Jonathan Rougier

School of Mathematics
University of Bristol

APTS Lecture notes on Statistical Inference

Our mission: To help people make
better choices under uncertainty.

Version 5, compiled on December 3, 2017.

1
Statistics: another short introduction

In Statistics we quantify our beliefs about things which we would like to know in the light of other things which we have measured, or will measure. This programme is not unique to Statistics: one distinguishing feature of Statistics is the use of *probability* to quantify the uncertainty in our beliefs. Within Statistics we tend to separate Theoretical Statistics, which is the study of algorithms and their properties, from Applied Statistics, which is the use of carefully-selected algorithms to quantify beliefs about the real world. This chapter is about Theoretical Statistics.

If I had to recommend one introductory book about Theoretical Statistics, it would be Hacking (2001). The two textbooks I find myself using most regularly are Casella and Berger (2002) and Schervish (1995). For travelling, Cox (2006) and Cox and Donnelly (2011) are slim and full of insights; Stigler (2016) likewise. For a snapshot of the current state of the art, especially where Statistics meets Machine Learning, see Efron and Hastie (2016).

1.1 Statistical models

This section covers the nature of a statistical model, and some of the basic conventions for notation.

A *statistical model* is an artefact to link our beliefs about things which we can measure to things we would like to know. Denote the values of the things we can measure as Y, and the values of the things we would like to know as X. These are all *random quantities*, indicating that their values, ahead of taking the measurements, are unknown to us. I will refer to X as the *predictands*, Y as the *observables*, and y^{obs} as the *observations*; the observations are actual values.

The convention in Statistics is that random quantities are denoted with capital letters, and particular values of those random quantities with small letters; e.g., x is a particular value that X could take. This sometimes clashes with another convention that matrices are shown with capital letters and scalars with small letters. A partial resolution is to use normal letters for scalars, and bold-face letters for vectors and matrices. However, I have stopped adhering to this convention, as it it usually clear what X is from the

context. Therefore both X and Y may be collections of quantities.

I term the set of possible (numerical) values for X the *realm* of X, after Lad (1996), and denote it $\mathcal{X}$. This illustrates another convention, common throughout Mathematics, that sets are denoted with ornate letters. The realm of (X, Y) is the cartesian product $\mathcal{X} \times \mathcal{Y}$. Where the realm is a cartesian product, then the margins are denoted with subscripts. So if $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, then $Z_1 = X$ and $Z_2 = Y$. The most common example is where $X = (X_1, \ldots, X_m)$, and the realm of each X_i is $\mathcal{X}$, so that the realm of X is $\mathcal{X}^m$.

In the definition of a statistical model, 'artefact' denotes an object made by a human, e.g. you or me. There are no statistical models that don't originate inside our minds. So there is no arbiter to determine the 'true' statistical model for (X, Y)—we may expect to disagree about the statistical model for (X, Y), between ourselves, and even within ourselves from one time-point to another.[1] In common with all other scientists, statisticians do not require their models to be true. Statistical models exist to make prediction feasible (see Section 1.3).

[1] Some people refer to the unknown *data generating process (DGP)* for (X, Y), but I have never found this to be a useful concept.

Maybe it would be helpful to say a little more about this. Here is the usual procedure in 'public' Science, sanitised and compressed:

1. Given an interesting question, formulate it as a problem with a solution.
2. Using experience, imagination, and technical skill, make some simplifying assumptions to move the problem into the mathematical domain, and solve it.
3. Contemplate the simplified solution in the light of the assumptions, e.g. in terms of robustness. Maybe iterate a few times.
4. Publish your simplified solution (including, of course, all of your assumptions), and your recommendation for the original question, if you have one. Prepare for criticism.

MacKay (2009) provides a masterclass in this procedure.[2] The statistical model represents a statistician's 'simplifying assumptions'.

[2] Many people have discussed the "unreasonable effectiveness of mathematics", to use the phrase of Eugene Wigner; see `https://en.wikipedia.org/wiki/The_Unreasonable_Effectiveness_of_Mathematics_in_the_Natural_Sciences`. Or, for a more nuanced view, Hacking (2014).

