudes and Ecologically Responsible Consumption. August 1996. 159 pp.

219. **Seppo Laaksonen** (ed.), International Perspectives on Nonresponse. Proceedings of the Sixth International Workshop on Household Survey Nonresponse. December 1996. 240 pp.

220. **Jukka Hoffrén,** Metsien ekologisen laadun mittaaminen. Elokuu 1996. 79 s.

221. **Jarmo Rusanen – Arvo Naukkarinen** – Alfred Colpaert – Toivo Muilu, Differences in the Spatial Structure of the Population Between Finland and Sweden in 1995 – a GIS viewpoint. March 1997. 46 pp.

222. **Anna-Maija Lehto,** Työolot tutkimuskohteena. Marraskuu 1996. 289 s.

223. **Seppo Laaksonen** (ed.), The Evolution of Firms and Industries. June 1997. 505 pp.

224. **Jukka Hoffrén,** Finnish Forest Resource Accounting and Ecological Sustainability. June 1997. 132 pp.

225. **Eero Tanskanen,** Suomalaiset ja ympäristö kansainvälisestä näkökulmasta. Elokuu 1997. 153 s.

226. **Jukka Hoffrén,** Talous hyvinvoinnin ja ympäristöhaittojen tuottajana –Suomen ekotehokkuuden mittaaminen. Toukokuu 1999. 154 s.

227. **Sirpa Kolehmainen,** Naisten ja miesten työt. Työmarkkinoiden segregoituminen Suomessa 1970–1990. Lokakuu 1999. 321 s.

228. **Seppo Paananen,** Suomalaisuuden armoilla. Ulkomaalaisten työnhakijoiden luokittelu. Lokakuu 1999. 152 s.

229. **Jukka Hoffrén,** Measuring the Eco-efficiency of the Finnish Economy. October 1999. 80 pp.

230. **Anna-Maija Lehto – Noora Järnefelt** (toim.), Jaksaen ja joustaen. Artikkeleita työolotutkimuksesta. Joulukuu 2000. 264 s.

231. **Kari Djerf,** Properties of some estimators under unit nonresponse. January 2001. 76 pp.

232. **Ismo Teikari,** Poisson mixture sampling in controlling the distribution of response burden in longitudinal and cross section business surveys. March 2001. 120 pp.

233. **Jukka Hoffrén,** Measuring the Eco-efficiency of Welfare Generation in a National Economy. The Case of Finland. November 2001. 199 pp.

234. **Pia Pulkkinen,** ”Vähän enemmän arvoinen” Tutkimus tasa-arvokokemuksista työpaikoilla. Tammikuu 2002. 154 s.

235. **Noora Järnefelt – Anna-Maija Lehto,** Työhulluja vai hulluja töitä? Tutkimus kiirekokemuksista työpaikoilla. Huhtikuu 2002. 130 s.

236. **Markku Heiskanen,** Väkivalta, pelko, turvattomuus. Surveytutkimusten näkökulmia suomalaisten turvallisuuteen. Huhtikuu 2002. 323 s.

237. **Tuula Melkas,** Sosiaalisesta muodosta toiseen. Suomalaisten yksityiselämän sosiaalisuuden tarkastelua vuosilta 1986 ja 1994. Huhtikuu 2003. 195 s.

238. **Rune Höglund – Markus Jäntti – Gunnar Rosenqvist (eds.),** Statistics, econometrics and society: Essays in honour of Leif Nordberg. April 2003. 260 pp.

239. **Johanna Laiho – Tarja Nieminen (toim.),** Terveys 2000 -tutkimus. Aikuisväestön haastatteluaineiston tilastollinen laatu. Otanta-asetelma, tiedonkeruu, vastauskato ja estimointi- ja analyysiasetelma. Maaliskuu 2004. 95 s.

240. **Pauli Ollila,** A Theoretical Overview for Variance Estimation in Sampling Theory with Some New Techniques for Complex Estimators. September 2004. 151 pp.

241. **Minna Piispa.** Väkivalta ja parisuhde. Nuorten naisten kokeman parisuhdeväkivallan määrittely surveytutkimuksessa. Syyskuu 2004. 216 s.

242. **Eugen Koev.** Combining Classification and Hedonic Quality Adjustment in Constructing a House Price Index. (Tulossa).

