

Statistical inference: Paradigms and controversies in historic perspective

1. Five paradigms

We will cover the following five lines of thought:

1. Early Bayesian inference and its revival

Inverse probability – Non-informative priors – “Objective”

Bayes (1763), Laplace (1774), Jeffreys (1931), Bernardo (1975)

2. Fisherian inference

Evidence oriented – Likelihood – Fisher information - Necessity

Fisher (1921 and later)

3. Neyman- Pearson inference

Action oriented – Frequentist/Sample space – Objective

Neyman (1933, 1937), Pearson (1933), Wald (1939), Lehmann (1950 and later)

4. Neo - Bayesian inference

Coherent decisions - Subjective/personal

De Finetti (1937), Savage (1951), Lindley (1953)

5. Likelihood inference

Evidence based – likelihood profiles – likelihood ratios

Barnard (1949), Birnbaum (1962), Edwards (1972)

Classical inference as it has been practiced since the 1950's is really none of these in its pure form. It is more like a pragmatic mix of 2 and 3, in particular with respect to testing of significance, pretending to be both action and evidence oriented, which is hard to fulfill in a consistent manner. To keep our minds on track we do not single out this as a separate paradigm, but will discuss this at the end.

A main concern through the history of statistical inference has been to establish a sound scientific framework for the analysis of sampled data. Concepts were initially often vague and disputed, but even after their clarification, various schools of thought have at times been in strong opposition to each other. When we try to describe the approaches here, we will use the notions of today.

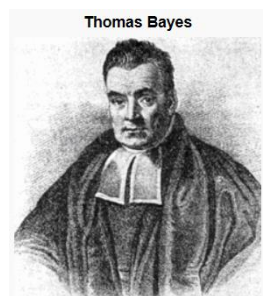
All five paradigms of statistical inference are based on modeling the observed data x given some parameter or “state of the world” θ , which essentially corresponds to stating the conditional distribution $f(x|\theta)$ (or making some assumptions about it). The five paradigms differ in how they make use of the model. For examples illustrating the difference between the five in the case of parameter estimation, see Appendix A. Some other issues that may point forward are discussed in the following appendices.¹

¹ The significant progress in algorithmic modeling the last decades is outside our scope, as these are mainly based on prediction accuracy criteria and pay no attention to inference paradigms (e.g. statistical trees). However, these developments have provided added insights to model selection in general. Key words: accuracy, simplicity, interpretability, variable selection, variable importance, curse of dimensionality, multiplicity of good models.

1.1 The early Bayes paradigm

Bayesian inference refers to the English clergyman Thomas Bayes (1702-1761), who was the first who attempted to use probability calculus as a tool for inductive reasoning, and gave his name to what later became known as Bayes law. However, the one who essentially laid the basis for Bayesian inference was the French mathematician Pierre Simon Laplace (1749-1827). This amounts to having a non-informative prior distribution on the unknown quantity of interest, and with observed data, update the prior by Bayes law, giving a posterior distribution where the expectation (or mode) could be taken as estimate of the unknown quantity. This way of reasoning was frequently called inverse probability, and was picked up by Gauss (1777-1855), who demonstrated that the idea could be used for measurement problems as well. In some sense we may say that objective Bayes ruled the ground up to about 1920, although the notion Bayes was never used. However, some writers were uncomfortable by the use of Bayes law to infer back, as probability is about the future.

Substantial progress in theoretical statistics was made during the 1800's, involving new concepts and modes of analysis, in particular correlation and regression. Key contributors were Francis Galton (1822-1911) and Karl Pearson (1857-1936), the latter responsible for the idea of chi-square tests. Their contributions was in a sense outside the sphere of inverse probability, but was important for the paradigm change soon to come, which had to embrace their developments.

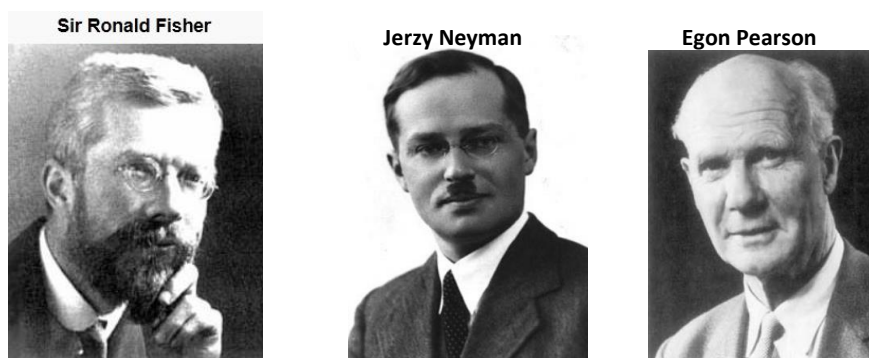


1.2 The Fisherian paradigm

The major paradigm change came with Ronald A. Fisher (1890-1962), probably the most influential statistician of all times, who laid the basis for a quite different type of objective reasoning. About 1922 he had come to the conclusion that inverse probability was not a suitable scientific paradigm. In his words few year later “the theory of inverse probability is founded upon an error, and must be wholly rejected”. Instead he advocated the use of the method of maximum likelihood and demonstrated its usefulness as well as its theoretical properties. One of his supportive arguments was that inverse probability gave different results depending on the choice of parameterization, while the method of maximum likelihood did not, e.g. when estimating an unknown odds instead of the corresponding probability. Fisher clarified the notion of a parameter, and the difference between the true parameter and its estimate, which had often been confusing for earlier writers. He also separated the concepts of probability and likelihood, and introduced a number of new and important concepts, most notably information, sufficiency, consistency and efficiency. However, he did not develop the likelihood basis further into what later became a paradigm of its own, the likelihood paradigm. For testing significance Fisher advocated the P-value as measure of evidence for and against a hypothesis (against when the P-value is small).

Fisher defined statistics as the study of populations, variation and data reduction. His scientific basis was applied problems of biology and agriculture, for which he developed his ideas. He had immediate and lasting success and influenced scientists in many fields. His main book *Statistical Methods for Research Workers* came in 14 editions, the first one in 1925, and the last one in 1970. Fisher himself remained certain that when his methods were so successful in biology they would be so in other sciences as well, and he was vigorously debating against many of his contemporary statisticians, among them Karl Pearson and Harold Jeffreys.

Harold Jeffreys (1891-1989), who mainly worked on problems of geophysics and astronomy, devoted his attention to problems where there were no population, and so felt that the Fisherian ideas did not fit. Moreover, while Fisher held the view that every probability had a “true value” independent of human knowledge, Jeffreys regarded probabilities as representing human information. As a Bayesian Jeffreys was able to give solutions to problems outside the population-sample context, which had no satisfactory solution or not treated at all within the Fisherian framework (among others problems with many nuisance parameters). He was largely ignored by the statistical community, partly due to vigorous opposition by Fisher and a modest Jeffreys. The follow up of Fisher’s works nurtured many statisticians for a long time, even in decades after his death in 1962. However, Jeffreys lived long and eventually got his deserved honor.



1.3 The Neyman-Pearson paradigm

In the late 1920's the duo Jerzy Neyman (1890-1981) and Egon Pearson (1895-1980) arrived on the scene. They are largely responsible for the concepts related to confidence intervals and hypothesis testing. Their ideas had a clear frequentist interpretation, with significance level and confidence level as risk and covering probabilities attached to the method used in repeated application. While Fisher tested a hypothesis with no alternative in mind, Neyman and Pearson pursued the idea of test performance against specific alternatives and the uniform optimality of tests. Moreover, Neyman imagined a wider role for tests as basis for decisions: “Tests are not rules for inductive inference, but rules of behavior”. Fisher strongly opposed these ideas as basis for scientific inference, and the fight between Fisher and Neyman went on for decades. Neyman and Pearson had numerous followers and Fisher seemingly lost the debate (he died in 1962). This bolstered the Neyman-Pearson paradigm as the preferred scientific paradigm, with the now common and well known concepts of estimation and hypothesis testing. However, the P-values are frequently used when reporting, pretending they represent a measure of evidence, in line with Fisher. This symbiosis is what we frequently name classical statistics.

With the optimality agenda the field of theoretical statistics gave room for more mathematics, and this gave name to Mathematical statistics as a separate field of study. According to Neyman (1950):

“Mathematical statistics is a branch of the theory of probability. It deals with problems relating to performance characteristics of rules of inductive behavior based on random experiments”

A solid axiomatic foundation for probability that fitted very well with the emerging scientific paradigm for inferential statistics with sample space as frame, was laid in 1933 by Andrey Kolmogorov (1903-1987). From then on the developments in probability went parallel to developments in statistics and establishing the notion mathematical statistics. A central scholar who unified these lines of thought was the Swede Harald Cramer (1893-1985) with his influential book *Mathematical Methods of Statistics* (1946).