A statistical model takes the form of a *family of probability distributions* over $\mathcal{X} \times \mathcal{Y}$. I will assume, for notational convenience, that $\mathcal{X} \times \mathcal{Y}$ is countable.[3] Dropping Y for a moment, let $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots\}$. The complete set of probability distributions for X is

$$\mathcal{P} = \left\{ p \in \mathbb{R}^k : p \geq 0, \sum_{i=1}^{k} p_i = 1 \right\}, \tag{1.1}$$

[3] Everything in this chapter generalizes to the case where the realm is uncountable.

where $p_i = \mathbb{P}(X = x^{(i)})$, and $k = |\mathcal{X}|$, the number of elements of $\mathcal{X}$. A family of distributions is a subset of $\mathcal{P}$, say $\mathcal{F}$. In other words, a statistician creates a statistical model by ruling out many possible probability distributions.

The particular way in which statisticians specify a subset of all distributions originates with Ronald Fisher in the 1920s; Stephen Stigler calls it "One of Ronald A. Fisher's subtlest innovations"

(Stigler, 2016, p. 180). The family is denoted by a *probability mass function (PMF)* f_X, a *parameter* θ, and a *parameter space* Ω, such that

$$\mathcal{F} = \left\{ p \in \mathcal{P} : \forall i \; p_i = f_X(x^{(i)}; \theta) \text{ for some } \theta \in \Omega \right\}. \tag{1.2}$$

For obvious reasons, we require that if $\theta' \neq \theta''$, then

$$f_X(\cdot\,; \theta') \neq f_X(\cdot\,; \theta''); \tag{1.3}$$

such models are termed *identifiable*.[4] Taken all together, it is convenient to denote a statistical model for X as the triple

$$\mathcal{E} = \{\mathcal{X}, \Omega, f_X\}, \tag{1.4}$$

termed a *parametric model*. For example, the Poisson family is

$$\text{Poisson} = \{\mathbb{N}, \mathbb{R}_+, f_X\} \text{ where } f_X(x; \theta) = e^{-\theta} \frac{\theta^x}{x!},$$

although it is common in this case to use 'λ' rather than 'θ' as the label for the parameter.[5] Where $\mathcal{X}$ is embedded in a larger set, it is understood that $f_X(x; \cdot) = 0$ for $x \notin \mathcal{X}$. This would allow us to define the Poisson distribution over the realm $\mathbb{R}$, if that turned out to be convenient.

Most statistical procedures start with the specification of a statistical model for (X, Y),

$$\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Omega, f_{X,Y}\}. \tag{1.5}$$

The method by which a statistician chooses $\mathcal{F}$ and then $\mathcal{E}$ is hard to codify, although experience and precedent are obviously relevant. See Davison (2003) for a book-length treatment with many useful examples. Some procedures start with a more general specification for f_X, termed *non-parametric* statistical models. The most common is that $f_X(x_1, \ldots, x_m)$ is a symmetric function of $(x_1, \ldots, x_m)$, termed *exchangeable*.

1.2 Hierarchies of models

The concept of a statistical model was crystalized in the early part of the 20th century. At that time, when the notion of a digital computer was no more than a twinkle in John von Neumann's eye, the 'f_Y' in the model $\{\mathcal{Y}, \Omega, f_Y\}$ was assumed to be a known analytic function of y for each θ.[6] As such, all sorts of other useful operations are possible, such as differentiating with respect to θ. Expressions for the PMFs of specified functions of set of random quantities are also known analytic functions: sums, differences, and more general transformations.

This was computationally convenient—in fact it was critical given the resources of the time—but it severely restricted the models which could be used in practice, more-or-less to the models found today at the back of every textbook in Statistics (e.g. Casella and Berger, 2002), or simple combinations thereof. Since about the

[4] Some more notation. f_X is a function; formally, $f_X : \mathcal{X} \times \Omega \to [0, 1]$. Two functions can be compared for equality: as functions are sets of tuples, the comparison is for the equality of two sets. $f_X(\cdot\,; \theta)$ is also a function, $f_X(\cdot\,; \theta) : \mathcal{X} \to [0, 1]$ but different for each value of θ. It is a convention in Statistics to separate the argument x from the parameter θ using a semi-colon.

