Summer Seminar: Philosophy of Statistics

Lecture Notes 5: From Probability Theory to Statistical Inference

Aris Spanos [Summer 2019]

1 Introduction

In chapter 2 we began a long journey exploring *probability theory*, with a view to set up a mathematical framework for modeling *stochastic phenomena*. The theory of probability discussed so far has been purely mathematical in nature, despite our best efforts to bring out the connections between probabilistic assumptions and real-world data in chapters 5-7. Center stage in this discussion has occupied the concept of a *statistical model*, which provides the cornerstone of *statistical inference* aiming to help us 'learn from data' about stochastic phenomena of interest; the subject matter of chapters 10-15. The primary objective of this chapter is to built bridges between probability theory and statistical inference that deals with actual data.

What might seem surprising at first sight is that the nature of statistical inference depend crucially on the *interpretation* of mathematical probability adopted, which determines its underlying *reasoning* for learning from data, and guides their *effectiveness*. This is because what constitutes evidence for or against an inferential claim depends in the interpretation of probability. Although there several interpretations of probability discussed in the literature, we will focus primarily on two:

(i) the frequentist, (ii) the degrees of belief,

that dominate current statistics**.** We will largely ignore the *classical* (games of chance) interpretation of probability, based on a finite number of equally likely outcomes, The **classical** interpretation was used in the context of games of chance and was viewed as stemming from *equally likely and finite set of outcomes* based on some sort of physical symmetry; see Laplace (1814). This is only of historical interest since it plays no role in current statistical inference; see Hacking (2006). In addition, the term classical interpretation of probability often leads to confusion with regard to the *classical* approach to statistical inference which is based on the relative *frequency* interpretation of probability.

2 Mathematical Probability: a brief summary

2.1 Kolmogorov's Axiomatic Approach

The axiomatic approach to probability is specified by a probability space $(S, \Im, \mathbb{P}(.))$:

(a) S denotes the set of all possible distinct outcomes.

(b) $\Im$ denotes a set of subsets of S, called *events* of interest, endowed with the mathematical structure of a σ-field, that is, it satisfies the following conditions:

(i) $S \in \Im$, (ii) if $A \in \Im$, then $\overline{A} \in \Im$, (iii) if $A_i \in \Im$ for $i=1, 2, \ldots, n, \ldots$, then $\bigcup_{i=1}^{\infty} A_i \in \Im$.

(c) $\mathbb{P}(.)$**:** $\Im \to [0, 1]$ denotes a set function that satisfies the axioms in table 10.1.

Table 10.1: Kolmogorov Axioms of Probability

[A1]	$\mathbb{P}(S)=1$, for any outcomes set S,
[A2]	$\mathbb{P}(A) \geq 0$, for any event $A \in \Im$,
[A3]	*Countable Additivity.* For a countable sequence of mutually exclusive events, i.e., $A_i \in \Im$, $i=1, 2, ..., n, ..$ such that $A_i \cap A_j = \varnothing$, for all $i \neq j$, $i, j=1, 2, ..., n, ...$, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

This formalization renders probability a sub-field of *measure theory* concerned with assigning size, length, content, area, volume, and etc. to sets. In this sense, the probability space $(S, \Im, \mathbb{P}(.))$ provides an idealized description of the stochastic mechanism that gives rise to the events of interest and related events $\Im$, with $\mathbb{P}(.)$ assigning probabilities to events in $\Im$.

2.2 Random Variables and Statistical Models

In chapters 3-4, the initial Kolmogorov formalism based on $(S, \Im, \mathbb{P}(.))$ was extended by introducing the concept of a *random variable:* a real-valued function:

$$X(.)\text{: } S \to \mathbb{R}, \text{ such that } \{X \leq x\} \in \Im \text{ for all } x \in \mathbb{R}.$$

That is, $X(.)$ assigns numbers to the elementary events in S in such a way so as to preserve the original event structure of interest ($\Im$). This extension is important for bridging the gap between the mathematical model $(S, \Im, \mathbb{P}(.))$ and the observable stochastic phenomena of interest, because observed data usually come in the form of *numbers* on the real line. The key role of the random variable $X(.)$ is to transform the original $(S, \Im, \mathbb{P}(.))^n$ into a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ defined on the real line:

$$(S, \Im, \mathbb{P}(.))^n \quad \stackrel{X(.)}{\longrightarrow} \quad \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) = \{f(\mathbf{x}; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m\},\ \mathbf{x} \in \mathbb{R}_X^n,\ m<n, \tag{1}$$

where $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$ denotes the joint distribution of the sample $\mathbf{X}:=(X_1,\ldots,X_n)$, and Θ the parameter space. Two of the most widely used statistical models are given in tables 10.2-3.

Table 10.2: The simple Bernoulli model

$\mathcal{M}_\theta(\mathbf{x})$: $X_k\backsim\mathsf{BerIID}(\theta,\theta(1-\theta)),\ x_k=0,1,\ \theta\in[0,1],\ k\in\mathbb{N}:=(1,2,\ldots)$

Table 10.3: The simple Normal model

$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: $X_k\backsim\mathsf{NIID}(\mu,\sigma^2),\ x_k\in\mathbb{R},\ \mu\in\mathbb{R},\ \sigma^2>0,\ k\in\mathbb{N}$

The statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ can be viewed as a parameterization of the stochastic process $\{X_k,k\in\mathbb{N}:=(1,2,...,n,...)\}$ whose probabilistic structure is chosen so as to render data $\mathbf{x}_0\mathbf{:}=(x_1,\ldots,x_n)$ a *typical realization* thereof.

The question that naturally arises is:

▶ **can the above Kolmogorov formalism be given an interpretation by assigning a meaning to the primitive mathematical concept probability**?

The different interpretations of probability aim to answer the question: what does probability correspond to in the real world? This correspondence is crucial because it determines the kind of inductive procedures one would follow with a view to "learn from data" about phenomena of interest. That is, it determines the nature of the inductive (statistical) inference called for. The dominating interpretations of probability are currently the 'frequentist' (or frequency) and 'degrees of belief' to be considered next.

3 Frequentist Interpretation(s) of probability

3.1 'Randomness' (stochasticity) is a feature of the real world

What distinguishes the frequentist interpretation from the others is the fact that probability is viewed as directly related to observed relative frequencies in real data; an objective feature of the real world. It refers to the limit of the relative frequency of the occurrence of an event by repeating the experiment that brings about the particular event. In this sense frequentist probability has always been associated with the statistical analysis of data. The frequency interpretation of probability can be traced back to the *statistical regularities* established during the 18th and 19th centuries. After the initial impetus provided by Grant's Bills of Mortality in 1662, there was a concerted effort to collect more and more demographic, economic and social (crimes, violent deaths, etc.) data. The descriptive analysis of these data led to the surprising conclusion that "despite the unpredictability at the individual level (people, firms etc.) there was a remarkable stability of the relative frequencies at the aggregate level (groups) over long periods of time."

By the 1830s the main field of application became *social statistics*: numerical science of society. Its focus was the unpredictability of individual human action and behavior and the search for order (statistical regularity) in larger groups. The main conclusion arising from these studies was that: *regularity could emerge from disorder and irrationality!* Society could be characterized by relatively stable rates of height, weight, education, intelligence, fertility, marriage, crime, suicides and deaths. This, in turn, led to the search for effects whose causes could be discerned in large numbers in an attempt to facilitate the discovery of *laws* analogous to those of *Newtonian mechanics* in the domain of society; see Porter (1986). The protagonist in this search was the Belgian polymath **Quetelet** who, by the 1870s, amassed an impressive collection of evidence of such large-scale statistical regularities; see Stigler (1986), Porter (1986). So much so that the idea of disorder at the individual level leading to order at the aggregate was brought into Physics. Maxwell and Boltzmann, in their attempt to justify their statistical interpretation of *gas laws,* invoked the idea of a model of numerous autonomous and unpredictable individuals (insignificant compared with the assemblage), where regularities characterizing the assemblage and can be used to explain macroscopic behavior, an idea borrowed from Quetelet's *social physics*. This idea led an important pillar of modern Physics known as **Statistical Mechanics**; see Von Plato (1994).

3.2 Model-based frequentist interpretation of probability

As argued in Spanos (2013a; 2017a), the modern variant of the frequentist interpretation of probability, we call *model-based* because it revolves around the concept of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, was initially articulated by Cramer (1946), p. 332, by synthesizing Kolmogorov's mathematical probability with the Fisher-Neyman-Pearson frequentist statistics:

"The mathematical theory belongs entirely to the conceptual sphere, and deals with purely abstract objects. The theory is, however, designed to form a model of a certain group of phenomena in the physical world, and the abstract objects and propositions of the theory have their counterparts in certain observable things, and relations between things. If the model is to be practically useful, there must be some kind of general agreement between the theoretical propositions and their empirical counterparts."

The frequentist interpretation relates to a specific *objective*: modeling observable phenomena of interest exhibiting chance regularity patterns with a view to learn from data about such phenomena. Neyman (1952), p. 27, described statistical modeling process as follows:

"The application of the theory involves the following steps:

(i) If we wish to treat certain phenomena by means of the theory of probability we must find some element of these phenomena that could be considered as random, following the law of large numbers. This involves a construction of a mathematical model of the phenomena involving one of more probability sets.

(ii) The mathematical model is found satisfactory, or not. This must be checked by observation.

(iii) If the mathematical model is found satisfactory, then it may be used for deductions concerning phenomena to be observed in the future." (Neyman, 1952, p. 27)

In (i) Neyman demarcates the domain of statistical modeling to stochastic phenomena that exhibit chance regularities, in the form of the long-run stability of relative frequencies. In (ii) he provides a clear statement concerning the nature of specification and model validation, and in (iii) he brings out the role of ascertainable error probabilities in assessing the optimality of inference procedures.

The current model-based frequentist approach to statistical modeling and inference was pioneered by Fisher (1922a) and extended by Neyman and Pearson (1933). The key to Fisher's approach to statistics was the concept of **a prespecified parametric statistical model** that provides the proper context for assigning probabilities to the relevant events associated with data.

The statistical model can be viewed as a stochastic Generating Mechanism (GM) with prespecified premises that give rise to deductively derived inference propositions, in the form of 'optimal' estimators, tests and predictors. In contrast to deduction, the validity of the inductive inferences requires the **soundness** of the prespecified premises vis-á-vis the data: the probabilistic assumptions imposed on the data. This is often insufficiently appreciated by the current discussions of the frequentist interpretation of probability. These developments provided a coherent grounding for the frequentist interpretation of probability anchored on stable "long-run" frequencies that is grounded on the **Strong Law of Large Numbers** (SLLNs).

3.2.1 The Strong Law of Large Numbers revisited

The probability of an event A, say, $\mathbb{P}(A)=p$, is directly related to the relative frequency of the occurrence of event A, as defined in the context of $(S,\Im,\mathbb{P}(.))$. The traditional frequentist interpretation is articulated in terms of the "long-run" metaphor which states that in a sequence of trials the relative frequency s_n of occurrence of event A will approximate (oscillate around) the value of its true probability $p=\mathbb{P}(A)$. This metaphor, however, carries the seeds of a potential confusion between the stochastic process $\{X_k,\ k\in\mathbb{N}\}$ itself and one its finite realizations $\{x_k\}_{k=1}^n$.

How is such an interpretation formally justified? The simple answer is that it is justified by invoking the SLLN; see chapter 9. Under probabilistic assumptions (restrictions) on the stochastic process $\{X_k,\ k\in\mathbb{N}\}$, the most restrictive being that it is IID, one can prove *mathematically* that:

$$\mathbb{P}(\lim_{n\to\infty}(\tfrac{1}{n}\textstyle\sum_{k=1}^{n}X_k)=p)=1 \tag{2}$$

known as *convergence almost surely* (a.s). The SLLN asserts that for any IID process $\{X_k,\ k\in\mathbb{N}\}$ and any event $A\in\Im$, the relative frequency of occurrence of A converges to $\mathbb{P}(A)$ with probability one. Borel (1909) proved the original SLLN using a Bernoulli

IID process (table 10.2) and these assumptions have been weakened considerably since then. For instance, when $\{X_k,\ k\in\mathbb{N}\}$ is a *martingale difference* process (chapter 9) the result in (2) holds. In an attempt to delineate what this result asserts more clearly, let us bring out what it does *not* claim.

First, the result in (2) does *not* involve any claims that the **sequence of numbers** $\{s_n\frac{1}{n}\sum_{k=1}^{n}x_k\}_{n=1}^{\infty}$ converges to p.

Second, (2) refers only to **what happens at the limit** $n=\infty$, and asserts nothing about the accuracy of $(\frac{1}{n}\sum_{k=1}^{n}x_k)$ as an approximation of $\mathbb{P}(A)$ for a given $n<\infty$; for that one needs to use the Law of Iterated Logarithm (LIL) (chapter 9).

Third, one can appeal to the SLLN only after the **invoked assumptions are validated** vis-á-vis the data; see Spanos (2013a).

3.2.2 Relating the SLLN to chance regularities

The *frequentist interpretation* identifies the probability of an event A with the (probabilistic) *limit* of the relative frequency of its occurrence, $s_n=\frac{1}{n}\sum_{k=1}^{n}x_k$, in the context of a well defined stochastic mechanism represented by the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. The proposed frequentist interpretation has several crucial features:

(i) it revolves around the notion of a statistical model $\mathcal{M}_{\theta}(\mathbf{x})$,

(ii) it is firmly anchored on the SLLN,

(iii) it is justified on empirical and not a priori grounds, and (iv) the key link between the SLLN and the observed stochastic phenomena comes in the form of the *stipulated provisions* in table 10.4.

Table 10.4: Model-based frequentist rendering: interpretative provisions

[i] data $\mathbf{x}_0{:=}(x_1,x_2,\ldots,x_n)$ is viewed as a 'truly typical' finite realization of the process $\{X_k,\ k\in\mathbb{N}\}$ underlying the statistical model $\mathcal{M}_{\theta}(\mathbf{x})$, and
[ii] the 'typicality' of $\mathbf{x}_0$ (e.g. IID) can be evaluated using misspecification testing.

That is, the validity of the model assumptions secures the meaningfulness of identifying the limit of the relative frequencies $\{s_n\}_{n=1}^{\infty}$ with the probability p by invoking Equation (2). Given that the probabilistic assumptions Bernoulli, IID are testable vis-à-vis data $\mathbf{x}_0$, the frequentist interpretation is justifiable on *empirical* and not on *a priori*, grounds.

In the current literature on the various interpretations of probability, the dominating frequentist interpretation is not the model-based articulated above, but the frequentist interpretation put forward by von Mises (1928). The two frequentist interpretations differ in several respects, but most importantly, the model-based circumvents the criticisms leveled against the von Mises interpretation. For that reason, it is important to compare the two interpretations and bring out their differences explicitly.

4 Degree of belief interpretation(s) of probability

4.1 'Randomness' is in the mind of the beholder

Our interest in the 'degree of belief' interpretation of probability stems from the fact that it gives rise to an approach to statistical inference known before the 1950s as the *inverse probability* approach and since then as the *Bayesian approach*.

During the 17th, 18th and most of the 19th centuries, the frequentist and degrees of belief interpretations of probability coexisted happily even in the writings of the same mathematician such as James Bernoulli. Poisson (1837) was the first to make explicit the distinction between the 'objective' (physical) and 'subjective' (non-physical) interpretations of probability, and the battle lines between frequentists and Bayesians were set for the next two centuries. To this day, the focus of the disagreement is primarily on which interpretation of probability is 'objective' and which is 'subjective', instead of focusing the discussion on the underlying reasoning and their primary objective in learning from data about stochastic phenomena of interest.

4.2 Degrees of subjective belief

The subjective degree of belief interpretation of probability has been championed by Ramsey, F.P. (1903–1930), De Finetti, B. (1906–1985) and Savage, L.J. (1917–1971); see Ramsey (1926), De Finetti (1974) and Savage (1954). This interpretation considers the probability of an event A as based on the personal judgment of whoever is assigning the probability; the personal judgement being based on the individual's knowledge and experience. In this sense the probability of event A is based on the person's beliefs and information relating to the experiment giving rise to event A.