Andrey Kolmogorov



Harald Cramér



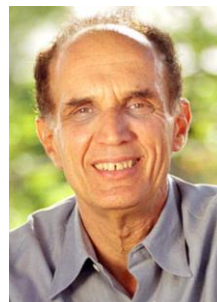
Abraham Wald



The scope of statistics was widened by Abraham Wald (1902-1950), who started to look at the field from a decision oriented perspective, based on loss functions and loss criteria beyond expected squared loss. His followers (like Blackwell and Girsick) developed this further, and provided deep insights to the links between statistics and decisions. However, this may still be regarded to be within or an extension of the Neyman-Pearson paradigm. At this point some defined the field of Statistics itself as the science of Decision making under uncertainty. This was kind of extravagant even at that time, since many aspects related to decision making were outside the focus of statisticians, and more so later, when these aspects came on the agenda of other than statisticians (economists, operations research, psychology etc.).

A central scholar in the development of the Neyman-Pearson theory was Erich Lehmann (1917-2009), who summarized its mathematical content in two influential books, one on estimation and one on hypothesis testing. He was also one of the champions in robust statistics, which was the natural follow-up when parametric statistics (based on normality or other parametric assumptions) had been nearly fully explored. Important contributions to statistical theory, not contrary to the classical tradition, continued to appear in the 1960's and onwards, some of it due to the appearance of the steadily increasing power of computers. The most important general idea was probably “the bootstrap” introduced by Bradley Efron (1938-), which opened up for judging preciseness of estimates also in situations where exact formulas were out of reach.

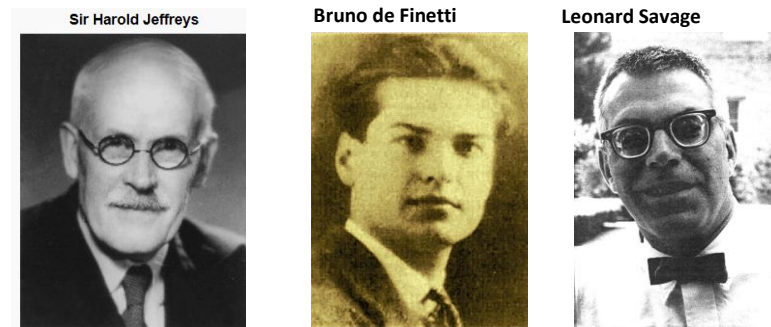
Erich Lehmann



Bradley Efron

1.4 The neo-Bayesian paradigm

In the period 1920-1960 the Bayes position was dormant, except for a few writers, most notably the before-mentioned Harold Jeffreys (1891-1989) on the issue of “objective” non-informative priors, and the Italian actuary Bruno de Finetti (1906-1985) and Leonard Savage (1917-1971) on “subjective” personal probability.



Important for the reawakening of Bayesian ideas was the new axiomatic foundation for probability and utility, established by von Neumann and Morgenstern (1944). This was based on axioms of coherent behavior. When Savage and his followers looked at the implications of coherency for statistics, they demonstrated that a lot of common statistical procedures, although fair within the given paradigm, they were not coherent. On the other hand they demonstrated that a Bayesian paradigm could be based on coherency axioms, which implied the existence of (subjective) probabilities following the common rules of probability calculus. Of interest here is also the contributions of Howard Raiffa (1924-) and Robert Schlaifer (1914-1994) who looked at statistics in the context of business decisions. They found classical statistics inadequate, but found the Bayes paradigm well suited, and could with some right argue that a Bayesian answer was what the decision maker wanted to hear.

The Bayesian paradigm was of course rejected by classical frequentist statisticians as a basis for scientific statistical inference. Serious science could not open up for subjectivity!

1.5 The Likelihood paradigm

This paradigm is the one of the five with the shortest history, and is formulated by statisticians George Barnard (1915-2002), Alan Birnbaum (1823-1976) and Anthony Edwards (1935-). They also found the classical paradigms unsatisfactory, but for different reasons than the Bayesians. Their concern was the role of statistical evidence in scientific inference, and they found the decision oriented classical paradigm sometimes worked against a reasonable scientific process. The Likelihoodists saw it as a necessity to have a paradigm that provides a measure of evidence, regardless of prior probabilities and regardless of available actions and their consequences. Moreover, this measure should reflect the collection of evidence as a cumulative process. They found the classical paradigms (Neyman-Pearson or Fisher) with its significance levels and P-values not able to provide this, and they argued that evidence of observations for a given hypothesis has to be judged against an alternative (and thus Fishers claim that P-values fairly represent a measure of evidence is invalid). They suggested instead to base the scientific reporting on the likelihood-function itself, and use likelihood ratios as measures of evidence. This gives a kind plausibility guarantee, but no direct probability guarantees as with the other paradigms, except that probability considerations can be performed when planning the investigation and its size. This paradigm has its traces back to Fisher,

but he did not follow this path and his use of the likelihood-function is different. A good account of these issues is given by Royall (1997).

1.6 The paradigm disputes

In the history of statistics we have had two major paradigm disputes: The one between Fisher and Neyman, both frequentists, and the one between the frequentists and the Bayesians. We will here briefly look at some of the arguments used in the two debates, and add some of arguments used in the likelihoodist debate. There are also some other major disputes related to paradigms to be discussed in section 2.

A brief note on the logic of inference: the philosophical basis for scientific statistical inference has largely been outside the interests of the general philosophy of science community. Maybe this is due to their adherence to strict non-probabilistic logic and universal validity (and lack of mathematical basis). Notable exceptions are Rudolf Carnap (1881-1970) and Karl Popper (1902-1994) having opposing views. While there are well established principles for deductive logic, this is not so for inductive logic. While Carnap explored various modes of inductive logic, aiming at establishing credibility for hypothesis, Popper held the view that induction is impossible and advocated the method of falsification. Fisher and Jeffreys agreed that statistical inference was an inductive process, going from data to the real world. This was contrary to Jerzy Neyman who held the view that statistical inference is deductive, closer to the view of the philosophers Hume and Popper, that induction is impossible. The phrase “inductive behavior” by Neyman in section 1.3 therefore has to be taken as weaker than inductive in the logical sense. Today Popper is often mentioned as support for the classical research tradition, while Carnap is taken as supportive for Bayesianism and likelihoodism. However, note that statisticians and philosophers may not be in full agreement on the use of the words inductive and deductive, as long as we have strict logic in mind.

1.6.1 Fisher versus Neyman-Pearson

The major difference between the Fisher paradigm and the Neyman-Pearson paradigm is with respect to the testing of hypothesis, i.e the set-up, interpretation and reporting of the result. As we have said above Fisher was more evidence oriented and Neyman-Pearson action oriented. In principle the difference is given in the following table, where the naming significance testing and hypothesis testing is frequently used to distinguish.

Significance testing (Fisher)	Hypothesis testing (Neyman-Pearson)
Purpose: For a single hypothesis H , to measure the evidence against H based on the observation x .	Purpose: To choose one of two specified hypothesis H_1 and H_2 on the basis of an observation x .
Elements: <ol style="list-style-type: none"> 1. One hypothesis H named the null hypothesis. 2. A real-valued function $T(x)$ that gives an ordering of the possible outcomes as evidence against H: $T(x_1) > T(x_2)$ means that x_1 is stronger evidence against H than x_2. 3. The result is a number named P-value or significance level. Interpreted as a measure of evidence against H: The smaller the P-value the stronger the evidence. 	Elements: <ol style="list-style-type: none"> 1. Two hypothesis H_1 and H_2 2. A test function $\delta(x)$ that specifies which hypothesis to choose when x is observed: If $\delta(x)=1$ choose H_1, if $\delta(x)=2$ choose H_2. 3. The result is a decision or action: Choose H_1 or choose H_2, based on action limits determined by predetermined probabilities of false choices. (Note. Here significance level is something else)

Fisher defined P-value as the probability of getting the observed result or a more extreme result when the hypothesis H is correct. The concept was still somewhat fuzzy, since there is no reference to an alternative hypothesis, and what should be regarded as extreme. It was criticized by William Gosset (1876-1937) early on, but is still the common way of scientific reporting to the dismay of many statisticians, Within the action-oriented Neyman-Pearson paradigm of testing statistical hypothesis there is really an obligation to look at both type I and type II risks and the research agenda became to derive the best test for various testing situations under different model assumptions. An important tool in this was the Neyman-Pearson Lemma saying that for testing a simple null-hypothesis against a simple alternative the optimal test is to reject the null-hypothesis when the likelihood ratio LR is beyond a constant c determined by the chosen significance level. By optimality is meant that for a given type I risk α (significance level²) to minimize the type II risk β (or equivalently maximize the power $1-\beta$). For a composite alternative it often turns out that the test so derived is the best test against all alternatives, and thus is a uniformly most powerful test (UMP). Neyman used the words acceptance and rejection of hypothesis, but said that this should not be taken as a statement that the hypothesis was true or false in the strict sense.