243. **Henna Isoniemi – Irmeli Penttilä (toim.),** Perheiden muuttuvat elinolot. Artikkeleita lapsiperheiden elämänmuutoksista. Syyskuu 2005. 168 s.

244. **Anna-Maija Lehto – Hanna Sutela – Arto Miettinen (toim.),** Kaikilla mausteilla. Artikkeleita työolotutkimuksesta. Toukokuu 2006. 385 s.

245. **Jukka Jalava – Jari Eloranta – Jari Ojala (toim.)** Muutoksen merkit – Kvantitatiivisia perspektiivejä Suomen taloushistoriaan. Tammikuu 2007. 373 s.

246. **Jari Kauppila.** The Structure and Short-Term Development of Finnish Industries in the 1920s and 1930s. An Input-output Approach. Elokuu 2007. 274 s.

247. **Mikko Myrskylä.** Generalised Regression Estimation for Domain Class Frequencies. Elokuu 2007. 137 s.

248. **Jukka Jalava.** Essays on Finnish Economic Growth and Productivity, 1860–2005. Joulukuu 2007. 154 s.

249. **Yrjö Tala.** Kirkon vai valtion kirjat? Uskontokuntasidonnaisuuden ongelma Suomen väestökirjanpidossa 1839–1904. 317 s.

250. **Hanna-Kaisa Rättö.** Hyvinvointi ja hyvinvoinnin mittaamisen kehittäminen. Huhtikuu 2009. 82 s.

251. **Pertti Koistinen (toim.),** Työn hiipuvat rajat. Tutkielmia palkkatyön, hoivan ja vapaaehtoistyön muuttuvista suhteista. Huhtikuu 2009. 159 s.

252. **Kirsti Ahlqvist.** Kulutus, tieto, hallinta. Kulutuksen tilastoinnin muutokset 1900-luvun Suomessa. Maaliskuu 2010. 314 s.

253. **Jukka Hoffrén (editor) – Eeva-Lotta Apajalahti –Hanna Rättö.** Economy-wide MFA with Hidden Flows for Finland: 1945–2008. 89 pp.

254. **Hannu Pääkkönen.** Perheiden aika ja ajankäyttö. Tutkimuksia kokonaistyöajasta, vapaaehtoistyöstä, lapsista ja kiireestä. Toukokuu 2010. 260 s.

255. **Anna Pärnänen.** Organisaatioiden ikäpolitiikat:strategiat, instituutiot ja moraali. Helmikuu 2011. 291 s.

256. **Ilja Kristian Kavonius.** Kädestä suuhun – Makro- ja mikrotaloudellinen tarkastelu suomalaisten kotitalouksien säästämisestä ja sen mittaamisesta 1950-luvulla. Kesäkuu 2011. 193 s.

257. **Vesa Kuusela.** Paradigms in Statistical Inference for Finite Populations. Up to the 1950s. Elokuu 2011. 236 pp.

The Research Reports series describes Finnish society in the light of up-to-date research results. Scientific studies that are carried out at Statistics Finland or are based on the datasets of Statistics Finland are published in the series.

This study describes the historical development of statistical inference for finite populations starting from the second half of the 18th century up to the beginning of the 1950s when the theory was documented in famous textbooks on survey sampling. The development was interplay between two different tasks: how to draw representative samples from populations, and how to estimate population parameters from the samples. The emergence of statistical thinking in the 19th century was a significant propellant. However, only when digital computers became available for statisticians, sampling techniques obtained their current significance.

9 789522 443151

Tilastokeskus, myyntipalvelu
PL 2 C
00022 TILASTOKESKUS
puh. 09 1734 2011
faksi 09 1734 2500
myynti@tilastokeskus.fi
www.tilastokeskus.fi

Statistikcentralen, försäljning
PB 2 C
FI-00022 STATISTIKCENTRALEN
tfn +358 9 173 420 11
fax +358 9 173 425 00
sales@stat.fi
www.stat.fi

Statistics Finland, Sales Services
PO Box 2 C
FI-00022 Statistics Finland
Tel. +358 9 1734 2011
Fax +358 9 1734 2500
sales@stat.fi
www.stat.fi

ISSN 0355-2071
= Research Reports
ISBN 978-952-244-316-8 (pdf)
ISBN 978-952-244-315-1 (print)