Example 10.3. In the case of tossing a fair coin a person is likely to assign the subjective probability $\mathbb{P}(H)=\frac{1}{2}$, because with no information about the chance mechanism involved the two outcomes seem a priori equally likely. In the case where the person in question has additional information relating to the mechanism, such as the coin is bent, the subjective probability is likely to change.

In view of the fact that a person's assessment of probabilities is inextricably bound up with his/her environmental and psychological experiences, probabilities can only be on an individual's experience and knowledge.

According to de Finetti (1974), the most appropriate way to think of probabilities and operarionalize them as **subjective degrees of belief** is in terms of betting odds. The idea is that an agent's (rational individual) 'degree of belief' in the occurrence of event A is equal to p, if and only if (iff), p units of utility is the price at which the agent is willing to buy or sell a bet that pays 1 unit of utility if A occurs and 0 if $\overline{A}$ occurs. The odds *for* a particular event A reflect the 'probability' that the event will occur, while odds *against* reflect the 'probability' that $\overline{A}$ will occur.

Example 10.4. Let us consider the case where somebody offers odds 2 to 1 *for* event A occurring, or in terms of odds ratio, $odds(A)=\frac{1}{2}$. This amounts to claiming

that the person taking the bet has one chance to win and two chances to lose. If the person whose degrees of subjective belief we aim to assess thinks that these are *fair odds*, then we can proceed to evaluate the person's subjective probability for event A, say $\Pr(A)=p$, via:

$$odds(A)=\frac{\Pr(A)}{\Pr(\overline{A})}=\frac{p}{(1-p)} \Rightarrow p=\frac{odds(A)}{1+odds(A)}.$$

Hence, in the above example where $odds(A)=\frac{1}{2}$, $\Pr(A)=\frac{\frac{1}{2}}{1+\frac{1}{2}}=\frac{1}{3}$.

This suggests that the subjective dimension of this probability arises from the fact that it is the decision of a particular individual whether the odds are fair or not. Another individual might consider as fair the odds ratio $odds_1(A)=\frac{1}{3}$, which implies that her subjective probability p_1 is likely to be different from p. This is not surprising because the personal experiences and knowledge, which influence judgement, are often different between individuals.

The Dutch book argument. A Dutch book is a system of bets guaranteeing that taking these bets one will lose no matter what happens. In this sense, a rational agent will not be rational if he/she falls prey to such a loosing scheme of bets. Hence, the subjective degrees of belief interpretation of probability stems from assuming a 'rational agent' who is:

(a) *informed* about an exhaustive set of events of interests, and assigns a degree of belief to each such event in a way that gives rise to a complete ordering,

(b) *consistent* in judgement in the sense that if $\Pr(A)>\Pr(B)$ and $\Pr(B)>\Pr(C)$, then $\Pr(A)>\Pr(C)$, and

(c) *coherent* in judgement in the sense that he/she cannot be fooled by a Dutch book sequence of bets.

According to Ramsey (1926) and De Finetti (1937), a 'rational agent' should be *consistent* and *coherent* in the sense that the probabilities resulting from what such an agent considers as fair odds pertaining to the relevant events of interest should satisfy the axioms of probability [**A1**]-[**A3**], except [**A3**] should be replaced with:

> [**A3**]*. *Finite Additivity.* For a finite sequence of mutually exclusive events, i.e., $A_i \in \Im$, $i=1,2,...,n<\infty$, such that $A_i \cap A_j=\varnothing$, for all $i\neq j$, $i,j=1,2,...,n$, then $\mathbb{P}(\bigcup_{i=1}^{n} A_i)=\sum_{i=1}^{n}\mathbb{P}(A_i)$.

This change raises several technical difficulties because it does away with the continuity of the continuity of $\mathbb{P}(.)$ (see chapter 2).

Example 10.5. Consider an individual whose subjective probabilities associated with two *independent* events A and B are:

$$\Pr(A)=.6,\ \Pr(B)=.3,\ \Pr(A\cap B)=.2.$$

This particular individual is *incoherent* because $\Pr(A\cap B)=.2 \neq \Pr(A)\cdot\Pr(B)=.18$.

Experimental and observational evidence about the above notion of degrees of belief rationality indicate that individuals faced with different bets do not behave rationally as the above description claims; see Kahneman (2013).

A key feature of the subjective degrees of belief interpretation of probability, touted by de Finetti (1974), is that: "Probability does not exist." (p. x), by which he meant that probability in an 'objective sense', as a feature of the real world, does not exist. Rather, probability exists only subjectively within the minds of individuals. de Finetti's followers view probability "as a quantitative characterization of randomness which is intrinsic in nature" (Press, 2003, p.28), and the **degree of belief probability reflects the individual's assessment of the uncertainty stemming from this intrinsic randomness in nature**. What is particularly interesting about this idea is that the same followers of de Finetti reject the argument that such 'randomness' is reflected in the data generated by such natural phenomena in a way that 'randomness', in the form of chance regularities exhibited by data, is rendered objective; a feature that is testable vis-a-vis the data and independent of one's beliefs.

Savage (1954) proposed a seven axiom system based on finite additivity that provides the basis for a theory of rational decision making grounded on Bayesian principles. This system includes axioms that frame the scaling of preferences and the use of lotteries to assess the subjective probabilities of decision makers. Savage's framing was influenced significantly by the pioneering work of von Neumann and Morgenstern (1947) on evaluating the outcomes of game theoretic decision making using expected utility. Indeed, optimal decision making under uncertainty based on maximizing expected utility or minimizing expected loss has become an integral part of modern economics.

4.3 Degrees of 'objective belief': logical probability

Another question with regard to the degree of belief interpretation of probability that comes to mind is ▶ whether one could find some way to establish that a particular odds ratio will be considered fair by any *rational person*; assuming a formal definition of *rational*. In such a case the personal dimension of the interpretation will change to a more objective one. The first to attempt such a recasting was the economist John Maynard Keynes (1883–1946), in Keynes (1921), and followed by the philosopher Rudolf Carnap (1891–1970) in Carnap (1950). The interpretation of subjective probability based on odds ratios which will be considered fair by any rational person is often called *logical probability*, as opposed to the *personal* subjective probability. The scientist who straddles the logical and the degrees of belief interpretations is the geologist Harold Jeffreys (1891–1989) in Jeffreys (1939); see Gavalotti (2005). In the context of the logical interpretation of probability, events and relations among events in the Kolmogorov mathematical framing in terms of a probability space $(S, \Im, \mathbb{P}(.))$ have a primary logical character: one may assign to each event $A \in \Im$ a proposition

relating to its occurrence. This will render logical relations between propositions into relations between events. The logical interpretation views the probability of a proposition or event A 'occurs' be a function of another event or proposition B relating to 'evidence'. When the evidence B is not sufficient to deductively lead to A, one may be able to evaluate the degree of support (or confirmation) of A by evidence B using logical probability. That is, for the logical interpretation of probability:

(a) probabilities are determined a priori (they cannot be justified empirically),

(b) probability is viewed as a logical relation between propositions (events), and

(c) a probabilistic assignment is always relative to a given evidence (proposition);

see Barnett (1999) and Fine (1973) for further discussion.

To bring the crucial difference between the frequentist, subjective degrees of belief and the logical interpretation of probability, let us consider the following example.

Example 10.6. Consider a weather forecast asserting that 'the probability of the event A-it will rain this afternoon at 3:30pm', is $\Pr(A)$=.6. The model-based frequentist interpretation would require the existence of a statistical model that could be estimated by past data on observables pertaining to the atmospheric pressure, rainfall, sunshine and other weather-related phenomena. What justifies the reliability of the forecast is the statistical adequacy of the estimated model that needs to be secured beforehand. In such a case $\Pr(A)$=.6 denotes the probability of event A that would often include a degree of uncertainty in the form of a forecast interval. The degree of subjective beliefs interpretation would require a rational agent to quantify the uncertainty associated with the intrinsic randomness associated with whether phenomena, such as raining, and output the forecast $\Pr(A)$=.6. The only justification needed is the individual's expertise and knowledge about such phenomena; the rational agent uses such knowledge to pick a number in the interval $[0, 1]$. The logical interpretation of probability would require the forecaster to declare the body of evidence E upon which the degree of support $\Pr(A)$=.6 it was based.

4.4 Which interpretation of probability?

In light of the fact that the interpretation of probability will determine the type of approach for statistical modeling and inference, it is important to address the question of 'which' interpretation of probability gives rise to 'what' type of statistical modeling and inference. The answer will inevitably be based on the author's knowledge and experience with such endeavors. In that sense, the reader is advised to interpret the comments that followed in that light.

From the statistical modeling and inference perspective, there is no doubt in this author's mind that the frequentist interpretation of probability and the associated approach to modeling and inference is superior to the Bayesian approach in several respects. As argued in chapter 1, a statistical model is built upon the systematic information contained in the observed data in an attempt to provide an appropriate description of the stochastic mechanism that gave rise to the data. The stance that

observed data $\mathbf{x}_0$:=$(x_1, ..., x_n)$ contain systematic statistical information in the form of chance regularity patterns, 'stochasticity' (randomness) is a feature of real-world phenomena and exists independently of one's beliefs; its appropriateness can be tested against the data. Moreover, the frequentist interpretation of probability provides a way to relate these regularities to abstract statistical models in a manner that renders the probabilistic assumptions of the model testable vis-a-vis data $\mathbf{x}_0$. In addition, learning from data about observable stochastic phenomena cannot be solely in the mind a particular individual, having to do with revising an individual's degrees of belief represented by a prior and a posterior distribution. Scientific knowledge needs to be testable and independent of one's beliefs.

5 Frequentist vs. Bayesian statistical inference

The two main approaches to statistical inference, associated with the frequentist and degrees of belief interpretations of probability can be usefully distinguished on three main grounds give in table 10.6.

Table 10.6: Delimiting features of different approaches to inference

[a] the interpretation of mathematical probability,
[b] the role and nature of the relevant information for inference purposes,
[c] inductive reasoning: how we learn from data, and the nature of learning.

5.1 The frequentist approach to statistical inference

The father of modern frequentist statistics is without a doubt R.A. Fisher. According to a noted historian of statistics: "Fisher was a genius who almost single-handedly created the foundations for modern statistical science, without detailed study of his predecessors." (Hald, 1998, p. 738).

In the terms of the three main grounds stated above, for the frequentist approach:

[a] The interpretation of probability is *frequentist*: the relative frequencies associated with the long-run metaphor (in a hypothetical set up) reflect the corresponding probabilities; the formal link comes in the form of the SLLN.

[b] The chance regularities exhibited by data $\mathbf{x}_0$:=$(x_1, x_2, \ldots, x_n)$ constitute the *only relevant statistical information* for selecting the statistical model. The probabilistic information that aims to account for all the chance regularities in data $\mathbf{x}_0$ is specified in the form of a statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,$$

where $\boldsymbol{\Theta}$ denotes the parameter space, $\mathbb{R}^n_X$ the sample space, and $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$ the (joint) distribution of the sample. Equivalently, the data $\mathbf{x}_0$ are viewed as a

'typical realization' of a stochastic mechanism described by a statistical model, The *substantive* information comes primarily in the form of a substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ whose parameters $\boldsymbol{\varphi}$ are related to statistical parameters $\boldsymbol{\theta}$ via $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$. The latter implies restrictions that should not be imposed at the outset so that they can tested against the data. $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ provides the cornerstone of all forms of inference which will be framed in terms of the unknown parameters $\boldsymbol{\theta}\in\Theta$.

[c] The primary aim of the frequentist approach is to *learn from data* about the "true" statistical data GM: $\mathcal{M}_{\boldsymbol{\theta}^*}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\},\ \mathbf{x}\in\mathbb{R}^n_X$

The expression "$\boldsymbol{\theta}^*$ denotes the true value of $\boldsymbol{\theta}$" is a shorthand for saying that "data $\mathbf{x}_0$ constitute a typical realization of the sample $\mathbf{X}$ with distribution $f(\mathbf{x};\boldsymbol{\theta}^*)$".

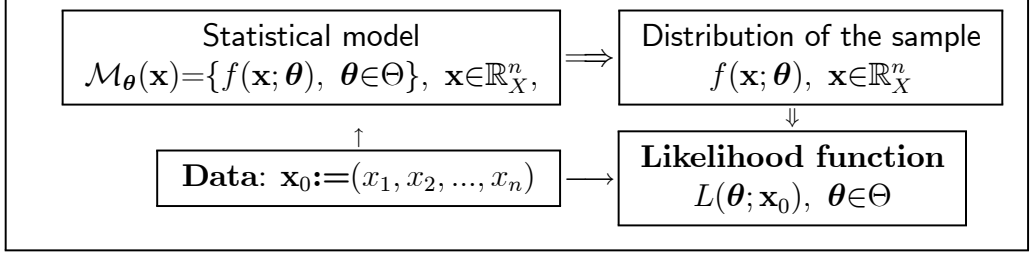
The frequentist interpretation of probability is inextricably bound up with what Neyman (1977), p. 99, called *stable long-run relative frequencies* of events of interest. $\mathcal{M}_{\theta}(\mathbf{x})$ is assumed to represent an idealized GM that could have given rise to data $\mathbf{x}_0$. The existence of the probabilities described by $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$, depends crucially on being able to demonstrate that one can estimate consistently the *invariant features* of the phenomenon being modeled, as reflected in the constant but unknown parameters $\boldsymbol{\theta}$. A crucial role in the frequentist 'learning from data' is played by the *likelihood function*:

$$L(\boldsymbol{\theta};\mathbf{x}_0)\propto f(\mathbf{x}_0;\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta,$$

where $\propto$ reads 'proportional to' the distribution of the sample evaluated at $\mathbf{x}_0$.

The frequentist approach to statistical inference is summarized in table 10.7.

Table 10.7: The frequentist approach to statistical inference



Example 10.7. Consider the *simple Bernoulli model* specified in table 10.2. The IID, Bernoulli assumptions imply that $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, takes the form:

$$\begin{aligned} f(\mathbf{x};\theta) &\overset{\text{I}}{=}\prod_{k=1}^n f_k(x_k;\theta_k)\overset{\text{IID}}{=}\prod_{k=1}^n f(x_t;\theta)\overset{\text{Ber}}{=}\prod_{k=1}^n \theta^{X_k}(1-\theta)^{1-X_k}= \\ &=\theta^{\sum_{k=1}^n X_k}(1-\theta)^{\sum_{k=1}^n(1-X_k)},\ \mathbf{x}\in\{0,1\}^n. \end{aligned} \tag{3}$$

Hence, the Likelihood Function (LF) for the simple Bernoulli model is:

$$L(\theta;\mathbf{x}_0)\propto\theta^{\sum_{k=1}^n x_k}(1-\theta)^{\sum_{k=1}^n(1-x_k)}=\theta^{n\overline{x}}(1-\theta)^{n(1-\overline{x})},\ \theta\in[0,1], \tag{4}$$

where $\overline{X}=\frac{1}{n}\sum_{k=1}^{n}X_k$. Note that $f(\mathbf{x};\theta)$, $\mathbf{x}\in\{0,1\}^n$ is a discrete density function for $\overline{X}$, but the Likelihood Function (LF), $L(\theta;\mathbf{x}_0)$, $\theta\in[0,1]$, is a continuous function of $\theta\in[0,1]$.