Fisher strongly opposed the way Neyman and Pearson transformed “his” significance testing into “acceptance procedures”. He also strongly opposed their interpretation of confidence intervals, and had his own interpretation, named fiducial limits, which may be seen as an effort to establish the idea of inverse probability without a prior at all. Fisher and Neyman debated these issues vigorously in the late 1930’s and into the 1950’s. Some of the disagreement was real, but some may have been just semantic differences. As for fiducial probability Fisher regarded this as his “jewell”, but it did not catch on, and has not become an integral part of Fisherian statistics (by many regarded as his major blunder).

What is currently taught at universities and schools named classical statistics is a curious mix of Fisher and Neyman-Pearson, sometimes named “rejection trials”. The reason may be an effort to cater to both scientific evidence and decisions, without serving any of them very well. It may also be seen as an effort to get closer to the philosophy of science, i.e. the doctrine that hypothesis can only be falsified (Popper).³ It does not matter what is taken as the null hypothesis and the alternative in a full decision framework, but with rejection trials and how it is practiced it clearly matters⁴, and the null hypothesis is the one up for possible falsification.

1.6.2 Classical versus Bayes

Classical statisticians and Bayesians were fighting through the 1950’s and the following decades, but classical statistics remained the common way of scientific analysis and reporting (confidence statements and P-values). It was also unfortunate that untrained people often interpreted frequentist probability statements wrong, and often in a Bayesian way, for instance take a confidence statement as if the parameter is random and not the confidence limits, which is allowed only if you are a Bayesian. Over the years many writers have come up with examples showing rather strange consequences of classical frequentist theory, often dismissed by frequentists as (extreme) cases of little practical interest. On the other hand the Bayesians have been harassed by the so-called *marginalization paradox*, where two seemingly equivalent approaches to a problem with nuisance parameters lead to different answers.

² Note that the word significance level is used different from Fisher, which may cause some confusion.

³ Elementary statistics textbooks may look Fisherian with its P-values, while more advanced textbooks are typically closer to Neyman-Pearson. The reason may be just that the elementary ones have limited scope or stop short of the “hard stuff”.

⁴ The α -risk is conventionally taken as 5% or 1% and the β -risk usually higher say 10% or 20%. You will hardly see these the other way around, say α -risk 20% and β -risk 1%.

Besides being challenged as unscientific, the Bayesian position was set back by its lack of readymade and well developed theory for more complicated situations, and lack of computational methodology to handle any prior, not just convenient conjugate families. However, the Bayesian camp made steady improvement on these issues and a great leap forward came when Markov chain Monte Carlo methods came to help in the 1990's. Now the Bayesians claimed that they covered many situations that the frequentists could not handle.

Frequentists are also blamed for putting their efforts in pre-data considerations (e.g. the sampling distribution of estimators), and in some cases pretend they are post-data, or interpreted that way by non-statisticians. The Bayesian claim that what matters is just the post- data consideration, in which sampling distributions and concepts like unbiasedness play no role. Some find it strange that within classical theory we do not just consider what it actually observed, but also what could have been observed. Others find it a necessity to do so, and then establish whether it is legitimate to condition solely on what is actually observed. Various views on the conditioning issue exist among classical statisticians. A related issue that separates the statisticians into two camps is their adherence or not to the likelihood principle saying that different modes of gathering data should lead to the same conclusion if the likelihood function is the same (up to a multiplicative constant), which resolves the conditioning issue. Bayesian inference fulfills this principle, while Neyman-Pearson type inference not necessarily does so.⁵

The classical statistician has a model, here taken to be a probability distribution $f(x|\theta)$, and as function of θ named likelihood-function . The Bayesian has both a likelihood-function $f(x|\theta)$ and a prior distribution $\pi(\theta)$. When Bayesians are criticized for a subjective $\pi(\theta)$ their reply may be that $f(x|\theta)$ may be just as subjective, and the classical statistician quite often take it for granted (e.g. normality). This raises an Important questions in relation to modeling in general:

- How is the set of models under study established in the first place?
- How are new models, not previously imagined, taken into account?

To address these issues the Bayesians apparently have to depart from the Bayesian updating scheme, and introduce some informal ad hoc procedures, at least in order to bring Bayesians closer to current scientific practice. It seems that Bayesians will have a hard time to live up to coherence, which is their very basis and their claim for being superior to the incoherent frequentists. However, a scientifically sound approach for establishing models is a general issue and a challenge for all breeds of statisticians and applied workers in most sciences. When statisticians could not agree, many sciences had to develop their own practices. The long lasting dispute between frequentists (objectivists) and Bayesians (subjectivists) may have contributed to covering up this issue for a long time.

From the viewpoint of philosophy of science the frequentist and Bayesian approach differ in one important aspect with respect to models: A frequentist approach with emphasis on significance testing gives support for possible model falsification, but not support for model confirmation in the strict sense (deductive reasoning). This is in line with philosopher Karl Popper's doctrine of falsification. In contrast, within a

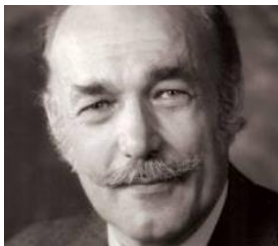
⁵ Here it should be mentioned the notions empirical Bayes estimates and Stein estimates, which had some Bayes features and, to the dismay of some classical statisticians, outperformed some classical methods, even using the classical criteria of judgment.

Bayesian framework with probabilities attached to the truthiness of a model, and this model will never be completely falsified. What we have is a weighted average of the models we imagined at the outset. A frequentist statistician will typically be concerned about model fit, but a Bayesian equipped with posteriors and odds may feel no (or minor needs) to check model fit, which may seem contrary to scientific thinking.

The intense dispute between the frequentists and the Bayesians faded out for various reasons:

- Not much could be added to the old debate and the differing views thrived very well within their own camp (with plenty of work to do for all)
 - The frequentists were all along a majority, leaning on a tremendous track record. Inconsistencies were overlooked, regarded as not relevant or minor and bearable.
 - The Bayesians got renewed spirit in last decades of the century, due to inventive use of high speed computing, which raised their self-esteem
- Many felt that the two existing main approaches were tailored to different contexts and thus kept both in their toolbox.

However, some statisticians kept the debate on the scientific basis alive. Among them was Dennis Lindley (1923-), a very active and outspoken proponent of Bayesianism for more than 50 years. He predicted early on a Bayesian 21th century. At the beginning of the 21th century this was not seen as impossible anymore.



Dennis Lindley

1.6.3 The Likelihood paradigm against the rest

A likelihoodist may start by asking the following three rethoric questions: Is basic science a question of decisions? Isn't it rather to provide evidence? Are the decision rules derived from NP-theory appropriate for interpreting data as evidence? Then argue that the classical paradigm sometimes work against a reasonable scientific process. As example is given the researcher who, based on the data gathered so far, sees an interesting effect that is not statistically significant. She then collects more data and hope to arrive at enough evidence to have documented the effect. A classical statistician may advocate that she now has to disregard the first data in renewed testing, as it will destroy the frequentist guarantees provided by the test. He may also say that the only alternative to this would have been to consult a statistician at the very beginning and embed the research in a sequential data-gathering process which could provide the overall guarantee against false significance.

The Neyman-Pearson way of reporting is by stating conclusions in relation to specific confidence level or α -risk and β -risk, the latter often given implicitly by the choice of number of observations. This is taken as measures of the strength of evidence in the data for the hypothesis supported by the data. Strictly speaking this is false, and the likelihoodists point to examples where the hypothesis accepted (i.e. not rejected) is not the one supported by the data from an evidence point of view. Moreover, they have demonstrated that the required number of observations derived from the common NP-planning formulas is inadequate in the

evidence perspective, and a (too) striking asymmetry in how the error risks are distributed among the competing hypothesis.

1.6.4 Reconciliations: “Objective” Bayes and “Error statistics”

Contrary to the above some statisticians started in the 1970’s efforts to reconcile the differing scientific basis for classical frequentist inference and Bayesian inference. At this time a new breed of Bayesians had entered the scene who advocated non-subjective prior distributions, a line of thought going back to Jeffrey’s non-informative priors in the 1920’s, which had stalled temporarily at the conclusion that non-informative priors do not exist. The project of these Bayesians was to create a framework of pure undisputed scientific inference where updating based on new information could go by Bayes law, starting with a reference or default prior distribution derived from some basic (hopefully agreed) principle, with no intention that this represented prior beliefs. Rather it should be a prior that “lets only the data speak”, and thus provide a comparison with what you get by assuming a prior representing available other knowledge, i.e. a sensitivity analysis. In many cases this yielded posterior distributions with characteristics that are consistent with frequentist ideas and interpretations. A champion in this area from about 1975 and onwards has been José Bernardo. However, this has also sparked disputes among the Bayesians themselves, and it will be interesting to see if these ideas will consolidate to a scientific paradigm of wide support. Worth to mention is also a recent development on the philosophical basis for statistical inference along frequentist lines provided by Deborah Mayo and Aris Spanos named “Error statistics”. This is an effort to unify estimation and hypothesis testing within a common frequentist framework.