[5] $\mathbb{N}$ denotes the set of natural numbers, and $\mathbb{R}_+$ the set of non-negative real numbers. Mathematicians are flexible about whether $0 \in \mathbb{N}$: in our case it is.

[6] That is, a function which can be evaluated to any specified precision using a finite number of operations, like the Poisson PMF or the Normal probability density function (PDF).

1950s—the start of the computer age—we have had the ability to evaluate a much wider set of functions, and to simulate random quantities on digital computers. As a result, the set of usable statistical models has dramatically increased. In modern Statistics, we now have the freedom to specify the model that most effectively represents our beliefs about the set of random quantities of interest. Therefore we need to update our notion of statistical model, according to the following hierarchy.

A. Models where f_Y has a known analytic form.

B. Models where $f_Y(y;\theta)$ can be evaluated.

C. Models where Y can be simulated from $f_Y(\cdot;\theta)$.

Between (B) and (C) exist models where $f_Y(y;\theta)$ can be evaluated up to an unknown constant, which may or may not depend on θ.

To illustrate the difference, consider the Maximum Likelihood Estimator (MLE) of the 'true' value of θ based on Y, defined as

$$\hat{\theta}(y) := \sup_{\theta\in\Omega} f_Y(y;\theta). \tag{1.6}$$

Eq. (1.6) is just a sequence of mathematical symbols, waiting to be instantiated into an algorithm. If f_Y has a known analytic form, i.e. level (A) of the hierarchy, then it may be possible to solve the first-order conditions,[7]

$$\frac{\partial}{\partial\theta} f_Y(y;\theta) = 0, \tag{1.7}$$

uniquely for θ as a function of y (assuming, for simplicity, that Ω is a convex subset of $\mathbb{R}$) and to show that $\frac{\partial^2}{\partial\theta^2} f_Y(y;\theta)$ is negative at this solution. In this case we are able to derive an analytic expression for $\hat{\theta}$. Even if we cannot solve the first order conditions, we might be able to prove that $f_Y(y;\cdot)$ is strictly concave, so that we know there is a unique maximum. This means that any numerical maximization of $f_Y(y;\cdot)$ is guaranteed to converge to $\hat{\theta}(y)$.

But what if we can evaluate $f_Y(y;\theta)$, but do not know its form, i.e. level (B) of the hierarchy? In this case we can still numerically maximize $f_Y(y;\cdot)$, but we cannot be sure that the maximizer will converge to $\hat{\theta}(y)$: it may converge to a local maximum. So the algorithm for finding $\hat{\theta}(y)$ must have some additional procedures to ensure that all local maxima are ignored: this is very complicated in practice, very resource intensive, and there are no guarantees.[8] So in practice the Maximum Likelihood algorithm does not necessarily give the MLE. We must recognise this distinction, and not make claims for the MLE algorithm which we implement, that are based on theoretical properties of the MLE.

And what about level (C) of the hierarchy? It is very tricky indeed to find the MLE in this case, and any algorithm that tries will be very imperfect. Other estimators of θ would usually be preferred. This example illustrates that in Statistics it is the choice

[7] For simplicity and numerical stability, these would usually be applied to $\log f_Y$ not f_Y.

[8] See, e.g., Nocedal and Wright (2006). Do not be tempted to make up your own numerical maximization algorithm.

of algorithm that matters. The MLE is a good choice only if (i) you can prove that it has good properties for your statistical model,[9] and (ii) you can prove that your algorithm for finding the MLE is in fact guaranteed to find the MLE for your statistical model. If you have used an algorithm to find the MLE without checking both (i) and (ii), then your results bear the same relation to Statistics as Astrology does to Astronomy. Doing Astrology is fine, but not if your client has paid you to do Astronomy.

[9] Which is often very unclear; see Le Cam (1990).