Example 10.8. For the *simple Normal model* (table 10.3), the NIID assumptions imply that:

$$f(\mathbf{x};\boldsymbol{\theta}) \overset{\text{I}}{=} \prod_{k=1}^{n} f_k(x_k;\boldsymbol{\theta}_k) \overset{\text{IID}}{=} \prod_{k=1}^{n} f(x_t;\boldsymbol{\theta}) \overset{\text{NIID}}{=} \prod_{k=1}^{n} \tfrac{1}{\sqrt{2\pi\sigma^2}}\exp(-\tfrac{(x_k-\mu)^2}{2\sigma^2}) = \\ =(\tfrac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left\{\tfrac{1}{2\sigma^2}\textstyle\sum_{k=1}^{n}(x_k-\mu)^2\right\},\ \mathbf{x}\in\mathbb{R}^n. \tag{5}$$

Hence, the LF for the simple Normal model takes the form:

$$L(\theta;\mathbf{x}_0)\propto(\tfrac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left\{\tfrac{1}{2\sigma^2}\textstyle\sum_{k=1}^{n}(x_k-\mu)^2\right\},,\ \boldsymbol{\theta}:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+. \tag{6}$$

Learning from data. When probability is interpreted in frequency terms, the objective of learning from data is attained by employing *effective* and *reliable* inference procedures to learn about $\boldsymbol{\theta}^*$, the 'true' value of $\boldsymbol{\theta}\in\Theta$. The effectiveness and reliability is evaluated using ascertainable *error probabilities* associated with different procedures. The underlying inductive reasoning comes in two forms:

(i) **factual**: the true state of nature ($\boldsymbol{\theta}^*$) (estimation and prediction), whatever that happens to be, and

(ii) **hypothetical**: various hypothetical scenarios are compared to what actually happened (hypothesis testing). Note that the observed data $\mathbf{x}_0$ are assumed to have been generated by $\mathcal{M}^*(\mathbf{x})$.

In the context of the broader scientific inquiry one begins with substantive questions of interest that pertain to phenomena of interest, and the ultimate objective of frequentist inference (estimation, testing, prediction) is to use data $\mathbf{x}_0$ to learn about the true data-generating mechanism $\mathcal{M}^*(\mathbf{x})$, $\mathbf{x}\in\mathbb{R}^n_X$.

The cornerstone of the proposed frequentist interpretation of probability is provided by the concept of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, $\mathbf{x}\in\mathbb{R}^n_X$, which plays a pivotal role in both modeling and inference because:

(i) it specifies the inductive premises of inference,

(ii) it determines what constitutes *a legitimate* event,

(iii) it assigns probabilities to all legitimate events via $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$,

(iv) it defines what are legitimate hypotheses and/or inferential claims,

(vi) it designates what constitute legitimate data $\mathbf{x}_0$ for inference purposes,

(v) it determines the relevant error probabilities in terms of which the optimality and reliability of inference methods is assessed. This is achieved by determining the sampling distribution of any statistic (estimator, test statistic, and predictor), say $T_n=g(X_1,\ldots,X_n)$, via:

$$F(t;\boldsymbol{\theta})=\mathbb{P}(T_n\leq t;\boldsymbol{\theta})=\underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ g(\mathbf{x})\leq t;\ \mathbf{x}\in\mathbb{R}^n_X\}} f(\mathbf{x};\boldsymbol{\theta})dx_1dx_2\cdots dx_n \tag{7}$$

This indicates that in frequentist inference $f(\mathbf{x};\boldsymbol{\theta})$ provides the sole source of relevant probabilities, including error probabilities used to evaluate the reliability of inferential procedures; see Spanos (1986).

The learning from data in the frequentist framework is achieved in three steps.

Step 1. *Specify* a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ with a twofold objective:

(i) $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ fully accounts for the chance regularities in data $\mathbf{x}_0$, and

(ii) the particular reparameterization of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is selected with a view to pose the substantive questions of interest by parametrically nesting substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

Step 2. Secure the *statistical adequacy* (validity vis-a-vis data $\mathbf{x}_0$) of the probabilistic assumptions underlying the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ using thorough Mis-Specification (M-S) testing. If the original $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is statistically misspecified, *respecify* it to account for all the statistical information in $\mathbf{x}_0$.

Step 3. After securing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the modeler probes for the *empirical validity* of the substantive model, say $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$, where $\boldsymbol{\varphi}\in\Phi$ denotes the substantive parameters of interest. The latter aims to explain the essential features of the phenomenon of interest. The probing takes the form of relating the *statistical* ($\boldsymbol{\theta}$) and *substantive* ($\boldsymbol{\varphi}$) parameters, via $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$, and testing any restrictions such a mapping entails. In addition, the modeler can use the statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ to probe:

(a) **substantive adequacy**: does the model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ shed adequate light on (describe, explain, predict) the phenomenon of interest?

It is very important to contrast the above question with:

(b) **statistical adequacy**: does $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ adequately account for the chance regularities in $\mathbf{x}_0$?

Widely quoted slogans such as: "All models are wrong, but some are useful." (Box, 1979), are also misleading. The people invoking such slogans *confuse* substantive with statistical inadequacy. It is one thing to say that a structural model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ is a crude approximation of the reality it aims to capture, and entirely another to claim that the assumed statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ could *not* have generated data $\mathbf{x}_0$, which is what statistical inadequacy amounts to. Hence, a structural model may always come up short in securing a substantively adequate $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ for the phenomenon of interest, but $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ may be perfectly adequate for answering substantive questions of interest.

5.2 The Bayesian approach to statistical inference

In order to avoid any misleading impressions it is important to note that there are numerous variants of Bayesianism; more than 46656 varieties of Bayesianism according to Good (1971)! In this section we discuss some of the elements of the Bayesian approach which are shared by most variants of Bayesianism.

Bayesian inference, like frequentist inference, begins with a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, but modifies the inferential set up in two crucial respects:

(i) the unknown parameter(s) $\boldsymbol{\theta}$ are now viewed as *random variables* (not unknown constants) with their own distribution, known as the *prior distribution*:

$$\pi(.)\colon\ \Theta \to [0,1],$$

which represents the modeler's assessment of how likely the various values of $\boldsymbol{\theta}$ in Θ are *a priori*, and

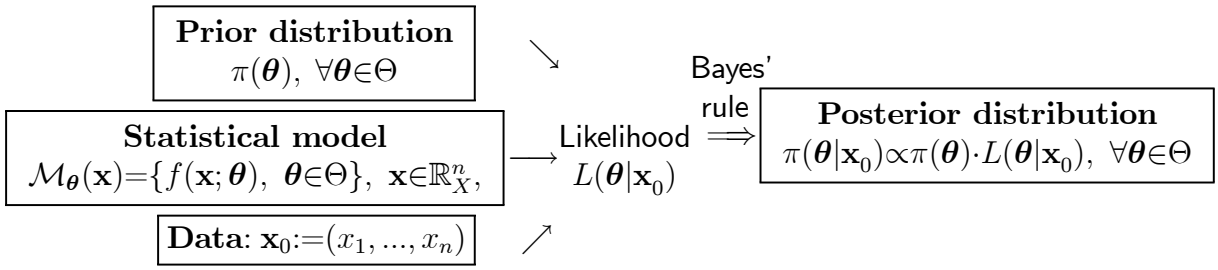
(ii) the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta})$ is re-interpreted by Bayesians to be defined as *conditional* on $\boldsymbol{\theta}$, and denoted by $f(\mathbf{x}|\boldsymbol{\theta})$.

Taken together these modifications imply that there exists a *joint distribution* relating the unknown parameters $\boldsymbol{\theta}$ and a sample realization $\mathbf{x}$:

$$f(\mathbf{x},\boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta}),\ \ \forall\boldsymbol{\theta}\in\Theta.$$

Bayesian inference is based exclusively on the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$ which is viewed as the revised (from the initial $\pi(\boldsymbol{\theta})$) *degrees of belief* for different values of $\boldsymbol{\theta}$ in light of the summary of the data by $L(\boldsymbol{\theta}|\mathbf{x}_0)$.

Table 10.8: The Bayesian approach to statistical inference



Example 10.9. Consider the *simple Bernoulli model* (table 10.2), and let the **prior** $\pi(\theta)$ be $\mathsf{Beta}(\alpha,\beta)$ distributed with a density function:

$$\pi(\theta)=\tfrac{1}{\mathsf{B}(\alpha,\beta)}\theta^{(\alpha-1)}(1-\theta)^{\beta-1},\ \alpha>0,\ \beta>0,\ 0<\theta<1. \tag{8}$$

Combining the likelihood in (4) with the prior in (8) yields the **posterior** distribution:

$$\begin{aligned}\pi(\theta|\mathbf{x}_0) &\propto \left(\tfrac{1}{\mathsf{B}(\alpha,\beta)}\theta^{(\alpha-1)}(1-\theta)^{\beta-1}\right)\left[\theta^{n\overline{x}}(1-\theta)^{n(1-\overline{x})}\right]= \\ &= \tfrac{1}{\mathsf{B}(\alpha,\beta)}\left[\theta^{n\overline{x}+(\alpha-1)}(1-\theta)^{n(1-\overline{x})+\beta-1}\right].\end{aligned} \tag{9}$$

In view of the formula in (8), (9) as an 'non-normalized' density of a $\mathsf{Beta}(\alpha^*, \beta^*)$, where:

$$\alpha^*{=}n\overline{x}+\alpha, \quad \beta^*{=}n(1-\overline{x})+\beta. \tag{10}$$

As the reader might have suspected, the choice of the prior in this case was not arbitrary. The Beta prior in conjunction with a Binomial-type LF gives rise to a Beta posterior. This is known in Bayesian terminology as a *conjugate pair*, where $\pi(\theta)$ and $\pi(\theta|\mathbf{x}_0)$ belong to the same family of distributions.

Savage (1954), one of the high priests of modern Bayesian statistics, summarizes Bayesian inference succinctly by asserting that: 'Inference means for us the change of opinion induced by evidence on the application of Bayes' theorem." (p. 178).

In the terms of the main grounds stated above, the Bayesian approach:

[a] Adopts the degrees of belief interpretation of probability introduced via $\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}{\in}\Theta$.

[b] The relevant information includes both (i) the data $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$, and (ii) prior information. Such prior information comes in the form of a prior distribution $\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}{\in}\Theta$, which is assigned *a priori* and represents one's degree of belief in ranking the different values of $\boldsymbol{\theta}$ in Θ as more probable and less probable.

[c] The primary aim of the Bayesian approach is to **revise** the original *ranking* based on $\pi(\boldsymbol{\theta})$ in light of the data $\mathbf{x}_0$ by updating in the form of the *posterior distribution*:

$$\pi(\boldsymbol{\theta}|\mathbf{x}_0){=}\tfrac{f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\propto}L(\boldsymbol{\theta}|\mathbf{x}_0)\cdot\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}{\in}\boldsymbol{\Theta}, \tag{11}$$

where $L(\boldsymbol{\theta}|\mathbf{x}_0) \propto f(\mathbf{x}_0|\boldsymbol{\theta})$ denotes a *re-interpreted* likelihood function as being conditional on $\mathbf{x}_0$. The Bayesian approach is depicted in table 10.8. Since the denominator $m(\mathbf{x}_0){=}\int_{\theta\in(0,1)}\pi(\boldsymbol{\theta})f(\mathbf{x}_0|\boldsymbol{\theta})d\boldsymbol{\theta}$, known as the **predictive** distribution, derived by integrating out $\boldsymbol{\theta}$, can be absorbed into the constant of proportionality in (11) and ignored for most practical purposes. The only exception to that is when one needs to treat $\pi(\theta|\mathbf{x}_0)$ as a proper density function which integrates to one, $m(\mathbf{x}_0)$ is needed as a normalizing constant.

Learning from data. In this context, learning from data $\mathbf{x}_0$ takes the form revising one's degree of belief for different values of $\boldsymbol{\theta}$ [i.e. different models $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \boldsymbol{\theta}{\in}\Theta$], in light of data $\mathbf{x}_0$, the learning taking the form $\pi(\boldsymbol{\theta}|\mathbf{x}_0){-}\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}{\in}\Theta$.. That is, the learning from data $\mathbf{x}_0$ about the phenomenon of interest takes place in the head of the modeler. In this sense, the underlying inductive reasoning is neither *factual* nor *hypothetical*, it's *all-inclusive* in nature: it pertains to *all* $\boldsymbol{\theta}$ in Θ, as ranked by $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$. Hence, Bayesian inference does not pertain directly to the real world phenomenon of interest per se, but to one's beliefs about $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \boldsymbol{\theta}{\in}\Theta$.

5.2.1 The choice of prior distribution

Over the last two decades the focus of disagreement among Bayesians has been the choice of the prior. Although the original justification for using a prior is that it

gives a modeler the opportunity incorporate substantive information into the data analysis, the discussions among Bayesians in the 1950s and 1960s made computational convenience the priority for the choice of a prior distribution, and that led to *conjugate priors* that ensure that the prior and the posterior distributions belong to the same family of distributions; see Berger (1985). More recently, discussions among Bayesians shifted the choice of the prior question to 'subjective' vs. 'objective' prior distributions. The concept of an 'objective' prior was pioneered by Jeffreys (1939) in an attempt to address Fisher's (1921) criticisms of Bayesian inference that routinely assumed a Uniform prior for θ as an expression of ignorance. Fisher's criticism was that if one assumes that a Uniform prior $\pi(\theta),\ \forall\theta\in\Theta$ expresses ignorance because all values of θ are assigned the same prior probability, then a reparameterization of θ, say $\phi{=}h(\theta)$, will give rise to a *very informative* prior for ϕ.

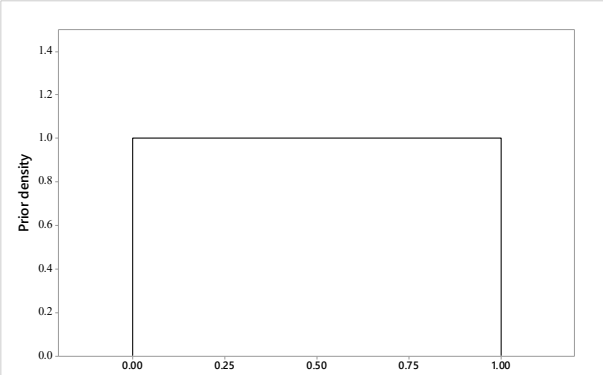


Fig. 10.1: Uniform prior density of θ

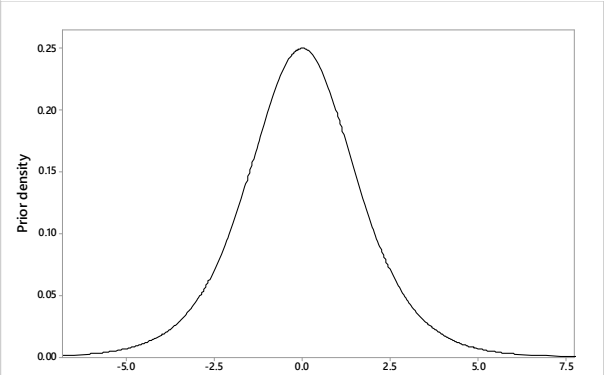


Fig. 10.2: Logistic prior density of ϕ

Example 10.10. In the context of the simple Bernoulli model (table 10.2), let the prior be $\theta\backsim\mathsf{U}(0,1),\ 0\leq\theta\leq 1$ (figure 10.1). Note that $\mathsf{U}(0,1)$ is a special case of the $\mathsf{Beta}(\alpha,\beta)$, for $\alpha{=}\beta{=}1$.

Reparameterizing θ into $\phi{=}\ln(\frac{\theta}{1-\theta})$, implies that $\phi\backsim\mathsf{Logistic}(0,1)$, $-\infty<\phi<\infty$. Looking at the prior for ϕ (figure 10.2) it becomes clear that the ignorance about θ has been transformed into substantial knowledge about the different values of ϕ.

In his attempt to counter Fisher's criticism, Jeffreys proposed a form of prior distribution that was invariant to such transformations. To achieve the reparameterization invariance Jeffreys had to use Fisher's information associated with the score

function and the Cramer-Rao lower bound; see chapters 11-12.