1.7 The choice of paradigm

Now feeling that there are a bewildering number of paradigm choices, let us ask the more basic question:
What is the objective of the statistical inference?

In practice we see three broad categories of objectives:

- A. Scientific evidence: What do these data say the public?
- B. Intellectual exploration: What do I believe based on these data?
- C. Decision-making: What should I (or we) do based on these data?

It is possibly too much to ask for a paradigm to embraces all three objectives, even if some proponents may say so. Here are the five paradigms again with an indication of their strongholds:

- 1. The objective Bayes paradigm: A, B
- 2. The Fisherian paradigm: A, B
- 3. The Neyman Pearson paradigm: A, B, C
- 4. The subjective Bayes paradigm: B, C
- 5. The Likelihood paradigm: A, B

Whatever paradigm chosen it is important to realize which of the three objectives you are currently facing. Although sometimes methods are common and computational formulas are the same (e.g. multiple regression), the interpretation and strength of conclusions may differ. This has mainly to do with how the data are gathered and how the analytic process is performed. In practice we often see misuse of methods, due to ignorance, misguiding or deliberate acts. However, the aim of a project may contain elements of all three perspectives, and it may be difficult to draw a sharp line between them. Note that B may also be

parts of science, but in different phases of the scientific process. By C we mean inference for decision support, with no ambitions like A and B; perhaps a method is used because there is no other simple or easily communicated method at hand.

But as we have seen above, all the paradigms are disputed for some reason, sometimes by stretching their frame too far. Then the question is: Do we have to choose paradigm and stick to it?

“There are advantages to the practicing statistician to embrace all.” (George Barnard)

“...we need mental conflict as much as mental coherence to spur us to creative thinking. When different systems give different answers it is a sign that we need to think deeper” (Stephen Senn, 2011).

The likelihood paradigm and the Bayesian paradigm cover A, B and C and can live well together, both satisfying the likelihood principle. It is therefore tempting to say: In cases with emphasis on evidence and no prior information be a likelihoodist, and in cases of decision and prior information, be a Bayesian. However, the likelihood-basis is also disputed, but some of the disagreement may be due to not keeping the concept of evidence apart from belief.

Some final notes: Paradigms of statistics are also linked to different conceptions of probability. Among the many there are two broad categories:

1. Aleatory uncertainty, representing randomness and variability, irreducible
2. Epistemic uncertainty, representing incertitude and ignorance degree of belief, possibly reducible

Bayesians have no reason to distinguish. Frequentists clearly distinguish between the two. The Neyman-Pearson theory is basically founded on aleatory uncertainty alone. Fisher tried to deal with epistemic uncertainty as well, based on his concept of fiducial probability, which he falsely claimed obeyed the common rules of probability. Recently this idea has seen revived and its credibility partly restored based on the concept of confidence distributions, see Schweder & Hjort (2014). They also invites you to take the confidence level as an epistemic probability (degree of belief) without a frequentist interpretation.

Three frequentist statisticians in 1975:



Significant, Not significant*, Significant

*In 2013 on the transfer list to become a Bayesian.

2. Other statistical inference issues

2.1 Extending the applicability

The pioneer developers of statistical inference theory (Fisher, Neyman, Pearson etc.) worked close to the natural sciences. They imagined a context where the researcher had a model based on theory and subject knowledge, and where unknown aspects of the model were sought explored by data. The way of data gathering was important and the pioneers devised modes that were scientifically sound, mainly controlled experiments. This way the analysis of the data could be based on statistical models where the unknown aspects typically were represented by parameters and the inference based on “exact” statistical theory. Economists in the 1930’s were eager to adopt the new statistical inference ideas in order to test their theories, and faced with challenges specific to their field, developed their own field econometrics (e.g Frisch, Haavelmo). However, in many areas there are plenty of data and not so well established theory, and the aim of the investigation is really to establish a relationship and, if lucky and clever, a theory. This is often the case in the social sciences, where the data seldom is obtained under well-defined conditions; they are rather observational data, also referred to as circumstantial data or “grab data”.

In order to serve as scientific evidence, the data should ideally be gathered under strictly controlled conditions. In sciences where controlled experiments, characterized by blocking and randomization, are not possible, one has to assess the following (Wang, 1993):

1. The randomness of the data
2. The stability of the phenomenon under study
3. The homogeneity of the population

In particular this applies to social science, where many do not take these criteria too seriously, and rather use the statistical inference machinery and vocabulary to give their work an undeserved scientific flavor. It should be clear that regression analysis as well as more advanced statistical techniques cannot in itself support causal statements.

During the latter half of the 20th century there was a steady application of standard statistical theory, as well as development of statistical models specific to the area that lead to problems which could be solved by the now well established general statistical principles, e.g. biology, agriculture, engineering, economics, insurance, psychology, educations, social science etc. Many areas offered new problems that inspired new models and new general theory, e.g. actuarial models, time series models and life-event models.

Some important basic inferential issues were still unresolved. While emphasis till now had been on inference for single data sets, there was a rising concern on how to combine data sets, possible from different sources or researchers, to an overall conclusion, possibly agreeable to all parties. Modes of so-called meta-analysis were developed in various sciences (social science, biosciences etc.). More intricate is the issue of combining results from different researchers or research groups that have come to conclusions

bared on different models. In particular this may be of great importance to environmental science, where some models may be disputed.⁶

Causality is another unresolved general research issue. Many sciences have been struggling with this and turned to statistics for help. The message from statisticians has been that without a controlled experiment (characterized by blocking and randomization), which can eliminate confounding factors, we only observe correlation and correlation does not imply causation. Statistical science initiated by R.A. Fisher in the 1920's offers extensive theories for controlled experimentation and analysis. For observational data the statistical profession had next to nothing to offer in terms of causal inference, except possibly in areas like epidemiology. In fact, causal modeling and analysis have been almost a taboo in statistical science. Among pioneer exceptions are Sewall Wright (1920's) in genetics, Haavelmo (1930's) in econometrics, Duncan (1950's), Wold (1960's) and Jöreskog (1970's) in social science.⁷ These contributions were not well known outside its own science. Although they sparked some discussion among statisticians, their general relevance were not understood and remained unexplored. In fact, some hard core statisticians rejected it altogether as unscientific (e.g. Freedman). Ironically methods to analyze causal schemes were unknowingly kept alive in parts of social science, however trying to avoid the label causal analysis.

In the 1990's causality and its relation to statistics finally came into a sharper focus, inspired by new theory and methods of graph modeling. Important contributions to the role of causality and the opportunities and limitations of inferring causality from observational data are given by Judea Pearl (2009), who picked up many of the forgotten ideas of the above scientists and developed them further. Methods for inferring causation from data in this context are now at hand, but have so far not had the impact on statistical practice hoped for. It will be interesting to see if it will be part of mainstream statistics in the future.

Philosophy of science has through the centuries also struggled with ideas of causality. The above deep-diving into causality is therefore a valuable contribution to the general philosophy of science as well.

2.2 Some specific controversies

We will now briefly look into three issues that have raised vigorous disputes at various points in the history of inferential statistics and are still debated among statisticians, users of statistics and their professional organizations. They relate to all sciences, but certain aspects more to social science than the natural sciences. They are

1. The significance test controversy:
 - a statistical significant result is not necessarily practically significant
2. Model snooping and inference bias
3. Inferring causal relationships

⁶ Means to solve this by Bayesian principles have been proposed in the mid 1990's by Adrian Raftery, but shown to have some defects (related to the so-called Borel paradox). The Norwegians Tore Schweder and Nils Lid Hjort have proposed a likelihood-based alternative without this defect.

⁷ We also mention Clive Granger who defined his causality concept for time series in the late 1960's, followed up by Schweder (1970)..

2.2.1 The significance testing controversy

Hypothesis testing is mainly associated with the classical paradigm of statistical inference, and offers the Bayesians plenty of opportunity to criticize. Although they have ways to handle this set up, they may prefer to phrase the problem differently. However, classical statisticians have over the years had heated discussions among themselves on the adequacy of the hypothesis testing framework, with Yates (1951) as an early critic. Arguments may range from the practical: *Its logic is difficult to grasp for lay practitioners leading to misconceptions and misuse*, to the conceptual: *It is not an adequate framework for scientific discovery and reporting*.