1.3 *Prediction and inference*

The applied statistician proposes a statistical model for (X, Y),

$$\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Omega, f_{X,Y}\}.$$

She then turns $\mathcal{E}$ and y^{obs} into a prediction for X. Ideally she uses an algorithm, in the sense that were she given the same statistical model and same observations again, she would produce the same prediction.

A statistical prediction is always a probability distribution for X, although it might be summarised, for example as the expectation of some specified function of X. From the starting point of the statistical model $\mathcal{E}$ and the value of an observable Y we derive the *predictive model*

$$\mathcal{E}^* = \{\mathcal{X}, \Omega, f_X^*\} \tag{1.8a}$$

where

$$f_X^*(\cdot; \theta) = \frac{f_{X,Y}(\cdot, y; \theta)}{f_Y(y; \theta)} \tag{1.8b}$$

$$\text{and } f_Y(y; \theta) = \sum_x f_{X,Y}(x, y; \theta); \tag{1.8c}$$

I often write '$*$' to indicate a suppressed y argument. Here f_X^* is the conditional PMF of X given that $Y = y$, and f_Y is the marginal PMF of Y. Both of these depend on the parameter θ. The challenge for prediction is to reduce the family of distributions $\mathcal{E}^*$ down to a single distribution; effectively, to 'get rid of' θ.

There are two approaches to getting rid of θ: *plug in* and *integrate out*, found in the Frequentist and Bayesian paradigms respectively, for reasons that will be made clear below. We accept, as our working hypothesis, that one of the elements of the family $\mathcal{F}$ is true. For a specified statistical model $\mathcal{E}$, this is equivalent to stating that exactly one element in Ω is true: denote this element as Θ.[10,11] Then $f_X^*(\cdot; \Theta)$ is the true predictive PMF for X.

For the plug-in approach we replace Θ with an estimate based on y, for example the MLE $\hat{\theta}$. In other words, we have an algorithm

$$y \mapsto f_X^*\big(\cdot\,; \hat{\theta}(y)\big) \tag{1.9}$$

to derive the predictive distribution for X for any y. The estimator does not have to be the MLE: different estimators of Θ produce different algorithms.

[10] Note that I do not feel the need to write 'true' in scare-quotes. Clearly there is no such thing as a true value for θ, because the model is an artefact (i.e. not true in any defensible sense). But once we accept, as a working hypothesis, that one of the elements of $\mathcal{F}$ is true, we do not have to belabour the point.

[11] I am following Schervish (1995) and using Θ for the true value of θ, although it is a bit clunky as notation.

For the integrate-out approach we provide a *prior distribution* over Ω, denoted π.[12] This produces a *posterior distribution*

$$\pi^*(\cdot) = \frac{f_Y(y;\cdot)\,\pi(\cdot)}{\mathrm{p}(y)} \tag{1.10a}$$

where

$$\mathrm{p}(y) = \int_\Omega f_Y(y;\theta)\,\pi(\theta)\,\mathrm{d}\theta \tag{1.10b}$$

(Bayes's theorem, of course). Here $\mathrm{p}(y)$ is termed the *marginal likelihood* of y. Then we integrate out θ according to the posterior distribution—another algorithm:

$$y \mapsto \int_\Omega f_X^*(\cdot\,;\theta)\,\pi^*(\theta)\,\mathrm{d}\theta. \tag{1.11}$$

Different prior distributions produce different algorithms.

That is prediction in a nutshell. In the plug-in approach, each estimator for Θ produces a different algorithm. In the integrate-out approach each prior distribution for Θ produces a different algorithm. Neither approach works on y alone: both need the statistician to provide an additional input: a point estimator, or a prior distribution. Frequentists dislike specifying prior distributions, and therefore favour the plug-in approach. Bayesians like specifying prior distributions, and therefore favour the integrate-out approach.[13]

* * *

This outline of prediction illustrates exactly how Statistics has become so concerned with *inference*. Inference is learning about Θ, which is a key part of either approach to prediction: either we need a point estimator for Θ (plug-in), or we need a posterior distribution for Θ (integrate-out). It often seems as though Statistics is mainly about inference, but this is misleading. It is about inference only insofar as inference is the first part of prediction.