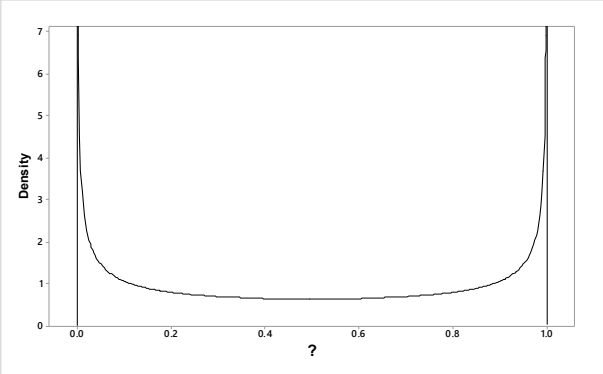


Fig. 10.3: Jeffreys prior

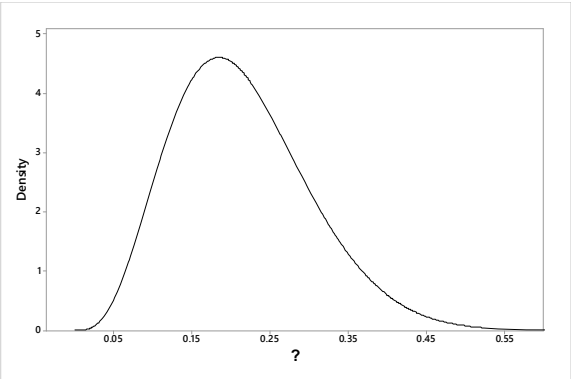


Fig. 10.4: Jeffreys posterior

Example 10.11. In the case of the simple Bernoulli model (10.2), the Jeffreys' prior is:

$$\pi_J(\theta) \backsim \mathsf{Beta}(.5, .5),\ \ \pi_J(\theta)=\frac{\theta^{-.5}(1-\theta)^{-.5}}{\mathsf{B}(.5,.5)}$$

In light of the posterior in (9), using $n\overline{x}$=4, n=20 gives rise to:

$$\pi_J(\theta|\mathbf{x}_0) \backsim \mathsf{Beta}(4.5, 16.5).$$

For comparison purposes, the Uniform prior $\theta \backsim \mathsf{U}(0, 1)$ in example 10.10 yields:

$$\pi(\theta|\mathbf{x}_0) \backsim \mathsf{Beta}(5, 17).$$

Attempts to extend Jeffreys prior to models with more than one unknown parameter initiated a variant of Bayesianism that uses what is called *objective* (default, reference) priors because they minimize the role of the prior distribution and maximize the contribution of the likelihood function in deriving the posterior; see Berger (1985), Bernardo and Smith (1994).

5.3 Cautionary notes on misleading Bayesian claims

(1) Bayesian textbooks sometimes begin their discussion of statistical inference by asking the reader to view it as analogous to the situation facing a physician seeking to come up with a diagnosis based on certain evidence as well as his knowledge and experience. The physician applies a variety of medical tests to a patient with a view to find out whether he/she suffers from a certain medical disorder. This analogical reasoning, however, is highly misleading. A moment's reflection suggests that the situation facing the physician is not at all analogous to the situation facing an empirical modeler seeking to understand stochastic phenomena of interest, such as the great recession of 2008.

The idea behind this argument is to convince the reader that statistical inference is all about individual decision making and the associated costs and benefits. The

truth of the matter is that learning from data about observable phenomena of interest has nothing to do with decisions and associated costs! As Fisher (1935a) perceptively argued: "In the field of pure research no assessment of the cost of wrong conclusions, or of delay in arriving at more correct conclusions can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence." (pp. 25-26).

Having said that, it is widely accepted that the Bayesian approach is better suited for modeling *decision making under uncertainty*. That, however, does not imply that it is also better for modeling and inference with observational data generated by complex systems such as the economy, or the physical universe.

(2) In their attempt to provide support for their preferred approach to modeling and inference, Bayesians criticize the frequentist approach on a number of different grounds that are often misplaced and misleading. Particularly pernicious are several examples used by Bayesians (Berger and Walpert, 1988) to point out several pathological flaws of the frequentist approach to estimation and testing. As argued in Spanos (2011a-b; 2012b; 2013b-d), all these examples revolve around statistical models that are intrinsically *pathological*, and thus the problems pointed out by Bayesians are primarily due to the flawed statistical models, and not to the weaknesses of frequentist inference.

(3) Bayesian inference makes use of all the available a priori information, but frequentist inference does not.

■ This is a highly misleading claim made often by Bayesians that conflate (intentionally?) *prior substantive matter information* and information in the form of a *prior distribution* $\pi(\boldsymbol{\theta})$, $\forall\boldsymbol{\theta}\in\Theta$. Substantive subject matter information usually comes in the form of the sign and magnitude of substantive parameters of interest, as well as relationships among such parameters. The quintessential example of that is the Simultaneous Equation model in econometrics; see Spanos (1990a). Frequentist inference is tailor-made to accommodate such information by relating the substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ and its parameters $\boldsymbol{\varphi}$, to the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and its parameters $\boldsymbol{\theta}$ via a system of restrictions $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$, $\forall\boldsymbol{\varphi}\in\Phi$, $\forall\boldsymbol{\theta}\in\Theta$. The crucial advantage of the frequentist approach is that the validity of such substantive information can be evaluated using the data by testing the validity of $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$. On the other hand, one will be very hard pressed to find a scientific field where substantive subject matter information comes in the form of a prior distribution $\pi(\boldsymbol{\theta})$, $\forall\boldsymbol{\theta}\in\Theta$, that is, in any way, testable against the data.

(4) Another claim made by Bayesians is that the frequentist approach is as subjective as the Bayesian because:

"... likelihoods are just as subjective as priors." (Kadane, 2011, p. 445)

■ This claim stems from a false equivalence. Objectivity in statistical inference stems from being able to critically evaluate the warrant of any inferential claim, by evaluating the extent to which a modeler has obviated the particular errors that could render the claim false. The adequacy of the probabilistic assumptions defining the

likelihood (the statistical model assumptions) vis-a-vis data $\mathbf{x}_0$, can be tested, not only by the modeler, but anybody else who would like to challenge the misspecification testing results of the modeler. How does one challenge the adequacy of the assumptions defining a prior $\pi(\boldsymbol{\theta})$, $\forall\boldsymbol{\theta}\in\Theta$?

6 An introduction to frequentist inference

6.1 Fisher and neglected aspects of frequentist statistics

R.A. Fisher (1890–1962) is the founder of modern frequentist statistics as a model-based approach to statistical induction anchored on the concept of a *statistical model* he devised:

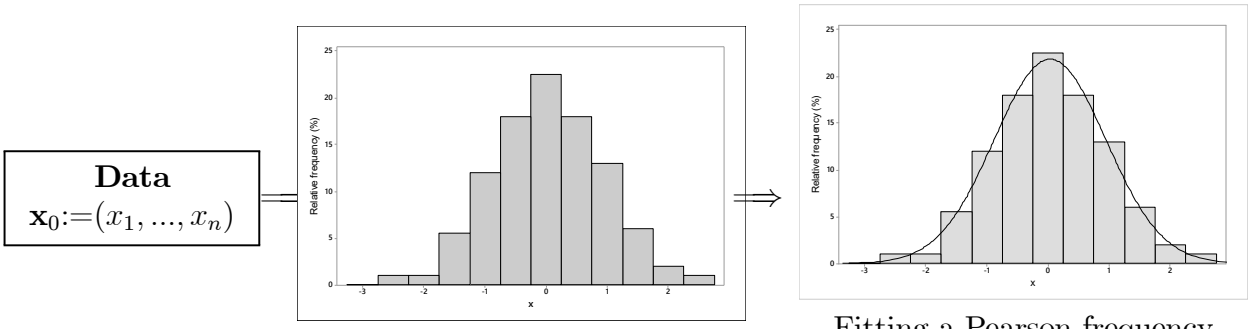
$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \Theta\subset\mathbb{R}^p,\ p<n, \tag{12}$$

where the (joint) distribution of the sample $f(\mathbf{x};\boldsymbol{\theta})$ 'encapsulates' the probabilistic information in the statistical model.

Fisher was able to recast Karl Pearson's (1857–1936) approach to statistics that commenced with data $\mathbf{x}_0{:=}(x_1,...,x_n)$ in search of a *frequency curve* to describe the *histogram* of $\mathbf{x}_0$ (figure 10.3). The choice of the particular frequency curve was narrowed down to, what we call today *the Pearson family of frequency curves* generated by a differential equation in four unknown parameters:

$$\frac{d\ln f(x)}{dx}=\frac{(x-\theta_1)}{\theta_2+\theta_3x+\theta_4x^2}. \tag{13}$$

Depending on the values of the parameters $(\theta_1,\theta_2,\theta_3,\theta_4)$, this equation can generate several frequency curves, including the Normal, the Student's t, the Beta, the Gamma, the Laplace, the Pareto, etc.; see Appendix 12.A for further details.



Histogram of the data

Fitting a Pearson frequency curve $f(x;\widehat{\theta}_1,\widehat{\theta}_2,\widehat{\theta}_3,\widehat{\theta}_4)$

Fig. 10.3: The Karl Pearson approach to statistics

Fisher's (1922a) recasting turned this approach on its head by commencing with a prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ (a 'hypothetical infinite population') that views $\mathbf{x}_0$ as a typical realization thereof. He envisaged the specification of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ as a response to the question: "Of what population is this a random sample?" (Fisher, 1922a,

p. 313), underscoring that: "the adequacy of our choice may be tested a posteriori." (p. 314). Fisher (1922a) identified the 'problems of statistics' to be: (1) **specification**, (2) **estimation** and (3) **distribution** and emphasized that addressing (2)-(3) depended crucially on dealing with (1) successfully first. Fisher classified all forms of inference based on sampling distributions under (3) distribution.

The formal apparatus of frequentist statistical inference, including estimation and testing was largely in place by the late 1930s. Fisher (1922a, 1925b, 1934), almost single-handedly, created the current theory of 'optimal' point estimation and formalized significance testing based on the p-value reasoning. Neyman and Pearson (1933) proposed an 'optimal' theory for hypothesis testing, by modifying/extending Fisher's significance testing; see Pearson (1966). Neyman (1937) proposed an 'optimal' theory for interval estimation analogous to N-P testing. Broadly speaking, the probabilistic foundations of frequentist statistics, as well as the technical apparatus associated with statistical inference methods, were largely in place by the late 1930s, but its philosophical foundations concerned with the proper form of the underlying inductive reasoning were in a confused state.

One of those foundational problems is that the traditional facets of statistical inference, estimation, testing and prediction, do not provide a complete picture of the frequentist approach to statistical modeling and inference. A crucial facet of modeling that received relatively little attention is that of statistical adequacy. Under the heading problems of Distribution, Fisher raises the problem of testing the *adequacy* of the specification (the postulated statistical model): "(iii) Problems of Distribution include the mathematical deduction of the exact nature of the distribution in random samples of our estimates of the parameters and other statistics designed to test the validity of our specification (test of Goodness of Fit)." (see Fisher (1925), p. 8).

In this book we consider the issue of statistical adequacy of paramount importance and view this aspect of *modeling* that includes stages 1-4 in table 10.9. It represents an extension/modification of the original Fisher's (1922a) framework (specification, estimation, distribution) needed to accommodate the M-S testing and respecification facets of modeling, with a view to secure the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

Table 10.9: Empirical modeling and Inference

Modeling	1. Specification 2. Estimation 3. Mis-Specification (M-S) Testing 4. Respecification $\therefore$ *Statistically adequate model*
Inference:	estimation, testing, prediction, simulation

6.2 Basic frequentist concepts and distinctions

Fisher (1922a), in his recasting of statistics, introduced numerous new concepts and ideas and brought out and addressed several confusions permeating Karl Pearson's approach to statistics. To render the preliminary discussion the basic concepts and important distinctions in frequentist inference less abstract, let us focus on a particular example.

Example 10.11. The basic elements of the simple Bernoulli model (table 10.2) are:

(i) the parameter space: $\Theta{:=}[1,0]$,

(ii) the sample space: $\mathbb{R}^n_X{:=}\{0,1\}^n{:=}\{0,1\}\times\{0,1\}\times\cdots\times\{0,1\}$,

A crucial distinction is that between a *sample* $\mathbf{X}{:=}(X_1,\ X_2,...,X_n)$, a set of random variables, and a *sample realization* $\mathbf{x}_0{:=}(x_1,x_2,...x_n)$ that represents just one point in $\mathbb{R}^n_X$.

Example 10.12. For the Bernoulli model (table 10.2) a *sample realization* $\mathbf{x}_0$, say $n{=}30$, would look like:

$$\mathbf{x}_0{:=}(0,0,1,0,1,1,0,0,1,0,0,0,1,1,0,1,0,1,0,0,1,0,1,0,0,0,1,1,0,0).$$

The distinction between a sample $\mathbf{X}$ and one realization $\mathbf{x}_0$ (out of many, often infinite possible ones) takes the form:

$$\begin{array}{rrccccccccc}
\text{Sample:} & \mathbf{X}{:=} & (X_1, & X_2, & X_3, & X_4, & X_5, & X_6, & \cdots & X_{30}) \\
 & & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \cdots & \downarrow \\
\text{Sample realization:} & \mathbf{x}_0{:=} & (0 & 0 & 1 & 0 & 1 & 1 & \cdots & 0)
\end{array}$$

Distribution of the sample vs. the likelihood function. As mentioned above, the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}{\in}\mathbb{R}^n_X$, is specified by the assumptions of a statistical model. This in turn determines the *likelihood function* via:

$$L(\boldsymbol{\theta};\mathbf{x}_0)\propto f(\mathbf{x}_0;\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}{\in}\Theta,$$

where $f(\mathbf{x}_0;\boldsymbol{\theta})$ is evaluated at the observed data point $\mathbf{x}_0$. In contrast to $f(\mathbf{x};\boldsymbol{\theta})$, which varies with $\mathbf{x}{\in}\mathbb{R}^n_X$, the likelihood function varies with $\boldsymbol{\theta}{\in}\Theta$, and measures the *likelihood* (proportional to the probability) associated with the different values of $\boldsymbol{\theta}$, to have been the 'true' parameter(s) $\boldsymbol{\theta}^*$ of the stochastic mechanism that gave rise to the particular sample realization $\mathbf{x}_0$.

Frequentist statistical inference procedures, such as estimation, testing and prediction, are based on the information summarized by $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}{\in}\mathbb{R}^n_X$. In this sense the appropriateness of these procedures depends crucially on the *validity of the assumptions* underlying the statistical model postulated. In cases where the underlying assumptions are invalid for $\mathbf{x}_0$, the inference results are likely to be unreliable because

the sampling distributions of an estimator, test statistic or predictor is determined by $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$ via (7).

Having introduced the basic concepts of the frequentist approach, we proceed to provide a bird's eye view of the main practical facets of statistical inference, traditionally known as estimation, testing and prediction, assuming the problems raised by the modeling facet have been addressed adequately.

6.3 Estimation: point and interval

▶ How could one estimate an unknown parameter θ?

Point estimation. In the context of a postulated statistical model $\mathcal{M}_\theta(\mathbf{x})=\{f(\mathbf{x};\theta),\ \theta\in\Theta\}$, $\mathbf{x}\in\mathbb{R}_X^n$, the data information comes in the form of a particular realization $\mathbf{x}_0$ of the sample $\mathbf{X}:=(X_1,X_2,...,X_n)$. Estimation pertains to the task of constructing a procedure (or a rule) that pin-points the true value θ^* of θ in Θ 'as best as possible'. This comes in the form of a mapping between the sample ($\mathcal{X}:=\mathbb{R}_X^n$) and the parameter ($\Theta$) spaces:

$$h(.):\ \mathcal{X}\to\Theta.$$

This mapping, referred to as a point *estimator* of θ, is denoted by (figure 11.4):

$$\text{Estimator:}\quad \widehat{\theta}(\mathbf{X})=h(X_1,X_2,...,X_n).$$

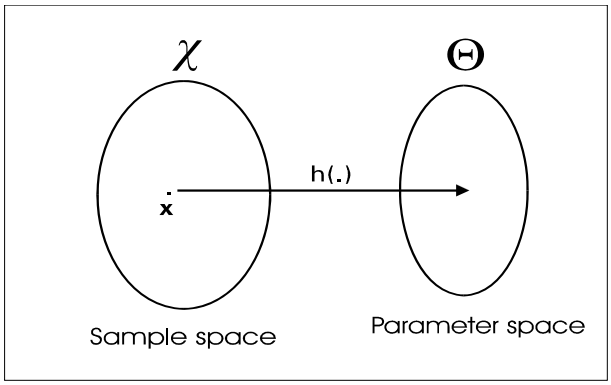


Fig. 10.4: Defining an estimator

The particular value taken by this estimator, based on the sample realization $\mathbf{X}=\mathbf{x}_0$, is referred to as a point *estimate*:

$$\text{Estimate:}\quad \widehat{\theta}(\mathbf{x}_0)=h(\mathbf{x}_0).$$

NOTE that to avoid cumbersome notation the same symbol $\widehat{\theta}$ is often used to denote both the estimator. When $\widehat{\theta}$ is used without the right hand side, the meaning should be obvious from the context.