Hypothesis testing is often misused or misinterpreted in practice. Issue: Waiting for significant results, and just publish those (p-values) and forget all non-significant results. The frequentist guarantees will then be illusory. This temptation to misuse is often taken as an objection to classical hypothesis testing. Reporting confidence intervals instead is often recommended. The Bayesians don't face the same trap.

A statistically significant difference is often falsely taken to be a difference of practical significance, i.e. a difference that matters. As example take two treatments, one standard and a new cheaper one. Statistically significant difference here only means that an observed difference is not likely to be just due to chance in a situation where there is no real difference in the expectation sense. Practical significance or not, on the other hand, has to be decided by the user of the result, which is not necessarily the observer or the statistician.

A common practice is to judge significance by P-values in relation to a chosen significance level (often routinely 5%). This way one relates to the risk of falsely claiming significant difference where there is no real difference (type I error). The risk of not claiming a difference when there is one is not addressed (type II error). Both risks should be of concern! It should be of concern (to someone) before gathering the data, or before wasting time on data which is too scarce to make reasonable inferences. Since the alternative to no difference typically is a range of possibilities the planner/observer/user will be forced to think about what is a difference that matters, which (s)he wants to discover with acceptable probability. The test may then be a balance between the two risks, say type I risk=5%, type II risk=10%. This determines jointly the necessary amount of observations and the critical value of the test statistic. In the case that we have to relate to scientific reporting we should also be aware of the role of sample size in the statements of significance or not and the consequence of blind adherence to common significance levels. In some cases it may be reasonable to increase the type I risk in order to reduce the type II risk. We cannot get it both ways! In some cases we realize that the risks are unacceptable and that the report is a waste and should not have been published.

In connection with common hypothesis testing we should always have in mind that

- With few observations the no difference hypothesis is favored, with a large type II risk as a consequence.
- With sufficiently many observations any tiny difference will be detected and statistically significant (with high probability).

In the latter case it is of course unfortunate if the reader interprets the report as if a practically significant difference has been found.

The sample size issue above may also be present when testing for model fit, e.g. by routine and unconscious use of the common chi-square test (formula). If we scale up the number of observations, say 100 times and imagine the same distribution of counts in the groups the test statistic Q will be scaled up by 100 as well, so that we may risk to reject any reasonable model. In this connection we point to a statement by G.E.P. Box: “No model is (exactly) true, but some models are useful!”. If a model is rejected by a goodness of fit based on massive data, it may still be useful, and to decide this we really have to check closer whether the difference matter for the problem at hand.

The following advice is often given: “Don’t look at the data before you test!” This is to avoid so called data snooping, where one search for patterns and then select something to test. On the other hand, it is argued that (Wang, 1993), in relation to the above, that one is better off by a two-step procedure:

- (1) Look at the data, and see if the difference is practically significant? If not, don’t bother about next step.
- (2) Is the difference statistically significant?

This advice is given for the context of a serious investigation based on models established prior to the data analysis, in which specific aspects are sought clarified by the data. We return to the data snooping problem later.

Although classical statistical theory often tries to cast hypothesis testing in a decision making context with its type I and type II error (and α and β risks), many argue that it is of little practical use. A major difficulty for many lay users of statistical theory is what to state as the null hypothesis and what to state as the alternative. Examples may be given where this is not so clear cut, and where reversing the roles may apparently give different answers, unless the whole apparatus of type I and type II errors are addressed. With these inherent difficulties in hypothesis testing, it is argued that it would be better to go for confidence intervals with a more modest ambition as decision support.

In the late 1900’s the role of significance testing also became a concern for many professional scientific organizations, e.g. education, psychology, medicine, and changes in publication practice in journals were asked for. Some argue that (objective) Bayesian reporting may resolve this issue.

2.2.2 Model snooping and inference bias

Science distinguishes between explorative and confirmative analysis. An explorative analysis of data may be used to suggest a possible model or a theory, and a confirmative analysis is about testing a proposed hypothesis or theory on new data. A basic principle in science is that we should never test a hypothesis by the same data that was used to establish in the hypothesis. However, quite often the researcher looks at data during the research process to get ideas on how to proceed. Then hypothesis are formed, and certain facets not consistent with this are discarded towards the final reported hypothesis testing, which are presented as confirmatory but really is just exploratory. Most data sets have some peculiar patterns which are entirely due to chance among the many other patterns that could also strike the eye. If we pick such a pattern and make a general hypothesis and then test it on the same data, the hypothesis will be self-confirming. In a way this is circular reasoning. If we fail to follow the no reuse principle we risk to present pure nonsense as scientific facts.

This happens all the time in scientific reporting, perhaps more so in the social sciences than elsewhere. In some cases it is not transparent that something like the above has actually happened. When a regression is

run with no prior theory and general conclusions are drawn based on significant variables this is so, unless we obtain new data for testing. Moreover, having plenty of explanatory variables the temptation is there to run many regressions and use variable significance or explanatory power in some sense as screening device. This is typically done in a stepwise manner: Forward inclusion, backward exclusion or full stepwise. This way many tests are performed and we are not sure to end up with something meaningful or anything of predictive value. In fact, with many variables we have the risk that even pure noise may turn out significant. Having such variables in a model (except in the error term) is more than useless. This horror has been clearly demonstrated by Freedman (1983). He established 51 data columns with 100 observations each, generated as Gaussian pure noise, i.e. independent, identically distributed standard normal variables. He then screened and refitted a regression and arrived at the following disturbing result: R-square=36%, P-value=0.5 and 6 coefficients significant at the 5% level. Faced with this result we have seemingly arrived at a relationship we can safely publish as research. However, we have explained pure noise by pure noise and therefore only contributed to pure nonsense.

Another backside of uncritical significance testing and reporting is the so-called “publication bias”, where only investigations with significant results are published. As example take an experiment where a new treatment is compared with the common one. Imagine that the experiment is performed by 20 researchers in each of 20 different countries, and 19 of them observed no difference and one observed a statistically significant improvement at the 5% level. Suppose just that one (the Norwegian) bothered to report the result or got the report published. We are then left with the impression that the new treatment was more effective, despite the fact that 1 out of 20 is exactly the expected number of false rejections of no difference when 5% significance level is used. Then imagine that all 20 had their results published. It would then be hilarious to say that the new treatment works better on Norwegians. Of course results like this have to be supported by more extensive repeated confirmatory experiments.

Publication bias is a reality which is extensively debated by scientists and professional organizations.

2.2.3 Inferring causal relationship

Many scientific investigations aim at clarifying causal relationships. Without special arrangement we can only observe correlation and as most students of statistics have been told “correlation does not imply causation”. A major contribution to science came with R.A. Fisher who laid the foundation for sound experimental practice, and contributed to the accompanying statistical theory. This included experimental designs with blocking and randomization as key tools in order to eliminate confounding factors. This became common practice in biology, medicine and pharmacy. However, in many sciences we cannot experiment and are left with data generated under less well defined conditions that defy inference guarantees of some kind.

A special event in the history of inferential statistics was the “Smoking and lung cancer controversy”. When statistics on a possible link started to appear in the 1950’s this was dismissed as evidence by prominent statisticians, most notably R. A. Fisher. Their argument was that the data was not gathered under controlled conditions, and possible confounding factors was not eliminated (e.g. genetics that encouraged both smoking and lung cancer). This was to the dismay of all who wanted to save new generations from a harmful habit. The controversy delayed the final conclusion by more than a decade, but sharpened the awareness of a fundamental scientific issue to the benefit of science itself, and possibly prevented a number of quick and unfounded conclusions elsewhere.

On the other hand, this contributed to the unfortunate situation that causality was a near taboo outside the context of well-designed experiments, and thus became outside the agenda of the mainstream

statistical science itself. Scattered efforts were made in some areas to develop methods for analyzing causal schemes, for instance path analysis in genetics in the 1920's (Sewall Wright). The general idea was to interpret an observed covariance structure in terms of an assumed causal scheme. This was picked up by in social science in the mid 1960's and developed further. The possibility to include unobserved (latent) variables in observational studies was found very attractive, and covariance structure analysis of some kind rapidly became a standard analysis tool in social science (e.g. Jöreskog, LISREL). However, since a given covariance structure may be consistent with very different causal schemes, it was emphasized that the analysis could possibly falsify a theory, but could never confirm a theory in the strict sense. Nevertheless, this kind of causal inference on observational data was intensely criticized by parts of the statistical profession. This is reflected by the statement: "No causation without manipulation" (D.R. Rubin). An outspoken critic was David Freedman, who explicitly expressed his critical view on the statistical practice in many areas in a series of papers in the 1980's. A compromise position is that a causal statement in an observational study can be justified if the research follows the refutation process as expressed by Wang (1993):

"A causal statement is formulated to describe a general characteristic that is consistent with existing knowledge and is deemed useful for other applications. Still, the causal statement is provisional and is subject to further refutation. If other research finding conflict with the statement, then the statement has to be reexamined or discarded. Otherwise the statement will be regarded as valid and useful".