Ideally, algorithms for inference should only be evaluated in terms of their performance as components of algorithms for prediction. This does not happen in practice: partly because it is much easier to assess algorithms for inference than for prediction; partly because of the fairly well-justified belief that algorithms that perform well for inference will produce algorithms that perform well for prediction.[14] I will adhere to this practice, and focus mainly on inference. *But not forgetting that Statistics is mainly about prediction.*

1.4 Frequentist procedures

As explained immediately above, I will focus on inference. So consider a specified statistical model $\mathcal{E} = \{\mathcal{Y}, \Omega, f_Y\}$, where the objective is to learn about the true value $\Theta \in \Omega$ based on the value of the observables Y.

[12] For simplicity, and almost always in practice, π is a probability density function (PDF), given that Ω is almost always a convex subset of Euclidean space.

[13] We often write 'Frequentists' and 'Bayesians', and most applied statisticians will tend to favour one approach or the other. But applied statisticians are also pragmatic. Although a 'mostly Bayesian' myself, I occasionally produce confidence sets.

[14] Often, this is because prediction questions can be expressed in terms of specified functions of the parameters.

We have already come across the notion of an *algorithm*, which is represented as a function of the value of the observables; in this section I will denote the algorithm as 'g'. Thus the domain of g is always $\mathcal{Y}$. The co-domain of g depends on the type of inference (see below for examples). The key feature of the Frequentist paradigm is the following principle.

Definition 1.1 (Certification)**.** For a specified model $\mathcal{E}$ and algorithm g, the *sampling distribution* of g is

$$f_G(v;\theta) = \sum_{y:g(y)=v} f_Y(y;\theta). \tag{1.12}$$

Then:

1. Every algorithm is certified by its sampling distribution, and
2. The choice of algorithm depends on this certification.

This rather abstract principle may not be what you were expecting, based on your previous courses in Statistics, but if you reflect on the following outline you will see that is the common principle underlying what you have previously been taught.

Different algorithms are certified in different ways, depending on their nature. Briefly, point estimators of Θ may be certified by their *Mean Squared Error function*. Set estimators of Θ may be certified by their *coverage function*. Hypothesis tests for Θ may be certified by their *power function*. The definition of each of these certifications is not important here, although they are easy to look up. What is important to understand is that in each case an algorithm g is proposed, f_G is inspected, and then a certificate is issued.

Individuals and user communities develop conventions about what certificates they like their algorithms to possess, and thus they choose an algorithm according to its certification. They report both $g(y^{\text{obs}})$ and the certification of g. For example, "$(0.73, 0.88)$ is a 95% confidence interval for Θ". In this case g is a set estimator for Θ, it is certified as 'level 95%', and its value is $g(y^{\text{obs}}) = (0.73, 0.88)$.

* * *

Certification is extremely challenging. Suppose I possess an algorithm $g : \mathcal{Y} \to 2^\Omega$ for set estimation.[15] In order to certify it as a confidence procedure for my model $\mathcal{E}$ I need to compute its coverage for every $\theta \in \Omega$, defined as

$$\text{coverage}(\theta; \mathcal{E}) = \mathbb{P}\{\theta \in g(Y); \theta\} = \sum_v \mathbb{1}_{\theta \in v} f_G(v; \theta), \qquad (1.13)$$

where '$\mathbb{1}_a$' is the indicator function of the proposition a, which is 0 when a is false, and 1 when a is true. Except in special cases, computing the coverage for every $\theta \in \Omega$ is impossible, given that Ω is uncountable.[16]

So, in general, I cannot know the coverage function of my algorithm g for my model $\mathcal{E}$, and thus I cannot certify it accurately, but only approximately. Unfortunately, then I have a second challenge. After much effort, I might (approximately) certify g for my model $\mathcal{E}$ as, say, 'level 83%'; this means that the coverage is at least 83% for every $\theta \in \Omega$. Unfortunately, the convention in my user community is that confidence procedures should be certified as 'level 95%'. So it turns out that my community will not accept g. I have to find a way to work backwards, *from* the required certificate, *to* the choice of algorithm.