Example 10.13. It is known (chapter 3) that in the Bernoulli model $\theta=E(X)$. This suggests that an obvious choice of a mapping as an estimator of θ is the sample mean: $\widehat{\theta}=\frac{1}{n}\sum_{i=1}^{n} X_i$.

Let us take the idea of an estimator a step further. In light of the fact that the estimator $\widehat{\theta}(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n} X_i$, is a function of the sample $\mathbf{X}$, $\widehat{\theta}(\mathbf{X})$ is a random variable itself, and thus the estimate $\widehat{\theta}(\mathbf{x}_0)$ is just one of the many (often infinitely many) values $\widehat{\theta}(\mathbf{X})$ could have taken. In the case of the above Bernoulli example for each sample realization, say $\mathbf{x}_{(i)}$, $i=1,2,...$ there is a different estimate, say:

$$\widehat{\theta}_{(1)}=.40,\ \ \widehat{\theta}_{(2)}=.43,\ \ \widehat{\theta}_{(3)}=.45,\ \ \widehat{\theta}_{(4)}=.51,\ \ \widehat{\theta}_{(5)}=0.35,$$

but all these are values of the same estimator $\widehat{\theta}(\mathbf{X})$. All possible values of $\widehat{\theta}(\mathbf{X})$, together with their corresponding probabilities are described by its sampling distribution $f(\widehat{\theta}(\mathbf{x});\theta)$, $\forall\mathbf{x}\in\mathbb{R}^n_X$ whose functional form is determined by $f(\mathbf{x};\theta)$ as in (7).

Example 10.14. For the simple Bernoulli model (table 10.2) the sampling distribution of $\widehat{\theta}(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n} X_i$ is a scaled Binomial distribution:

$$\widehat{\theta}(\mathbf{X})\backsim \text{Bin}\left(\theta,\ \tfrac{\theta(1-\theta)}{n};n\right),\forall\mathbf{x}\in\{0,1\}^n.$$

The idea is that $f(\widehat{\theta}(\mathbf{x});\theta)$, $\forall\mathbf{x}\in\{0,1\}^n$ is closely distributed around θ^*, the true θ, as possible. The various concepts associated with the optimality of an estimator will be discussed in chapter 11. Ideally, $f(\widehat{\theta}(\mathbf{x});\theta)$ assigns probability one to θ^*, i.e. it reduces to a degenerate distribution of the form $\mathbb{P}(\widehat{\theta}(\mathbf{X})=\theta^*)=1$.

Interval estimation. Another form of estimation is *interval estimation* which amounts to specifying a multi-valued (one-to-may) function whose range defines a region in Θ (see figure 11.5).

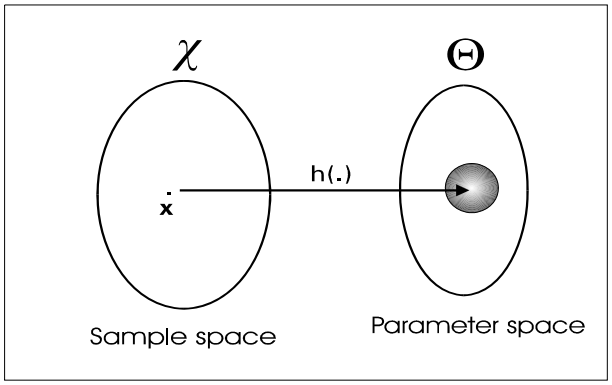


Fig. 10.5: Defining an interval estimator

Example 10.15. In the case of the simple Bernoulli model one might be interested in specifying an *interval estimator* of θ which hopefully includes the true value θ^* of θ, as much as possible. This amounts to specifying a 'random' interval:

$$[\widehat{\theta}_L(\mathbf{X}),\ \widehat{\theta}_U(\mathbf{X})], \tag{14}$$

where $\widehat{\theta}_L(\mathbf{X})$ and $\widehat{\theta}_U(\mathbf{X})$ denote the lower and upper bounds of the Confidence Interval (CI). These represent two mappings from $\mathcal{X}:=\mathbb{R}^n_X$ to Θ are often functions of an optimal point estimator of θ such that the interval in (14) covers (overlays) θ^* with a high probability, say .95, i.e.

$$\mathbb{P}(\widehat{\theta}_L(\mathbf{X}) < \theta^* \leq \widehat{\theta}_U(\mathbf{X}))=.95.$$

The above probabilistic statement asserts that in a long-run sequence of sample realizations, say $\mathbf{x}_{(i)}$, $i=1,2,...,N$, the resulting intervals defined by $[\widehat{\theta}_L(\mathbf{x}_{(i)}),\ \widehat{\theta}_U(\mathbf{x}_{(i)})]$, $i=1,2,...,N$, will to include θ^*, 95% of the time. In any one sample realization, however, it is not known whether the interval includes θ^* or not.

6.4 Hypothesis testing: a first view

Another form of statistical inference relates to testing hypotheses of interest about specific values or a range of values for θ.

Example 10.16. For the simple Bernoulli model $\theta\in\Theta:=[0,1]$. Assuming that the value of substantive interest is $\theta=.5$, one can define the null (H_0) and alternative (H_1) hypotheses to form:

$$H_0\text{: } \theta \leq .5 \text{ against } H_0\text{: } \theta > .5,$$

that partitions $\Theta:=[0,1]$ into $\Theta_0:=[0,.5]$ and $\Theta_1:=(.5,1]$. is to construct a test which enables us to decide whether the hypothesis that the true θ belongs to this subset, say $\Theta_0\subset\Theta$, is supported by the data.

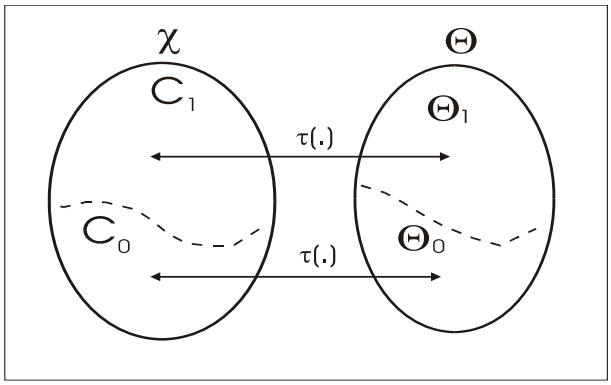


Fig. 10.6: Defining a Neyman-Pearson test

Neyman-Pearson test. A Neyman-Pearson test is defined in terms of test statistic $\tau(.)$: $\mathcal{X}\to\mathbb{R}$ that partitions the sample space into an acceptance $C_0=\{\mathbf{x}$: $\tau(\mathbf{x})\leq c\}$ and a rejection region $C_1=\{\mathbf{x}$: $\tau(\mathbf{x})>c\}$ (figure 10.6), corresponding to Θ_0 and Θ_1, respectively. When $\mathbf{x}_0\in C_0$, H_0 is accepted, and when $\mathbf{x}_0\in C_1$, H_0 is rejected.

The test statistic $\tau(\mathbf{X})$ is a random variable with its own *sampling distribution* derived via (7). This distribution can be used to calibrate its 'optimality' in terms of the error probabilities associated with rejecting H_0 when it is true, and accepting it when it is false.

6.5 Prediction (forecasting)

Prediction (or forecasting) is concerned with specifying appropriate functions of the sample $X_1, X_2, ..., X_n$ which enable us to predict beyond the data in hand such as the observation of X at $n+1$, denoted by X_{n+1}. That is, define an optimal function $q(.)$ such that:

$$\hat{X}_{n+1}=q(X_1, X_2, ..., X_n).$$

A natural choice of the function $q(.)$, which is optimal in a mean square sense is the conditional expectation of X_{n+1} given $X_1, X_2, ..., X_n$. As shown in chapter 7 the only function $q(X)$ which minimizes the mean of the squared error:

$$E\left[X_{n+1}-q(X)\right]^2,$$

is none other than the conditional expectation:

$$q(X)=E(X_{n+1}|X_1, X_2, ..., X_n).$$

Example 10.17. In the case of the Bernoulli model the obvious way to derive the predictor of X_{n+1} is to utilize the statistical GM (see chapter 7) which is indeed based on such a conditional expectation. From the statistical GM we can see that the best way to predict X_{n+1} is to extend it beyond the sample period, i.e. postulate:

$$X_{n+1}=\theta+u_{n+1}.$$

Given that θ is unknown and $E(u_{n+1})=0$ the natural predictor is: $\hat{X}_{n+1}=\widehat{\theta}(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n}X_i$. In this sense the mapping $q(.)$ is a composite mapping from $\mathcal{X}$ to Θ and then from Θ to the prediction space $(\mathcal{X}_p)$:

$$q(h(.)):\ \mathcal{X}\rightarrow\Theta\rightarrow\mathcal{X}_P.$$

From this we can see that $q(\mathbf{X})$ is also a random variable whose sampling distribution depends on that of $\widehat{\theta}$. Hence, any probabilistic statements about the accuracy of $\hat{X}_{n+1}$ will be based on the sampling distribution of $\widehat{\theta}$.

We conclude this section by re-iterating again that classical procedures of statistical inference are constructed and assessed through the sampling distributions of the relevant statistics. Moreover, the statistics used in Confidence Intervals (CI) and Hypothesis testing rely on optimal point estimators; an optimal CI and an optimal test rely on optimal point estimators.

6.6 Probability vs. frequencies: the empirical CDF

The best and most direct link between mathematical probability and the relative frequencies comes in the form of the empirical cumulative distribution function (ecdf) defined by:

$$\widehat{F}_n(x)=\frac{[\text{no. of } (x_1,x_2,....,x_n) \text{ that do not exceed } x]}{n}=\frac{1}{n}\sum_{k=1}^{n}\mathbb{I}_{(-\infty,x]}\left(X_k\right),\quad \forall x\in\mathbb{R},$$

where $\widehat{F}_n(x)$ refers to the *relative frequency* of the observations not exceeding the value x. When viewed as a function of the observations $\mathbf{x}_0$, $\widehat{F}_n(x)$ enjoys all the properties of its theoretical counterpart, the cdf $F(x)$.

Kolmogorov (1933a) posed and answered the crucial question:

▶ how well does $\widehat{F}_n(x)$ approximate the cdf $F(x)$?

Let $\{X_t,\ t\in\mathbb{N}\}$ be IID stochastic process with cdf $F(x)$.

(a) **Unbiased and consistent** (i) $E(\widehat{F}_n(x))=F(x)$, (ii) $Var(\widehat{F}_n(x))=\frac{F(x)[1-F(x)]}{n}$, $\forall x\in\mathbb{R}$.

(i)-(ii) imply that $\lim_{n\to\infty} Var(\widehat{F}_n(x))=0$, and thus $\widehat{F}_n(x) \xrightarrow{\mathbb{P}} F(x)$.

(b) **Kolmogorov's distance theorem**. For the Kolmogorov distance:

$$\Delta_n:=\sup_{x\in\mathbb{R}}|\widehat{F}_n(x)-F(x)|, \text{ for any } \textit{continuous } F(x):$$

$$\lim_{n\to\infty}\mathbb{P}(\sqrt{n}\Delta_n\leq z)=1-2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2z^2}, \tag{15}$$

where the *convergence is uniform* in $z>0$.

(c) **Dvoresky-Kiefer-Wolfowitz.** Upper bounds for its tail area probabilities:

$$\mathbb{P}(\Delta_n>\varepsilon)\leq 2e^{-2n\varepsilon^2}, \text{ for any } \varepsilon>0. \tag{16}$$

In addition, the SLLN holds for Δ_n.

(d) **Glivenko-Canteli theorem**. Let $\{X_t,\ t\in\mathbb{N}\}$ be an IID stochastic process with cdf $F(x)$. Then: $\mathbb{P}(\lim_{n\to\infty}\Delta_n=0)=1$ i.e. $\widehat{F}_n(x) \xrightarrow{a.s} F(x)$, uniformly in $x\in\mathbb{R}$.

A moment's reflection suggests that the above results indicate that $\widehat{F}_n(x)$ can be viewed as the bridge between the *relative frequencies* of stochastic phenomena and the probabilities stemming from $F(x)$. What renders this result important is that the assumptions that give rise to [a]-[d] are testable, IID, whose validity bestows empirical content to the cdf.

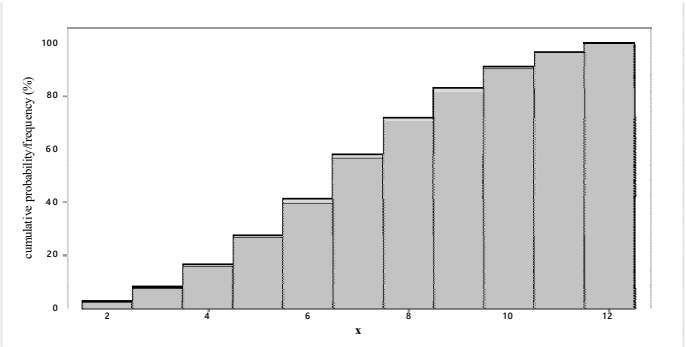


Fig. 10.7: The cdf vs. ecdf for the dice data

Example 10.18. Let us return to the dice data in table 1.1 and figure 1.1; chapter 1. In figure 1.5 it was conjectured that the data histogram is very close to

the probability distribution. As argued in chapter 5, however, the histogram $\widehat{f}_n(x)$ is an inconsistent estimator of $f(x)$, but a consistent estimator for $f(x)$, such as the kernel estimator, can be viewed as deduced indirectly using $\widehat{F}_n(x)$. A more sound confirmation of that is given in figure 10.7 where the discrepancies between the cdf and the ecdf are shown to be minor.

6.6.1 From the ecdf to estimators and tests

In the context of a statistical model, $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the parameters $\boldsymbol{\theta}$ are related to the raw moments via:

$$\mu'_r(\boldsymbol{\theta})=\int_{x\in\mathbb{R}_X} x^r dF(x;\boldsymbol{\theta}).$$

NOTE that in the case where the density function $f(x;\boldsymbol{\theta})$ is continuous:

$$dF(x)=f(x)dx \ \ \rightarrow \mu'_r(\boldsymbol{\theta})=\int_{x\in\mathbb{R}_X} x^r f(x;\boldsymbol{\theta})dx, \ \ x\in\mathbb{R}_X.$$

In this sense the unknown parameter(s) $\boldsymbol{\theta}$ can be viewed as a functions of $F(x;\boldsymbol{\theta})$: $\boldsymbol{\theta}=\mathbf{g}(F)$, assuming that $\mathbf{g}(.)$ is a Borel function. In this sense, an estimator of $\boldsymbol{\theta}$ can be viewed as a function of the the *empirical cdf* (ecdf): $\widehat{\boldsymbol{\theta}}(\mathbf{X})=\mathbf{g}(\widehat{F}_n)$,

where the ecdf can be defined by: $\widehat{F}_n(x)=\frac{\#\{x_k\leq x\}}{n}=\frac{1}{n}\sum_{k=1}^{n} H\left(x-x_k\right), \quad x\in\mathbb{R}$,

with $H(.)$ being the Heaviside function: $H(u)=\begin{cases} 0, & \text{if } u<0 \\ 1, & \text{if } u\geq 0 \end{cases}$

Example 10.19. Consider the case where the parameter of interest is $E(X)$:

$$E(X):=\mu=\int_{x\in\mathbb{R}_X} xdF(x).$$

One can estimate μ by replacing $F(x)$ with $\widehat{F}_n(x)$, giving rise to the sample mean since:

$$\begin{aligned}\overline{X}_n=\int_{x\in\mathbb{R}_X} xd\widehat{F}_n(x) \ &= \int_{x\in\mathbb{R}_X} xd\left(\tfrac{1}{n}\sum_{k=1}^{n} H\left(x-x_k\right)\right)= \\ &=\frac{1}{n}\sum_{k=1}^{n}\int_{x\in\mathbb{R}_X} xdH\left(x-x_k\right)=\frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned}$$

In light of the above results associated with the ecdf $\widehat{F}_n(x)$, it is clear that the latter provides the best link between the mathematical world of probability and distribution functions and the real-world of data and frequencies, in complete agreement with the frequentist interpretation of probability.