A new twist came in the late 1980's after graph modeling had become research area of itself in statistics. This also fitted well with Bayesian ideas, and led to the development of Bayesian Nets, which rapidly became a popular technique, mainly marketed by the emerging data-mining community. This inspired researchers with varied background to explore concepts of causality more deeply, and to pick up the lost threads of early writers, like Wright, Haavelmo, Duncan, Wold etc. In section 2.1 we have mentioned the major contribution of Judea Pearl in, having a prescription for causal inference that pays attention to Rubin's demand above. Of particular interest to social science and economics is the concept of structural equation models (SEM), originating in econometrics. They were originally (by Haavelmo, Marchak, Wold etc.) intended to be carriers of causal information, but by later econometricians mostly just taken to be carriers of probabilistic and thus covariate information (in line with the causal taboos).

Currently we see a restoration of the original intention (Pearl, 2009): "The structural equations are meant to define a state of equilibrium, and not strict mathematical equations. The equilibrium may be violated by outside interventions, and the equations represent an encoding of its effects and the mechanism of restoring equilibrium". According to him:

1. Causation is the summary of behavior under interventions.
2. The language of causal thought is the combination of equations and graphs.
3. Causation is revealed by "surgery" on the equations (guided by the graph) and predict its consequences.

For more on this see Appendix B.

Alternative and partly opposing schemes for causal modeling and interpretation exist, among them by Jim Heckman (2000).

A last question:

Is a situation with no fully agreed scientific basis for statistical inference bad or good?

Write your opinion here:

Appendix A: Examples of estimation in the five main paradigms

We will illustrate the five main inference paradigms by the problem of estimating a single unknown parameter θ . Our context is: Observe a random variable X (possibly vector) with distribution given by the density $f(x | \theta)$ for given θ in the continuous case, or point probability $p(x | \theta)$ in the discrete case. As functions of θ for given x they are named likelihoods. In the Bayesian case let $\pi(\theta)$ be the prior distribution of θ and let $\pi(\theta|x)$ be the posterior of θ given $X=x$, obtained by Bayes law.

We will consider two situations, one with discrete and one with continuous X :

- a. X Binomial(n, θ) $p(x | \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$
- b. X Normal(θ, σ^2) σ^2 known $f(x | \theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\theta)^2}$

The second situation will be extended to n independent, identically distributed observed X 's.

For each we want the opportunity to judge the preciseness of the estimate, typically by some kind of probability statement. In classical (and Fisherian) statistics parameters are non-random, while random in Bayesian statistics, and no need to differentiate it from other random variables. The kind of probability statement given will therefore be quite different. In Bayesian statistics it will be a direct probability statement about the quantity in question, while in classical statistics it will typically be given as a confidence interval with an attached confidence level guarantee, which may be interpreted in a frequency context (covering probability in repeated applications).

1. Early Bayesian inference

- Start with an non-informative prior distribution for the parameter θ (as if it was random).
- Then use data $X=x$ to compute the posterior distribution of θ .
- The (Bayes) estimate of θ is then given by the expected value of the posterior, i.e. $E(\theta|x)$.
- The choice of an non-informative prior prevents prejudice and is "scientifically objective".

Example 1a.

Prior of θ : Uniform[0,1]. Observed $X=x$.

Posterior of θ : Beta($x+1, n-x+1$).

Expectation (Bayes estimate) = $(x+1)/(n+2)$ (The Laplace rule)

Example 1b.

Prior of θ : Uniform $[-\infty, +\infty]$. (Improper prior!)

Observed $X=x$.

Posterior of θ : Normal(x, σ^2).

Expectation (Bayes estimate) = x

Note. In case of “flat” prior on $[-\infty, +\infty]$ $f(\theta) = 1$ and so $f(\theta | x) \propto f(x | \theta)$ i.e. the posterior is directly proportional to the likelihood (see next).

2. Fisherian inference

Consider the likelihood function

$$l(\theta|x) = f(x | \theta) \quad \text{as function of } \theta \text{ for given } x.$$

The maximum likelihood estimate of θ is then obtained by maximizing $l(\theta|x)$ wrt θ , i.e.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} l(\theta|x)$$

Under some regularity assumptions $\hat{\theta}$ is approximate $Normal(\theta, \frac{1}{I(x)})$ where

$$I(x) = - \frac{\partial^2}{\partial \theta^2} \log(l(\theta|x))$$

is an empirical version of the so-called Fisher-information. The approximate normality may be used to make fiducial limits for θ :

$$\hat{\theta} \pm 1.96 \cdot \frac{1}{\sqrt{I(x)}}$$

Note. In defining the likelihood function we may leave out any multiplicative constant not dependent on θ .

Example 2a.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} (\theta^x (1-\theta)^{(n-x)}) = x/n \qquad I(x) = \frac{n}{\frac{x}{n}(1-\frac{x}{n})}$$

$$\frac{x}{n} \pm 1.96 \cdot \sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}}$$

Example 2b.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} (e^{-\frac{1}{2\sigma^2}(x-\theta)^2}) = x \qquad I(x) = \frac{1}{\sigma^2}$$

$$x \pm 1.96 \cdot \sigma$$

In case of n independent, identically distributed observed X 's:

$$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

3. Neyman- Pearson inference (Classical inference)

Estimate may be derived from various principles (among them maximum likelihood).

Derived estimates are then judged by performance criteria:

- Unbiased, minimum variance, invariance
- Consistent, asymptotic optimal, robust etc
- Minimal loss (decision oriented)

The reporting may be as a point estimate or as a confidence interval.

Example 3a.

$$\hat{\theta} = \frac{X}{n} \quad E(\hat{\theta}) = \theta \quad \text{var}(\hat{\theta}) = \frac{\theta(1-\theta)}{n} \rightarrow 0$$

$$\tilde{\theta} = \frac{X+1}{n+2} \quad E(\tilde{\theta}) = \frac{n\theta+1}{n+2} \rightarrow \theta \quad \text{var}(\tilde{\theta}) = \frac{n\theta(1-\theta)}{(n+2)^2} \rightarrow 0$$

$$E(\tilde{\theta} - \theta)^2 = \frac{n\theta(1-\theta)+(1-2\theta)^2}{(n+2)^2} \rightarrow 0 \quad (\text{consistent})$$

Note. In general we have $E(\hat{\theta} - \theta)^2 = \text{var}(\hat{\theta}) + (E\hat{\theta} - \theta)^2$, so in principle for expected squared loss allowing for some bias may be compensated by smaller variance.

Confidence interval (with interpretation different from fiducial limits)

$$\frac{x}{n} \pm 1.96 \cdot \sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}}$$

Example 3b.

$$\hat{\theta} = X \quad E(\hat{\theta}) = \theta \quad \text{var}(\hat{\theta}) = \sigma^2$$

Example 3b'. X_1, X_2, \dots, X_n i.i.d. $\text{Normal}(\theta, \sigma^2)$ σ^2 known. \bar{X} sufficient for θ .

$$\hat{\theta} = \bar{X} \quad E(\hat{\theta}) = \theta \quad \text{var}(\hat{\theta}) = \sigma^2/n \rightarrow 0$$

\bar{X} is UMVU (Uniformly Minimum Variance Unbiased).

\bar{X} is BLUE (Best Linear Unbiased) for any distribution, Normal or not.

$$\bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

4. Neo - Bayesian inference

As with section 1 except, if warranted use an informative prior distribution reflecting the current knowledge of the parameter θ (for convenience stay within conjugate family of priors)

- Then use data $X=x$ to compute the posterior distribution of θ .
- The (Bayes) estimate of θ is then given by the expected value of the posterior, i.e. $E(\theta|x)$.

Example 4a.

Conjugate prior of θ : Beta[r,s]. Observed $X=x$.

Posterior of θ : Beta(r+x, s+n-x).

Expectation (Bayes estimate) = $(x+r)/(n+r+s)$

Note the cases $r=s=1$ (uniform prior) and $r=s=0$ (improper prior).

Example 4b.

Conjugate prior of θ : Normal(m_0, σ_0^2)

Observed $X=x$.

Posterior of θ : Normal($\frac{\frac{m_0}{\sigma_0^2} + \frac{x}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}$)

Expectation (Bayes estimate) = $\frac{\frac{m_0}{\sigma_0^2} + \frac{x}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}$

Note the case $\sigma_0^2 \rightarrow \infty$ corresponding to "flat" improper prior.

Example 4b'.

Conjugate prior of θ : Normal(m_0, σ_0^2)

Observed $x=(x_1, x_2, \dots, x_n)$. Computed \bar{x} (sufficient statistic).