7 Nonparametric inference

7.1 Parametric vs. nonparametric inference

A large part of current applied research in most fields is based on neither the frequentist nor the Bayesian approach. Instead, it is based on what is misleadingly described

as nonparametric modeling and inference. Practitioners are often motivated to use nonparametric inference because of certain egregiously misinformed assertions that are often taken at face value:

(a) 'The crucial difference between parametric and nonparametric inference is that the former relies on rather restrictive probabilistic assumptions but the latter does not'.

(b) 'When a modeler has a large enough sample size n data, there is no need to make any probabilistic assumptions about what generated the data because one can use the empirical cumulative distribution function (ecdf) $\widehat{F}_n(x)$ to provide a sound basis for inference.'

(c) 'Even when a modeler does not have a large enough sample size n, one can resample the data themselves to derive empirically based inferences that are more reliable than those based on parametric inference.'

The truth is that nonparametric inference differs from parametric inference in one crucial respect: the prespecified statistical model for the latter assumes a direct distribution assumption that parameterizes:$\mathcal{M}_\theta(\mathbf{x})=f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X$, completely, but that of the former, say $\mathcal{M}_{F}(\mathbf{x})$, replaces the distribution assumption with an unknown family of distributions F, assumed to satisfy certain indirect distributional assumptions that include: (i) the existence of certain moments up to order p, as well as (ii) smoothness restrictions on the unknown density function $f(z)$, $z\in\mathbb{R}_Z$ (bounded support, continuity, symmetry, differentiability, unimodality, boundedness and continuity of derivatives of $f(z)$ up to order m=3, except at the endpoints, etc.); see Thompson and Tapia (1990), Wasserman (2006). It is important, however, to emphasize that both $\mathcal{M}_\theta(\mathbf{x})$ and $\mathcal{M}_{F}(\mathbf{x})$ impose direct dependence and heterogeneity probabilistic assumptions.

It is argued that by imposing sufficient untestable smoothness restrictions on $f(z)$, presented as harmless mathematical restrictions imposed for convenience, is the surest way to untrustworthy evidence. This is often justified on the basis of instinctive impressions that 'weaker assumptions are less vulnerable to misspecification, and give rise more robust inferences since they rely on asymptotic approximations that render misspecification less pernicious as $n\to\infty$. That is, the nonparametric perspective encourages practitioners to adopt a combination of generic robustness stemming from broader inductive premises and asymptotic procedures as a way to alleviate the effects of potential statistical misspecification. What is not so obvious is whether a distribution-free $\mathcal{M}_{F}(\mathbf{x})$ provides 'insurance' against departures from a direct distribution assumption. The price of such insurance is often less precise inferences because broader inductive premises give rise to less precise inferences. In what follows a case is made that weak assumptions can be as fallible as strong ones, and the best way to guard against statistical misspecification is the testability of the model assumptions.

Buying insurance? An example of buying 'insurance' against departures from Normality can be found in the statistics literature relating to Wilcoxon-type tests vs. the t-test in the context of simple statistical models (IID) under various non-Normal

distributions; see Hettmansperger (1984). What is often insufficiently appreciated by this literature is that the various comparisons of 'asymptotic efficiency' (Bahadur, 1960) between nonparametric tests and the t-test are often dubious in both value and substance. For instance, comparing a Wilcoxon-type test with the t-test in the case where the true distribution is *Cauchy* is absurd since the t-test is based on the sample mean and variance neither of which exist in the case of the Cauchy distribution. Also, in the case of the *Uniform* distribution the relevant comparison for a nonparametric test should not be the t-test because there is an optimal parametric test, an F-type test given by Neyman and Pearson (1933). The only way to make sense of such comparisons is when one criticizes the 'Normality of bust' strategy in empirical modeling, which is clearly a derisory approach; see Spanos (2002).

More importantly, unlike natural hazards, in statistical modeling one has additional information in the form of $\mathbf{x}_0$ that can be used to evaluate potential departures from these assumptions. In the case of non-Normality, one can easily use the data to distinguish between realizations of IID processes from different families of distributions with sufficient confidence using simple t-plots. That is, buying insurance for a distribution assumption is superfluous because evaluating their validity in the case of an IID sample realization $\mathbf{x}_0$ is as trivial as glancing at a t-plot and a smoothed histogram; see chapter 5. What guards an inference from the hazards of statistical misspecification is the *testability* of the assumptions.

The real difference. The crucial difference between parametric and nonparametric inference is that the latter replaces the easily testable distribution assumptions with untestable indirect distribution assumptions, but relies on highly restrictive distribution and heterogeneity assumptions to retain the tractability of the sampling distributions of the relevant statistics. What guards the reliability of inference against misspecification is not buying insurance for easily detectable departures, but the objective testability of these assumptions. As argued by Fisher (1922a):

"For empirical as the specification of the hypothetical population [statistical model] may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts." (p. 314).

Weaker probabilistic assumptions, such as replacing Normality with the existence of the first two moments, usually imply broader but partly non-testable inductive premises, forsaking the possibility to secure the reliability of inference.

7.2 Are weaker assumptions preferable to stronger ones?

On closer examination, these instinctive impressions about the robustness of nonparametric inference turn out to be highly misleading. What the above assertions vaunting nonparametric inference seem to ignore is that a statistical model involves assumptions from all tree categories: distribution, dependence and heterogeneity. Moreover, departures from distribution assumptions are the least problematic from

the statistical misspecification perspective, since their effects on the reliability of inference are much less pernicious and decrease with the sample size n. This is not true for departures from the dependence and heterogeneity assumptions. For instance, the presence of heterogeneity in one's data usually worsens the reliability of inference as n increases; see Spanos and McGuirk (2001). What is ironic is that nonparametric inference often relies on the highly restrictive IID assumptions, as in the case of nonparametric inference based on the ecdf $\widehat{F}_n(x)$, order statistics and ranks, as well as kernel smoothing techniques; see Lehmann (1975), Wasserman (2006). If the motivation is to buy insurance against departures from probabilistic assumptions, it makes a lot more sense to do that for the IID assumptions.

Weaker probabilistic assumptions, such as replacing Normality with the existence of the first two moments, usually imply broader but partly non-testable assumptions, forsaking the possibility to secure the reliability of inference.

Example 10.20. This very point is illustrated in Bahadur and Savage (1956) who replaced the Normality assumption in the simple Normal model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: } X_k \backsim \mathsf{NIID}(\mu,\sigma^2),\ x_k\in\mathbb{R},\ \mu\in\mathbb{R},\ \sigma^2>0,\ k\in\mathbb{N}, \tag{17}$$

with a broader family of distributions F defined by the existence of the first two moments:

$$\mathcal{M}_{\mathit{F}}(\mathbf{x})\text{: } X_k \backsim \mathsf{IID}(\mu,\sigma^2),\ x_k\in\mathbb{R},\ k=1,2,...,n. \tag{18}$$

Then, they posed the question whether there is a reasonably reliable test within the family F for testing the hypotheses, H_0: $\mu=0$, vs. H_1: $\mu\neq 0$, analogous to the t-test based on (chapter 13):

$$d(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-0)}{s^2},\ C_1=\{\mathbf{x}\text{: } |d(\mathbf{x})|>c_\alpha\}\quad \overline{X}_n=\tfrac{1}{n}\sum_{k=1}^{n}X_k,\ s^2=\tfrac{1}{n-1}\sum_{k=1}^{n}\left(\overline{X}_n-X_k\right)^2. \tag{19}$$

The surprising answer is that no such test exists. Any t-type test based on F will be *biased* and *inconsistent* (Bahadur and Savage, 1956):

"It is shown that there is neither an effective test of the hypothesis that $\mu=0$, nor an effective confidence interval for μ, nor an effective point estimate of μ. These conclusions concerning μ flow from the fact that μ is sensitive to the tails of the population distribution; parallel conclusions hold for other sensitive parameters, and they can be established by the same methods as are here used for μ." (p. 1115).

The intuitive reason for this result is that the family F defined by the existence of its first two (or even all) moments is too broad to enable one to 'tame the tails' of the sampling distribution of $\overline{X}_n$ sufficiently to be able to evaluate relevant error probabilities associated with a t-type test. One can easily imagine density functions with raging shapes and untamed tails even when X_k takes values within a finite range $[a,b]$ so that all moments exist. In the case of a given n there is no reason to believe that the tails of $\overline{X}_n$ are tame enough to evaluate the relevant error (type I and II) probabilities. The existence of certain moments does not provide enough structure

for the density function to render the tail areas sufficiently smooth to evaluate with any accuracy the error probabilities.

In light of the limit theorems (WLLN, SLLN, CLT) in chapter 9, one might ask: (a) why Bahadur and Savage (1956) did not invoke the CLT to claim the following asymptotic result for their t-type test:

$$d(\mathbf{X}) \overset{\mu=\mu_0}{\underset{n\to\infty}{\sim}} \mathsf{N}(0,1)\,. \tag{20}$$

(a) why they did not invoke the SLLN to claim that $\overline{X}_n$ is a strongly consistent for μ, and thus "an effective point estimator".

The answer to (a) is that all limit theorems (WLLN, SLLN, CLT) hold at the limit $n{=}\infty$, and not for a given $n<\infty$. The best one can do in practice is to use the LIL for the LLN or the Berry-Esseen type upper bounds to evaluate how good is the approximation of the sampling distribution of $\overline{X}_n$ in (19) for the particular n, σ^2 and $\mu_3{=}E\,|X_k{-}\mu|^3$; see chapter 9. Hence, the claim that the approximation will be reasonably good irrespective of the true distribution of X_k within the family F is unwarranted. The fact of the matter is that invoking (20) is tantamount to imposing approximate Normality for the sampling distribution, which will be an accurate enough claim, when the log-likelihood can be approximated by a quadratic function of $\boldsymbol{\theta}{:=}(\mu,\sigma^2)$ [like the Normal]; see Geyer (2013). Believing that the validity of (20) stems from invoking the heuristic 'as $n\to\infty$' is just an illusion. The trustworthiness of any inference results invoking (20) stems solely from the approximate validity of the probabilistic assumptions imposed on $\mathbf{x}_0$ for the specific n.

The answer to (b) is that consistency is a *minimal* property of an estimator, i.e. it is a necessary but not a sufficient property for an effective (optimal) point estimator; see chapter 11.

The claim that broader inductive premises are less vulnerable to misspecification is analogous to the claim that a bigger fishing net will always produce a larger catch. This, however, ignores the fact that the size of the catch depends also on the pertinency of the casting location. In the case of a fully parametric model with testable assumptions, misspecification testing can guide the modeler toward a more appropriate 'location'; this adeptness is missing from a weak but non-testable set of probabilistic assumptions.

To illustrate the how the traditional approach to modeling favors weak probabilistic assumptions in conjunction with limit theorems, at the expense of the reliability of inference, consider the following example.

Example 10.21. A typical example of a traditional weak set of probabilistic assumptions for the AR(1) model, used as a basis for unit root testing, specified in a path breaking paper by Phillips (1987), is given in table 10.10. How would a practitioner decide that the probabilistic assumptions (i)-(iv) are appropriate for a data set $\mathbf{y}_0$? In practice, modelers will just take such assumptions at face value because they are not testable and *hope* that their asymptotic results will not be too

unreliable; their hope is unfounded.

Table 10.10: AR(1) model (Phillips, 1987)

	$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t, \quad t \in \mathbb{N} := (1, , 2, ..., n, ...)$
[i]	$E(u_t)=0$, for all $t \in \mathbb{N}$,
[ii]	$\sup_t E\|u_t\|^{\delta+\varepsilon} < 0$ for $\delta > 2, \varepsilon > 0$,
[iii]	$\lim_{n\to\infty} E(\frac{1}{n}(\sum_{t=1}^{n} u_t)^2) = \sigma^2_\infty > 0$,
[iv]	$\{u_t,\ t \in \mathbb{N}\}$ is *strongly mixing* with mixing coefficient $\alpha_m \underset{m\to\infty}{\to} 0$ such that $\sum_{m=1}^{\infty} \alpha_m^{1-\delta/2} < \infty$.

In contrast, table 10.11 gives a complete set of testable probabilistic assumptions for the same AR(1) model. It turns out that in practice the inference results invoking the non-testable assumptions [i]-[iv] will be reliable only to the extent that the testable assumptions [1]-[5] are approximately valid for the particular data $\mathbf{y}_0$.

Table 10.11: Normal, AutoRegressive [AR(1)] model

Statistical GM:		$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t, \quad t \in \mathbb{N}.$	
[1]	Normality:	$(y_t, y_{t-1}) \backsim \mathsf{N}(.,.)$,	} $t \in \mathbb{N}$.
[2]	Linearity:	$E\left(y_t \| \sigma(y_{t-1})\right) = \alpha_0 + \alpha_1 y_{t-1}$,	
[3]	Homoskedasticity:	$Var\left(y_t \| \sigma(y_{t-1})\right) = \sigma_0^2$,	
[4]	Markov:	$\{y_t,\ t \in \mathbb{N}\}$ is a Markov process,	
[5]	t-invariance:	$(\alpha_0, \alpha_1, \sigma_0^2)$ are *not* changing with t,	

$\alpha_0 = E(y_t) - \alpha_1 E(y_{t-1}) \in \mathbb{R}$, $\alpha_1 = \frac{Cov(y_t, y_{t-1})}{Var(y_{t-1})} \in (-1, 1)$, $\sigma_0^2 = Var(y_t) - \frac{Cov(y_t, y_{t-1})^2}{Var(y_{t-1})} \in \mathbb{R}_+$

Of course, practitioners invoking [i]-[iv] would not know that unless they test assumptions [1]-[5]! It is often conveniently forgotten, butlimit theorems invoked by Consistent and Asymptotically Normal (CAN) estimators and associated tests, also rely on probabilistic assumptions, such as (i)-(iv) above, which are usually non-testable, rendering the reliability of the resulting inferences dubious. Such asymptotic results tell us nothing about the trustworthiness of the inference results based on data $\mathbf{z}_0 := (z_1, ..., z_n)$. The latter is inextricably bound up with the particular $\mathbf{z}_0$ and n; see Spanos (2018).

7.3 Induction vs. deduction

At a more subtle level, the conventional wisdom favoring weaker assumptions is burdened with a fundamental confusion between *mathematical deduction* and *statistical induction*. Mathematical deduction puts a premium on results based on the weakest (minimal) set of assumptions comprising the deductive premises. A deductive

argument is logically 'valid' if it is impossible for its premises to be true while its conclusion is false. The logical validity of a deductive argument, however, does not depend on the 'soundness' of the premises. In statistical induction the empirical validity (soundness) of the premises is of paramount importance because the deductive argument (the inference procedure) will uphold that validity to secure the reliability of the inference procedures. Statistical induction puts the premium on the strongest (maximal) set of *testable* assumptions comprising the inductive premises. When validated vis-a-vis the data, such a maximal set provides the most effective way to learn from data because the inference procedures stemming from it are both reliable and optimal (precise).

In this sense, the weakest link when using CAN estimators and related inference procedures in practice is the conjuration of limit theorems which rely on mathematically convenient but empirically non-testable assumptions. Despite their value as purely deductive propositions, such limit theorems are insidious for modeling purposes when their assumptions turn out to be invalid for one's data.