Posterior of θ : Normal($\frac{\frac{m_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$)

Expectation (Bayes estimate) = $\frac{\frac{m_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$

Note 1. The case $\sigma_0^2 \rightarrow \infty$ corresponds to "flat" improper prior.

Note 2. The estimate is a weighted sum of m_0 and \bar{x} with weights dependent on σ_0, σ and n .

Note 3. The mean is a sufficient statistic, and thus the posterior given all observations and given the mean are the same.

Example 4c. $X \text{ Normal}(\mu, \sigma^2)$ σ^2 unknown. Parameter vector $\theta = (\mu, \sigma^2)$

Look up in the literature the cases: Improper prior on σ and conjugate prior on σ .

“Objective” Bayesian inference (Reference analysis)

Reference analysis provides a Bayesian solution that depends only on the model assumptions and the observed data. It applies a prior that reflects lack of knowledge in the sense that it is the worst possible with respect to the information gap between the prior and the posterior obtained by Bayes law. As a measure of lack of information is suggested the expected (Shannon) information $I(\pi, f) = EI(X; \pi, f) = \int I(x; \pi, f) f(x) dx$ where

$$I(x; \pi, f) = \int \pi(\theta|x) \log\left(\frac{\pi(\theta|x)}{\pi(\theta)}\right) d\theta$$

and $f(x)$ and $\pi(\theta|x)$ are implicitly given by the common conditioning rules (the integrals wrt x has to be replaced by sums in the case of discrete observations) . This gives rise to a variational problem to be solved for $\pi(\theta)$. In the one-parameter case (with a regularity condition) this leads to the simple solution (which corresponds to Jeffrey’s non-informative prior)

$$\pi(\theta) = I(\theta)^{1/2}$$

where $I(\theta)$ is the expected Fisher information of the model itself, i.e.

$$I(\theta) = - \int f(x|\theta) \frac{\partial^2}{\partial \theta^2} \log(f(x|\theta)) dx$$

Example 4d

$I(\theta) = \theta^{-1}(1 - \theta)^{-1}$ which gives $\pi(\theta) = \theta^{-1/2}(1 - \theta)^{-1/2}$ so that the Bayes estimate becomes $(x+1/2)/(n+1)$.

5. The likelihood paradigm

The likelihood paradigm is essentially, for given observed x , to examine the likelihood function $l(\theta|x)$ as function θ . The relative plausibility of θ_1 versus θ_2 is judged by the likelihood ratio $l(\theta_1|x)/l(\theta_2|x)$. The plausibility of an estimate $\hat{\theta}$ can be judged as function of the unknown θ by

$$R(\theta|x) = l(\theta|x) / l(\hat{\theta}|x)$$

Usually the estimate is taken to be the maximum likelihood estimate. In that case this function has one as its maximum.

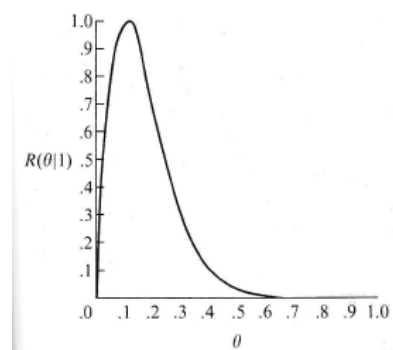
Example 5a $x \sim \text{bin}(n, \theta)$ n given

$$R(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}}{\hat{\theta}^x(1-\hat{\theta})^{n-x}} = \left(\frac{\theta}{\hat{\theta}}\right)^x \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n-x}$$

Note that the multiplicative factor common to the numerator and the denominator cancels out. If x is given and we observe n , the common factors are different, but cancels out giving the same R -function.

As an example take $n=10$ and $x=1$, so that the maximum likelihood estimate is $\hat{\theta}=0.1$. The graph of the relative likelihood as function of θ is then

We see the region of plausible (not unreasonable) values of θ ranges up to about 0.5.



Example 5b x_1, x_2, \dots, x_n independent $\text{Normal}(\theta, \sigma^2)$ σ^2 known

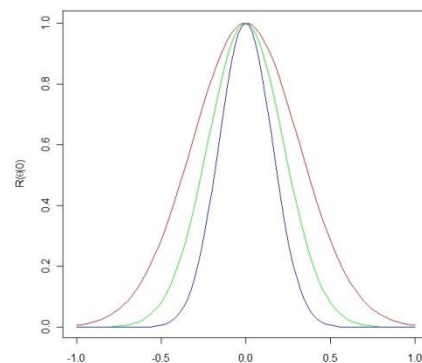
In this case (check!): $R(\theta|x) = \exp\left(-\frac{n}{2\sigma^2}(\theta - \bar{x})^2\right)$

Suppose $\sigma^2=1$ and $\bar{x} = 0.0$.
is a graph for the cases $n=10, 20, 40$.

We see that plausible values of θ are
 $n=10$ between -1.0 and $+1.0$
 $n=40$ between -0.5 and $+0.5$

Compare this with common standard error considerations.

Warning: The vertical axis in these graphs does not represent probabilities.



Here

For
For

Appendix B Evidence based (likelihood) testing of hypothesis

This brief account should be contrasted with classical hypothesis testing found in most introductory statistics textbooks and to Bayesian testing found in Bayesian texts.

Let x_1, x_2, \dots, x_n be independent identically distributed with density $f(x)$. We will test

$$H_1: f(x)=f_1(x) \text{ against } H_2: f(x)=f_2(x)$$

Let the likelihood-function in each of the two cases be $L_i = \prod_{j=1}^n f_i(x_j)$ for $i=1,2$.

For appropriately chosen k we will say that data is strong evidence

$$\text{for } H_1 \text{ if } \frac{L_2}{L_1} \geq k \quad \text{for } H_2 \text{ if } \frac{L_1}{L_2} \geq k \text{ i.e. if } \frac{L_2}{L_1} \leq \frac{1}{k}$$

and too weak evidence to decide if $\frac{1}{k} \leq \frac{L_2}{L_1} \leq k$.

The probability of wrongly claiming evidence for H_2 when H_1 is true is bounded (easily proved)

$$P_{H_1}\left(\frac{L_2}{L_1} \geq k\right) \leq \frac{1}{k}$$

with the same upper bound for wrongly claiming evidence for H_1 when H_2 is true.

Suggested choices of k are: $k=8$ for Strong evidence, $k=32$ for Quite strong evidence, matching approximately the choices of significance testing of 5% and 1% respectively.

If we have some prior information on the position of the two hypotheses, we may choose different k 's for the two possible claims of strong evidence.

Example $f(x)$ Normal(θ, σ^2) σ known. $H_1: \theta=\theta_1$ against $H_2: \theta=\theta_2$

We now have $\frac{L_2}{L_1} = \exp\left\{\left[\bar{x} - \left(\theta_1 + \frac{\delta}{2}\right)\right] \frac{n\delta}{\sigma^2}\right\}$ where $\delta = \theta_2 - \theta_1$

Strong evidence in favor of H_2 occurs when $\frac{L_2}{L_1} \geq k$ which is equivalent to

$$\bar{x} \geq \theta_1 + \frac{\delta}{2} + \frac{\sigma^2 \log(k)}{n\delta} = \bar{x}_u$$

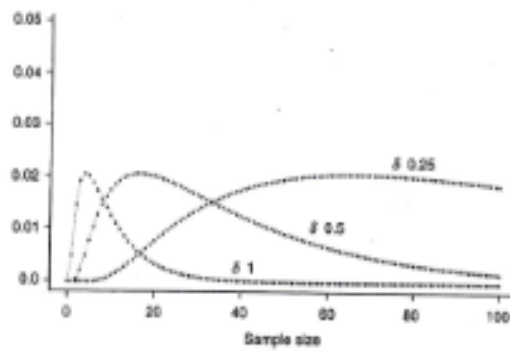
The probability that this happens when H_1 is true

$$P_{H_1}(\bar{x} \geq \bar{x}_u) = 1 - G\left(\frac{\sqrt{n}\delta}{2\sigma} + \frac{\sigma \log(k)}{\delta\sqrt{n}}\right)$$

where G is the cumulative standard normal distribution. We see that this expression is small for both small n and large n , and obviously has a maximum found by maximizing the parenthesis with respect to n . This maximum $1 - G(\sqrt{2 \log(k)})$ occurs for $n = 2 \log(k) \left(\frac{\sigma}{\delta}\right)^2$. For $k=8$ and $k=32$ the maximums are 0.021 and 0,004 respectively, much lower than the general upper bound $1/k$. The intuition here is of course that for n

small there is not enough evidence to support any of the hypothesis, and for n large it is no support for the wrong hypothesis.

A graph of this probability as function of the number of observations for various δ (scaled by σ) follows for the case $k=8$ (Royall, 1997):



Appendix C: Structural models and causality

In a system of structural equations each equation of type $y = f(x)$ really is an assumption $f(x) \rightarrow y$ and the system may be represented as a directed graph with variables as nodes.

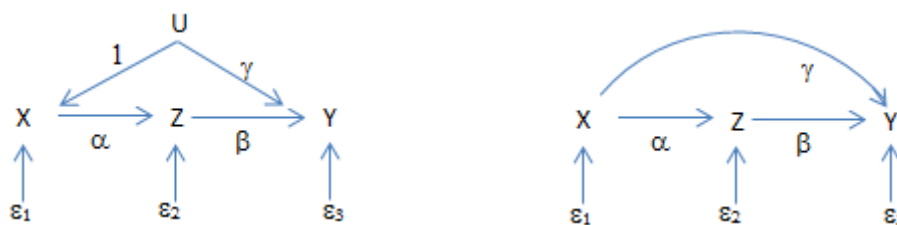
Pearl (2009) has provided a deeper understanding of such modeling, among others the following:

- (i) The necessity to differentiate between observed covariation and changes made by intervention, i.e. for probabilities the difference between $P(A | \text{see } B)$ and $P(A | \text{do } B)$. The first can be handled by ordinary probability calculus: the latter needs some more, in particular how to derive “do-probabilities” from “see-probabilities”.
- (ii) The opportunities and limitations when identifying and estimating the unknown model parameters from observed data, the aim being more on judging the effect of changes in variables by interventions on other variables rather than estimates of the individual parameters and global fit.
- (iii) The models will then help answer questions like; What else have to be observed? What extra assumptions have to be made?
- (iv) Since SEM’s are based on causal assumptions they cannot be tested by statistical tests. Having two competing models all you can say is that: “If you accept this, then...., but if you accept that, then....”. However, the consequence of having different models with the same covariance structure (which cannot be separated statistically) can now be explored.
- (v) The link between SEM and Bayesian Net modeling is clearly demonstrated, and their advantages and disadvantages of each are discussed.

We will illustrate by two simple models which explains an observable variable y in terms of two observable variables z and x . Here the ϵ_i ’s are unobservable error terms assumed to be independent with expectation zero. The variable u is also unobservable with expectation zero and independent of the ϵ_i ’s. Each model is represented in a directed graph.

Model M_1 :
 $x = u + \epsilon_1$
 $z = \alpha x + \epsilon_2$
 $y = \beta z + \gamma u + \epsilon_3$

Model M_2 :
 $x = \epsilon_1$
 $z = \alpha x + \epsilon_2$
 $y = \beta z + \gamma x + \epsilon_3$



The two models have the same covariance structure, giving rise to identical probabilistic prediction of y in terms of z and x , but differ slightly in their data generating mechanism. In M_2 is y (except error) determined by the observables directly by z and x , but also indirectly by x via z , while in M_1 is y (except error) determined by the observable z and the unobservable u , which is observed with error by x . We now have

$$M_1: E(Y | do X=x) = E(\beta(\alpha x + \varepsilon_2) + \gamma u + \varepsilon_3) = \beta \alpha x$$

$$M_2: E(Y | do X=x) = E(\beta(\alpha x + \varepsilon_2) + \gamma x + \varepsilon_3) = (\beta \alpha + \gamma)x \quad (= E(Y | see X=x) \text{ for both } M_1 \text{ and } M_2)$$

Note that for M_1 we have to “wipe out” the first equation (without replacing u by $x - \varepsilon_1$ in the third).

The parameters β , α and γ may be determined from observations of Y , Z and X by running the two regressions of M_2 . However, to do the prediction of Y for given (do or see) x in model M_2 it suffices to pick the estimate of the regression coefficient $\rho = \beta \alpha + \gamma$ of the regression of Y wrt X . If we want the do-prediction for model M_1 this has to be corrected by the formula $\beta \alpha = \rho - \gamma$, where the estimate of γ is obtained by the regression coefficient of X in the full regression of Y wrt Z and X .

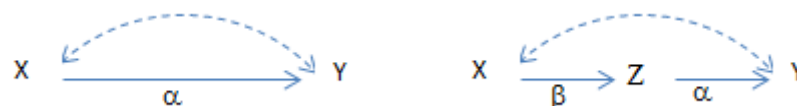
Equations are named structural if they can be given an interpretation as follows (Pearl, 2009):

For the equation $y = \beta x + \varepsilon$ imagine an experiment where we control X to x and any other set of variables Z to z , and the value y of Y is given by $\beta x + \varepsilon$, where ε is not a function of x and z .

The meaning of β in this expression is $\beta = \frac{\partial}{\partial x} E(Y | do X = x)$

The total effect of X on Y is given by the distribution of Y when X is held constant at x and all other variables vary naturally. The direct effect of X on Y is given by the distribution of Y when X is held constant at x and all the other observable variables are held constant as well.

In the above examples there were no two-way arrows indicating feedback between X and Y . This may occur in practice, as illustrated in two simple graphs. In the left graph X has a direct causal link to Y , but there is also a two-way link between them, which represents the possibility of unobserved (and unknown) variables that affect both X and Y , i.e. what is usually named confounding. In the right graph we see the same, but here we have a third observable variable Z , which intervenes the direct causal link between X and Y .



In the first case it is not possible to identify the direct effect α , but in the second case both direct effects turn out to be identifiable. Assuming linear models for both causal links, the β is estimated by the regression coefficient in the regression Z wrt X , and α is estimated by the regression coefficient of Z in the regression of Y wrt Z and X .

Modern causal theory provides the necessary and sufficient conditions for parameter identifiability, in general and for models with specific structure. These conditions are translated to simple graphical rules, which also help to clarify what more is needed to be observed to go from non-identifiability to identifiability.

It makes a difference how the equality sign “=” is interpreted in expressions like $y = a + bx + e$. Is it just equating two algebraic terms, a way of describing the conditional distribution of y given x , void of any causal content, or is “=” interpreted as “determined by” or caused by?

Literature

Berger, J.O. (1980) *Statistical Decision Theory and Bayesian Analysis*. Springer.

Bernardo, J.M. (1979) Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society B*, 41, 113-147. (with discussion).

Bernardo, J.M. (1997) Noninformative priors does not exist. *Journal of Statistical Planning and Inference* 65, 159-189 (with discussion).

Birnbaum, A. (1962) On the foundations of statistical inference. *Journal of the American Statistical Association*, 57, 296-306. (See also [link](#))

Birnbaum, A. (1972) More on concepts of statistical evidence. *Journal of the American Statistical Association*, 67, 858-861.

Breiman, L. (2001) Statistical modeling: The two cultures. *Statistical Science*, vol 18, No 3, 199-231 (with discussion).

Cox, D.R. (1958) Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29, 351-372.

Cox, D.R. (1988) Some aspects of conditional and asymptotic inference. *Sankhya*, A50, 314-330.

Cox, D.D. (2006) *Principles of Statistical Inference*. Cambridge University Press.

Cramer, H. (1946) *Mathematical Methods of Statistics*. Princeton University Press (republished 1999)

Efron, B. (1978) Controversis in the foundation of statistics. *American Mathematical Monthly*, 85, 231-246.

Fraser D.A.S. and Reid, N. (1990) *Statistical Inference: Some Theoretical Methods and Directions*, *Environmetrics*, 1(1): 21-35.

Hacking, I. (1965) *Logic of Statistical Inference*. Cambridge University Press.

Lehmann, E. L. (1959) *Testing Statistical Hypothesis*. New York: Wiley.

Lehmann, E. L. (1983) *Theory of Point Estimation*. New York: Wiley.

Lehmann, E. L. (1998) *Elements of Large-Sample Theory*. New York: Springer Verlag.

Lehmann, E.L. (2011) *Fisher, Neyman, and the Creation of Classical Statistics*. New York: Springer Verlag.

Mayo, D.G. (2013) *Philosophy of Statistics*. [Course homepage](#). See also [link](#).

Morrison, D.E. and Henkel, R.E. Eds. (1970) *The Significance Test Controversy – A Reader*. Butterworths.

Pearl, J. (2000) *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Robins, J. and Wasserman, L. (2000) Conditioning, Likelihood, and Coherence: A review of some fundamental concepts. *Journal of the American Statistical Association*, 95, 1340-1346.

Royall, R. (1997) *Statistical Evidence A. The likelihood paradigm*. Chapman & Hall.

Schweder, T. and Hjort, N. L. (2014) *Confidence, Likelihood, Probability*. Cambridge University Press.

Taper, M.L. and Lele, S.R. (2010) *The Nature of Scientific Evidence Statistical, Philosophical and Empirical Considerations*. University of Chicago Press.

Wang, C. (1993) *Sense and Nonsense of Statistical Inference*. New York: Marcel Dekker Inc.