In several papers in the 1940s and 1950s Fisher railed against "mathematicians" who view statistics as a purely deductive field, by ignoring its inductive roots stemming from the link between the probabilistic assumptions and the actual data; see Box (1978), pp. 435-8. Current textbook econometrics is sometimes taught as a purely deductive field. Their use of empirical illustrations aims primarily to demonstrate how to carry out the calculations. Their value as exemplars of empirical modeling that gives rise to learning from data about phenomena of interest is questionable because they are invariably based on non-validated inductive premises.

7.4 Revisiting generic robustness claims

Box (1953) defined robustness to refer to the sensitivity of inference procedures (estimators, tests, predictors) to departures from the model assumptions. According to Box, a procedure is said to be robust against certain departure(s) from the model assumptions when the inference is not *very sensitive* to the presence of *modest departures* from the premises; some assumptions 'do not hold, to a greater or lesser extent'. Since the premises of inference are never exactly 'true', it seems only reasonable that one should evaluate the sensitivity of the inference method to modest departures. When reading the above passage one is struck by the vagueness of the various qualifications concerning 'modest departures', 'degrees of insensitivity' and assumptions holding 'to a greater or lesser extent'. The problem is that establishing the degree of 'insensitivity' that renders the reliability of an inference procedure 'tolerable' under specific departures from the model assumptions is a very difficult task.

What is often insufficiently appreciated in practice is that departures from model assumptions can take an infinite number of forms, but there are no robustness results for generic departures, such as:

$$Corr(X_i, X_j)\neq 0, \text{ for all } i\neq j,\ i,j=1,...n. \tag{21}$$

This is because one cannot evaluate the potential discrepancy between nominal and actual error probabilities under generic departures such as (21). Worse, this discrepancy in the case of the t-test based on (19) will be very different, depending, not only on the particular form of dependence, say:

$$\text{(i) Exchangeable:} \quad Corr(X_i, X_j)=\rho,\ 0<\rho<1,\ \text{for all } i\neq j,\ i,j=1,...n,$$

$$\text{(ii) Markov:} \quad Corr(X_i, X_j)=\rho^{|i-j|},\quad -1<\rho<1,\ i\neq j,\ i,j=1,...n,$$

but also on the magnitude of ρ; see Spanos (2009). This implies that before one can provide a reasonable evaluation of this discrepancy one needs to establish the appropriateness of the specific departures for the particular data to demonstrate their potential relevance. The problem is that if one needs to establish the particular form of dependence appropriate for one's data, the whole robustness line of reasoning is undermined. This is because it becomes pointless to return to the original (misspecified) model if one were able to reach a (statistically adequate) model after respecification.

7.5 Inference based on asymptotic bounds

Le Cam (1986a) pointed out a serious flaw in relying on asymptotic results for inference:

"... limit theorems 'as n tends to infinity' are logically devoid of content about what happens at any particular n. All they can do is suggest certain approaches whose performance must then be checked on the case at hand. Unfortunately the approximation bounds we could get were too often too crude and cumbersome to be of any practical use." (p. xiv).

One of the most crucial features of the development of limit theorems since 1713 has been the continuing weakening of the premises – the invoked probabilistic assumptions on $\{X_k,\ k\in\mathbb{N}\}$ – giving rise to the various inferential conclusions; see chapter 9. The most important such weakening comes in the form of the distribution assumption, that has been, almost immediately, replaced by indirect distributional assumptions pertaining to the existence of moments; see chapter 3. Historically, it is interesting to note that the initial proof of the WLLN by Bernoulli (1713) was based on the finite sample distribution of $\overline{X}_n=\frac{1}{n}\sum_{k=1}^{n} X_k$ under the Bernoulli, IID assumptions:

$$\overline{X}_n \backsim \mathsf{Bin}\left(\theta,\ \theta(1-\theta); n\right). \tag{22}$$

That is, instead of upper bounds offered later by inequalities, such as Chebyshev's, Markov's and Bernstein's (Appendix 9A), Bernoulli used bounds based on the Binomial distribution in (22). When advocates of nonparametric modeling and inference extol the virtues of such an approach to modeling, they often claim weaker assumptions without qualifying that by admitting a great loss of precision in replacing a direct distribution assumption with an indirect mathematical conditions that are often non-testable one. In an attempt to bring more balance to such claims let us consider what happens to the precision of inference when that strategy is followed.

Example 10.22. For a Bernoulli, IID stochastic process $\{X_k,\ k\in\mathbb{N}\}$, let us consider the precision of inference in the case of evaluating tail areas of the form $\left(|\overline{X}_n-\theta|\geq\varepsilon\right)$ or equivalently $\left(|\overline{X}_n-\theta|<\varepsilon\right)$ under two different scenarios, when using [a] the finite sample distribution, and [b] an upper bound based on inequalities.

[a] Using an upper bound. Using the fact that $Var(X)=\theta(1-\theta)<\frac{1}{4}$ in conjunction with Chebyshev's inequality for n=1000 and $\varepsilon = .1$, we can deduce that:

$$\mathbb{P}\left(|\overline{X}_n-\theta|>\varepsilon\right)\leq\tfrac{1}{4(1000)(.1)^2}=\tfrac{1}{40}=.025\Leftrightarrow\mathbb{P}\left(|\overline{X}_n-\theta|\leq\varepsilon\right)\geq .975. \tag{23}$$

[b] Using the Normal approximation (22). As shown in section 2, the Binomial in (22), when standardized $Z_n=\frac{\overline{X}_n-n\theta}{\sqrt{n\theta(1-\theta)}}$ can be accurately approximated by the Normal for values of θ around .5. This means that the Normal approximation can achieve the same precision for the tail area in (23) based on n=1000, with only n=126 since $\frac{\varepsilon}{\sqrt{Var(\overline{X}_n)}}=\frac{.1}{\sqrt{\frac{.25}{126}}}=2.245$ and thus:

$$\mathbb{P}\left(|\tfrac{1}{n}\sum_{k=1}^{n}X_k-\theta|>.1\right)\leq 1-\int_{-2.245}^{2.245}(\tfrac{\exp(-.5x^2)}{\sqrt{2\pi}})dx=.025.$$

Example 10.23. To bring out the crudeness of the bound given by Chebyshev's inequality, consider increasing the bound from .025 to .0125, i.e. $\mathbb{P}\left(|\overline{X}_n-\theta|>.1\right)\leq .0125$, by increasing n.

[a] Using an upper bound. The upper bound suggests that this requires n=2000 since:

$$\tfrac{1}{4n(.1)^2}=.0125\Rightarrow n=2000. \tag{24}$$

[b] Using the Normal approximation (22). Since $\frac{\varepsilon}{\sqrt{Var(\overline{X}_n)}}=\frac{.1}{\sqrt{\frac{.25}{156}}}=2.498$, one would achieve the same precision as the bound in (24) based on 1000 additional observations, with only 30 more (n=156):

$$\mathbb{P}\left(|\tfrac{1}{n}\sum_{k=1}^{n}X_k-\theta|>.1\right)\leq 1-\int_{-2.498}^{2.498}(\tfrac{\exp(-.5x^2)}{\sqrt{2\pi}})dx=.025.$$

An advocate of nonparametric inference might object to the above example, by arguing that Chebyshev's inequality is not the sharpest to be used in this case. For instance, *Hoeffding's inequality* is likely to yield a much better upper bound.

Example 10.24. For the above Bernoulli, IID case, let ε=.1 and n=200.

Using Chebyshev's inequality yields: $\mathbb{P}\left(|\overline{X}_n-\theta|>.1\right)\leq\frac{1}{4(200)(.1)^2}=.125$.

On the other hand, Hoeffding's inequality $\mathbb{P}\left(|\overline{X}_n-\theta|>\varepsilon\right)\leq 2e^{-2n\varepsilon^2}$ (Appendix 9.A) yields a better upper bound: $\mathbb{P}\left(|\overline{X}_n-\theta|>.1\right)\leq 2e^{-2(200)(.1)^2}=.037$.

The problem with the above argument is twofold. First, the true tail area probability is considerably smaller than both: $\mathbb{P}\left(|\overline{X}_n-\theta|>.1\right)=.00468$.

Second, Hoeffding's is not always better than Chebyshev's. For ε=.1 and n=100:

$$\text{Hoeffding: } \mathbb{P}\left(|\overline{X}_n-\theta|>\varepsilon\right)\leq .271, \quad \text{Chebyshev: } \mathbb{P}\left(|\overline{X}_n-\theta|>\varepsilon\right)\leq .25.$$

Although the Bernoulli IID assumptions constitute a very special case, the lesson to be drawn from the above example is more general: using the finite sample distribution of $\overline{X}_n$ or even a good approximation of that, will always give rise to much more precise inferences. In the above case it is absurd to ignore the fact that $\overline{X}_n$ has a sampling distribution that is Binomial when there are only two outcomes because by definition the underlying distribution is Bernoulli.

The counter-argument advocates of nonparametric modeling and inference might deploy is that in practice we have no idea what the underlying distribution is, and thus such finite sampling distributions are not available. This argument ignores the fact that distribution assumptions are testable vis-a-vis the data, and such tests will be of great value even in cases where on uses nonparametric techniques because it can shed light on the appropriateness of the asymptotic results invoked as a basis of inference; see chapter 5. Worse, the consequences of misspecified dependence and heterogeneity assumptions are considerably more pernicious, even though nonparametric techniques often rely heavily on IID assumptions without testing their validity. Making explicit distribution assumptions and testing their validity provides a much better strategy for learning from data in practice.

7.6 Whither nonparametric modeling?

The question that naturally arises at this stage is whether there is a role for non-parametric inference in empirical modeling. The answer is unquestionably yes, but not for inference purposes. Nonparametric modeling is of paramount importance for the *modeling* facet: specification, estimation, M-S testing and respecification.

At the *specification stage* non-parametric techniques, such as kernel smoothing, can be invaluable as an integral part of the preliminary (exploratory) data analysis. 'Reading' data plots, such as t-plots and scatter plots constitutes a qualitative assessment. A number of non-parametric tests based on order statistics and ranks can be used to quantify the above qualitative inductive argument. P-P and Q-Q plots, based on order statistics, can provide particularly powerful ways to discern the chance regularity patterns in the data pertaining to the underlying distribution. Similarly, smoothing techniques can be used to reduce the data-specificity of data plots such as histograms and provide the modeler with heedful ideas about appropriate parametric models.

At the *M-S testing stage* the aim is to assess the empirical adequacy of the postulated parametric model by probing the validity of the model assumptions. The modeler has to assess null hypotheses whose negation (the implicit alternative) is by its very nature non-parametric (cannot be specified in terms of the parameters of the model under consideration). In this context non-parametric techniques are

invaluable because they provide general forms of alternative hypotheses. In addition, non-parametric M-S tests are useful supplements to parametric tests because they depend on different assumptions.

8 The basic bootstrap method

The basic bootstrap procedure was proposed by Efron (1979) who recognized that one could generate data to derive an empirical sampling distribution of any statistic of interest by 'resampling' the original data:

$$\mathbf{x}_0 := (x_1, x_2, ..., x_n),$$

with a view to construct several realizations of the sample which preserve the probabilistic structure of $\mathbf{x}_0$. A single *bootstrap sample* realization:

$$\mathbf{x}^* := (x_1^*, x_2^*, ..., x_m^*), \ m \leq n,$$

is constructed by selecting the values $(x_1^*, x_2^*, ..., x_m^*)$ using *simple random sampling with replacement* from $\mathbf{x}_0$. The bootstrap realization values constitute an *unordered* set of m observed values selected one at a time (with replacement) from $\mathbf{x}_0$, and thus repetitions of the same value x_k are possible. In practice one would generate N such bootstrap realizations of size $m \leq n$ whose the number is:

$$\binom{n+m-1}{m} = \frac{(n+m-1)!}{m!(n-1)!}$$

For instance, when n=10 and we want to construct bootstrap realizations of size n, the number of possible samples is very large:

$$\binom{2n-1}{n} = \binom{19}{10} = \frac{(19)!}{(10!)(10!)} = 92378.$$

In practice, one could generate $N \leq \binom{n+m-1}{m}$ bootstrap realizations of size m:

$$[\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_N^*],$$

and use them to learn about various features of the sampling distribution of a *statistic*, say: $Y_n = h(X_1, X_2, ..., X_n)$. The values of the statistic $h(\mathbf{X})$:

$$h(\mathbf{x}_1^*) \quad h(\mathbf{x}_2^*) \quad h(\mathbf{x}_3^*) \quad \cdots \quad h(\mathbf{x}_N^*),$$

for a large enough N can be used to approximate its *sampling distribution* and its empirical *moments* of interest; see fig. 10.8. The formal justification of the bootstrap procedure stems from the results on the ecdf in section 6.6.

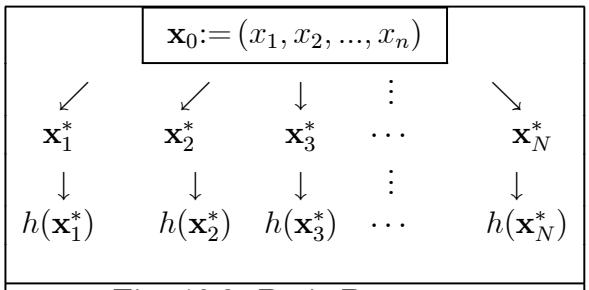


Fig. 10.8: Basic Bootstrap

Example 10.25. Consider the case of a simple statistical model in (18) could learn about the sampling distribution of $\overline{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$ by evaluating this for each of the N bootstrap realizations to generate the sample means:

$$[\overline{x}_1^*, \overline{x}_2^*, ..., \overline{x}_N^*]\,,$$

where $\overline{x}_i^*=\frac{1}{N}\sum_{i=1}^n x_i^*$ is the arithmetic average of $\mathbf{x}_i^*$, $i=1,2,...,N$. For a large enough N one can approximate the sampling distribution or any moment of $\overline{X}_n$. For instance, one could approximate the variance of $\overline{X}_n$ using:

$$Var_B(\overline{X}_n)\simeq\tfrac{1}{N-1}\sum_{i=1}^n(\overline{x}_i^*-\overline{x}^*)^2,\ \overline{x}^*=\tfrac{1}{N}\sum_{i=1}^n\overline{x}_i^*$$

The great advantage of this approximation is that one can apply it to any statistic of the original sample $Y_n=h\,(X_1, X_2, ..., X_n)$, (not just the sample mean) and approximate its variance via:

$$Var_B(Y_n)\simeq\tfrac{1}{N-1}\sum_{i=1}^n(y_i^*-\overline{y}^*)^2,\ \overline{y}^*=\tfrac{1}{N}\sum_{i=1}^n\overline{y}_i^*,$$

where $y_i^*=h\,(\mathbf{x}_i^*)$, $i=1,2,...,N$. By the same token one can construct a bootstrap approximation to any moment or function of the sampling distribution, including the density function itself.

8.1 Bootstrapping and statistical adequacy

The question that naturally arises at this stage is that the bootstrap seems to be too good to be true. The catch is that the bootstrap is a lot more vulnerable to statistical misspecification than any other procedure, because it *overexploits* the IID assumptions. When the IID assumptions are invalid for the original data $\mathbf{x}_0$ the above derivations will be highly unreliable.

Example 10.26. To illustrate what can go wrong, consider the original data with n=7 (table 10.12), together with its ordering and three bootstrap resamples of

the same size.

Table 10.12: Bootstrap resampling

t	$\mathbf{x}_0$	$\mathbf{x}_1^*$	$\mathbf{x}_2^*$	$\mathbf{x}_3^*$
1	1.13700	1.19121	1.20153	1.15740
2	1.15740	1.15740	1.17543	1.22341
3	1.17543	1.13700	1.13700	1.13700
4	1.19121	1.22341	1.22341	1.20153
5	1.20153	1.17543	1.17543	1.17543
6	1.21654	1.21654	1.13700	1.21654
7	1.22341	1.13700	1.21654	1.17543

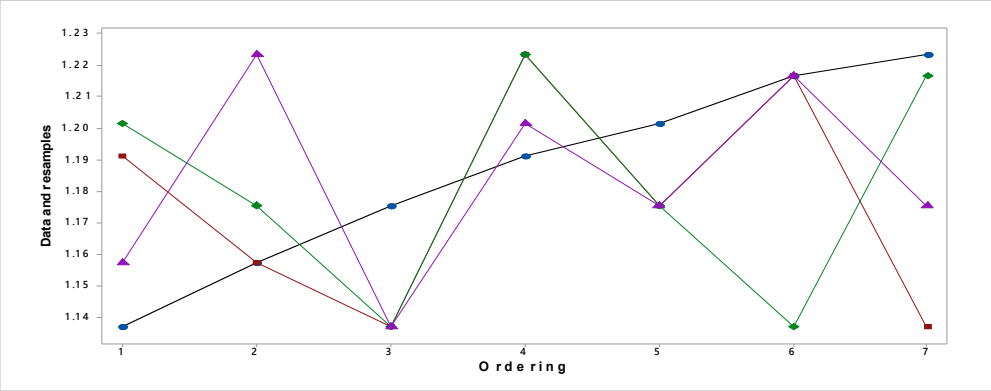


Fig. 10.9: t-plot of $\mathbf{x}_0$ (solid) and three resamples (dashes)

The t-plots of $\mathbf{x}_0, \mathbf{x}_1^*, \mathbf{x}_2^*, \mathbf{x}_3^*$ in fig. 10.9 reveal that $\mathbf{x}_0$ exhibits a distinct upward trend $[E(X_t)=1.12+.023t-.001t^2]$, but the bootstrap data appear to have a constant mean, which represents a serious distortion of the probabilistic structure of the original data $\mathbf{x}_0$. Hence, the bootstrap approximations are likely to give rise to highly misleading inferences.

In bootstrapping and other forms of resampling the primary goal is to generate artificial data that constitute faithful replicas of the original data $\mathbf{x}_0$, i.e. the artificial data $\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_N^*$ should have the same probabilistic structure as $\mathbf{x}_0$. When that is not the case, the artificial data will give rise to unreliable inferences concerning the stochastic mechanism that gave rise to $\mathbf{x}_0$. A simple way to assess whether the bootstrap-based artificial data constitute faithful replicas of the original data is to test whether the difference between simulated and actual data is 'non-systematic noise':

$$\mathbf{v}_i=(\mathbf{x}_0-\mathbf{x}_i^*) \backsim \text{IID}(0,\sigma^2),\ \ i=1,2,3$$

A t-plot of $(\mathbf{v}_{it},\ i=1,2,3,\ t=1,2,..,7)$ will show that they all exhibit trending means (verify!). Ignoring that would give rise to a misleading bootstrap standard error (BSE) since:

$$[\overline{x}_1^*, \overline{x}_2^*, \overline{x}_3^*] =[1.1769, 1.1809, 1.1838] \rightarrow \overline{x}^*=1.1805,$$

$$Var_B(Y_n)=\tfrac{1}{N-1}\sum_{i=1}^{n}(\overline{x}_i^*-\overline{x}^*)^2=.000012,\ \ BSE(\overline{X}_n)=.00346.$$

In contrast, the estimated standard error of $\overline{X}_n$ around its true mean based on $\mathbf{x}_0$ is $SE(\overline{X}_n)$=.001485; considerably smaller!

Data specific patterns. Another problem with the bootstrap method is that the inference results based on its approximations are too data specific; see Efron and Tibshirani (1993). That is, by increasing $N\to\infty$ the ecdf of the bootstrap realizations $[\mathbf{x}_1^*,\mathbf{x}_2^*,...,\mathbf{x}_N^*]$, say $\widetilde{F}_n(x^*)$, will converge to $F(x)$ only when $\mathbf{x}_0$ is a 'truly typical realization' of the process $\{X_t,\ t\in\mathbb{N}\}$ assumed to underlying the generation of the data. For instance, in the case where the data have a genuine *outlier*, the bootstrapping is likely to give rise to distorted results.

For economic data, the probabilistic assumptions of IID are usually invalid, and the question that arises is whether one could extend the simple bootstrap to data that exhibit dependence and/or heterogeneity. The answer is a qualified yes, because most of the proposed resampling and sub-sampling methods do not work in practice, despite their optimal asymptotic properties. The problem is to generate realizations which constitute faithful replicas of the original data; see Politis, Romano and Wolf (1999), Lahiri (2003).

9 Summary and conclusions

The primary objective of this chapter has been twofold. First, to discuss how the interpretation of probability plays a crucial role in determining the approach to inference aim to learn from data about the phenomena that gave rise to the data. It is argued that the different interpretations of probability call for alternative procedures in learning from data. The focus has been placed on the frequentist and degrees of belief interpretations of probability that give rise to the frequentist and Bayesian approaches to statistical modeling and inference, respectively. The frequentist interpretation of probability gives rise to 'optimal' (most effective) inference procedures whose capacity to learn about the 'true' data generating mechanism $\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\},\ \mathbf{x}\in\mathbb{R}_X^n$, is maximized, or equivalently, their proclivity to err is minimized. The expression "$\boldsymbol{\theta}^*$ denotes the true value of $\boldsymbol{\theta}$" is a shorthand for saying that "data $\mathbf{x}_0$ constitute a typical realization of the sample $\mathbf{X}$ with distribution $f(\mathbf{x};\boldsymbol{\theta}^*)$".

The degree of belief interpretation of probability assumes that any learning from data takes the form of revising one's beliefs about how different values of $\boldsymbol{\theta}$ in Θ could have given rise to data $\mathbf{x}_0$ by comparing the prior $\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$ to the posterior $\pi(\boldsymbol{\theta}|\mathbf{x}_0)\propto L(\boldsymbol{\theta}|\mathbf{x}_0)\cdot\pi(\theta),\ \forall\boldsymbol{\theta}\in\Theta$, distribution. In a nutshell, learning from data in Bayesian statistics is $\pi(\boldsymbol{\theta}|\mathbf{x}_0)-\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$, where the original ranking of different values of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$ is revised in light of the data to be: $\pi(\boldsymbol{\theta}|\mathbf{x}_0),\ \forall\boldsymbol{\theta}\in\Theta$. Contrary to de Finetti's (1974) proclamation that probability only exists in the minds of individuals, adopting the stance that chance regularity patterns is a crucial feature of real world phenomena of interest, rendered the frequentist the preferred approach to empirical

modeling. In relation to that, it is argued that the best bridge between empirical regularities in the real world and probability in Plato's world comes in the form of the empirical cumulative distribution function (ecdf), $\widehat{F}(x)$, and its relationship to its theoretical counterpart, the cdf $F(x)$, $\forall x \in \mathbb{R}$. On the other hand, when the focus of interest is not empirical modeling of stochastic phenomena of interest, but decision making under uncertainty, the Bayesian approach, in conjunction with the decision theoretic setup, can be very useful.

Second, to provide a short introduction to three different approaches to statistical modeling and inference. The frequentist approach is contrasted with the Bayesian and the nonparametric approaches in an attempt to bring out their similarities and differences. The focus in the next 5 chapters will be on the frequentist approach, but it is important to understand how it differs from the other two approaches. The above discussion brings out certain neglected aspects of frequentist modeling associated with the modeling facet, problems of specification, misspecification testing and respecification, with a view to secure a statistically adequate model in order to secure the reliability of procedures used in the inference facet. It is argued that none of the approaches to inference, frequentist, Bayesian or nonparametric, is immune to statistical misspecification. Moreover, the only effective way to secure the reliability of inference is to establish the statistical adequacy of the prespecified statistical model.

Additional references: Arnold (1990), Azzalini (1996), Davison (2003).

Important concepts

Frequentist interpretation of probability, degrees of belief interpretation of probability, chance regularities are a feature of the real world, model-based frequentist interpretation of probability, SLLN and empirical regularities, the circularity charge for the frequentist interpretation, the long-run metaphor charge, the Kolmogorov complexity interpretation of probability, propensity interpretation of probability, the Dutch book argument, decision theoretic framing of probability, the 'true' data generating mechanism, likelihood function, prior and posterior distributions, loss functions, Karl Pearson's approach to statistics, Fisher's approach to frequentist statistics, specification, estimation, misspecification testing, respecification, interval estimation, hypothesis testing, the distribution of the sample, the likelihood function, sample space, parameter space, point estimation, interval estimation, hypothesis testing, prediction, the estimated cumulative distribution function (ecdf), nonparametric inference, weak model assumptions, undue reliance on asymptotic results, basic bootstrap.

Crucial distinctions

Frequentist vs. degrees of belief interpretation of probability, model-based vs. von Mises frequentist interpretation of probability, subjective vs. logical (objective) degrees of belief, probability of events vs. betting odds, probabilistic incoherence, statistical vs. substantive adequacy, frequentist vs. Bayesian inference, Karl Pearson's vs. Fisher's approach to statistics, factual vs. hypothetical reasoning, modeling vs.

inference facets of empirical modeling, sample vs. sample realization, the distribution of the sample vs. the likelihood function, sample vs. parameter space, estimator vs. estimate vs. parameter, probability vs. empirical frequencies, the cdf vs. the ecdf, testable vs. non-testable probabilistic assumptions, parametric vs. nonparametric inference.

Essential ideas

- The interpretation of mathematical probability plays a crucial role in determining the type and nature of the statistical inference giving rise to learning from data about phenomena of interest. The key issue is whether the chance regularity patterns in data represent a feature of real world phenomena or not.

- The two main interpretations of probability, the frequentist and degrees of belief, give rise to two very different approaches to statistical inference, known as frequentist (or classical) and Bayesian. The key difference between the two approaches revolves around the way they give rise to learning from data about phenomena of interest.

- For frequentist inference, learning from data revolves around optimal (capable) procedures that shed light on the 'true' stochastic generating mechanism that gave rise to the particular data.

- For Bayesian inference, learning from data revolves around revising one's prior beliefs pertaining to the ranking of the different values of θ in light of the data using the posterior distribution.

- Be aware of Bayesian attempts to denigrate frequentist inference using (deliberately) misleading arguments based on misinterpretations of frequentist inference and pathological statistical models.

- The real difference between parametric and nonparametric inference is the latter replaces easily testable distribution assumptions with non-testable indirect distribution assumptions pertaining to the existence of moments and the smoothness of the unknown density function.

- The extent to which nonparametric inference can guard against statistical misspecification is often overvalued because weaker but non-testable mathematical conditions seem less vulnerable to misspecification; that's an illusion.

- For mathematical deduction weaker assumptions are always considered as superior to stronger assumptions giving rise to the same result. The reverse is true for statistical induction because the latter depends crucially on the validity of its premises, and often weaker assumptions are non-testable.

- Relying exclusively on asymptotic approximations, based on as $n \to \infty$, for inference is always a bad strategy, as asserted by Le Cam (1986a).
- The use of probabilistic inequalities as a basis of statistical inference invariably gives rise to imprecise results.
- The basic bootstrap procedure for resampling is highly vulnerable to departures from the IID assumptions.

10 Questions and Exercises

1. Explain why the interpretation of mathematical probability plays a crucial role in determining the type and nature of statistical modeling and inference.

2. (a) Compare and contrast the degrees of belief and model-based frequentist interpretations of probability.

(b) Explain why the criticism that the model-based frequentist interpretation of probability is of very limited scope because it is only applicable in the case of IID data is misplaced.

(c) Explain the 'long-run' metaphor associated with the frequency interpretation of probability by paying particular attention to the key feature of replicability, and discuss Keynes's comment that "in the long-run we will all be dead".

3. (a) Explain briefly the difference between the model-based and von Mises frequentist interpretations of probability.

(b) Using your answer in (a) explain why the charges: (i) the circularity charge, (ii) the long-run metaphor, (iii) the single event probability charge, are misplaced when leveled against the model-based frequentist interpretation as it relates to the SLLN and the interpretative provisions.

4. (a) Discuss the common features of the model-based frequentist interpretation of probability and Kolmogorov's complexity interpretation of probability.

(b) Discuss the relationship between the model-based frequentist interpretation of probability and the propensity interpretation of probability. Explain why there is no conflict between the two interpretations.

5. Explain why a person with subjective probabilities associated with two *independent* events A and B are: $\Pr(A)=.5,\ \Pr(B)=.7,\ \Pr(A \cap B)=.2$, is incoherent.

6. (a) Compare and contrast the subjective and logical (objective) 'degrees of belief' interpretations of probability.

(b) Briefly compare and contrast the frequentist and Bayesian approaches to statistical inference on the basis of the delimiting features in table 10.6.

7. In the case of the simple Bernoulli model, explain the difference between frequentist inference based on a sampling distribution of an estimator of θ, say:

$$f(\widehat{\theta}(\mathbf{x});\theta),\ \forall \mathbf{x} \in \mathbb{R}^n,$$

and Bayesian inference based on the posterior distribution $\pi(\theta|\mathbf{x}_0),\ \forall\theta\in\Theta$.

8. (a) "... likelihoods are just as subjective as priors." (Kadane, 2011, p. 445). Discuss.

(b) Discuss the following claims by Koop, Poirier and Tobias (2007), p. 2:"... frequentists, argue that situations not admitting repetition under essentially identical conditions are not within the realm of statistical enquiry, and hence 'probability' should not be used in such situations. Frequentists define the probability of an event as its long-run relative frequency. The frequentist interpretation cannot be applied to (i) unique, once and for all type of phenomena, (ii) hypotheses, or (iii) uncertain past events. Furthermore, this definition is nonoperational since only a finite number of trials can ever be conducted.'.

9. (a) Compare and contrast Karl Pearson's approach to statistics with that of R.A. Fisher, and explain why the former implicitly assumes that the data constitute a realization of an IID sample.

10. (a) Compare and contrast the following: (i) sample vs. sample realization, (ii) estimator vs. estimate, (iii) distribution of the sample vs. likelihood function.

(b) Explain briefly why frequentist estimation (point and interval), testing and prediction are primarily based on mappings between the sample and parameter spaces.

11. "The various limit theorems relating to the asymptotic behavior of the empirical cumulative distribution function, in conjunction with the validity of the probabilistic assumptions it invokes, bestow empirical content to the mathematical concept of a cdf $F(x)$". Explain and discuss.

12. For the random variable X, where $E(X)$=0 and $Var(X)$=$\frac{1}{3}$, derive an upper bound on the probability of the event $\{|X-.6|>0.1\}$. How does this probability change if one knows that $X\backsim \mathsf{U}(-1,1)$?

13. For the random variable X, where $E(X)$=0 and $Var(X)$=1, derive an upper bound on the probability of the event $\{|X-.6|>0.1\}$. How does this probability change if one knows that $X\backsim \mathsf{N}(0,1)$? How accurate is the following inequality?

$$\mathbb{P}\left(|X|\geq\varepsilon\right)\leq\tfrac{1}{\sqrt{2\pi}}\int_{\varepsilon}^{\infty}\tfrac{x}{\varepsilon}e^{-\frac{x^2}{2}}dx=\tfrac{1}{\sqrt{2\pi}}\left(\tfrac{1}{\varepsilon}\right)e^{-\frac{\varepsilon^2}{2}},\ \text{for } x>\varepsilon.$$

14. (a) (i) In example 10.24 with ε=.1, evaluate the required sample size n to ensure that: $\mathbb{P}\left(|\overline{X}_n-\theta|>\varepsilon\right)\leq$.020.

(ii) Calculate the increase in n needed for the same upper bound (.02) with ε=.05.

(iii) Repeat (i)-(ii) using the Normal approximation to the finite sample distribution and compare the results.

(b) Discuss briefly the lessons to be learned from the Bahadur and Savage (1956) example 10.20 pertaining to the use of limit theorems for inference purposes.

15. (a) "For modeling purposes specific distribution assumptions are indispensable if we need precise and sharp results. Results based on bounded moment conditions are invariably imprecise and blunt." Discuss.

(b) "For a large enough sample size n, one does *not* need to worry about distributional assumptions, because one can use the bounds offered by limit theorems to 'calibrate' the reliability of any inference." Discuss.