So Frequentist procedures require the solution of an intractable inverse problem: for specified model $\mathcal{E}$, produce an algorithm g with the required certificate. Actually, it is even harder than this, because it turns out that there are an uncountable number of algorithms with the right certificate, but most of them are useless. Most applied statisticians do not have the expertise or the computing resources to solve this problem to find a good algorithm with the required certificate, for their model $\mathcal{E}$. And so Frequentist procedures, when they are used by applied statisticians, tend to rely on a few special cases. Where these special cases are not appropriate, applied statisticians tend to reach for an off-the-shelf algorithm justified using a theoretical approximation, plus hope.

The empirical evidence collected over the last decade suggests that the hope has been in vain. Most algorithms (including those based on the special cases) did not, in fact, have the certificate that was claimed for them.[17] Opinion is divided about whether this is fraud or merely ignorance. Practically speaking, though, there is no doubt that Frequentist procedures are not being successfully implemented by applied statisticians.

1.5 *Bayesian procedures*

We continue to treat the model $\mathcal{E}$ as given. As explained in the previous section, Frequentist procedures select algorithms according to their certificates. By contrast, Bayesian procedures select algorithms mainly according to the prior distribution π (see Section 1.3), without regard for the algorithm's certificate.

A Bayesian inference is synonymous with the posterior distribution π^*, see (1.10). This posterior distribution may be summarized

[15] Notation. 2^Ω is the set of all subsets of Ω, termed the 'power set' of Ω.

[16] The special cases are a small subset of models from (A) in the model hierarchy in Section 1.2, where, for a particular choice of g, the sampling distribution of g and the coverage of g can be expressed as an analytic function of θ. If you ever wondered why the Normal linear model is so common in applied statistics (linear regression, z-scores, t-tests, and F-statistics, ANOVA, etc.), then wonder no more. Effectively, this family makes up most of the special cases.

[17] See Madigan et al. (2014) for one such study or, if you want to delve, google "crisis reproducibility science". There is even a wikipedia page, `https://en.wikipedia.org/wiki/Replication_crisis`, which dates from Jan 2015.

according to some method, for example to give a point estimate, a set estimate, do a hypothesis test, and so on. These summary methods are fairly standard, and do not represent an additional source of choice for the statistician. For example, a Bayesian algorithm for choosing a set estimator for Θ would be (i) choose a prior distribution π, (ii) compute the posterior distribution π^*, and (iii) extract the 95% High Density Region (HDR).

In principle, we could compute the coverage function of this algorithm, and certify it as a confidence procedure. It is very unlikely that it would be certified as a 'level 95%' confidence procedure, because of the influence of the prior distribution.[18] A Bayesian statistician would not care, though, because she does not concern herself with the certificate of her algorithm. When the model is given, the only thing the Bayesian has to worry about is her prior distribution.

[18] Nevertheless, there are theorems that give conditions on the model and the prior distribution such that the posterior 95% HDR is approximately a level 95% confidence procedure; see, e.g., Schervish (1995, ch. 7).

Bayesians see the prior distribution as an opportunity to construct a richer model for (X, Y) than is possible for Frequentists. This is most easily illustrated with a hierarchical model, for a population of quantities that are similar, and a sample from that population. Hierarchical models have a standard notation:[19]

$$Y_i \mid X_i, \sigma^2 \sim f_{\epsilon_i}(X_i, \sigma^2) \qquad i = 1, \ldots, n \tag{1.14a}$$
$$X_i \mid \theta_i \sim f_{X_i}(\theta_i) \qquad i = 1, \ldots, m \tag{1.14b}$$
$$\theta_i \mid \psi \sim f_\theta(\psi) \qquad i = 1, \ldots, m \tag{1.14c}$$
$$(\sigma^2, \psi) \sim f_0\,. \tag{1.14d}$$

[19] See, e.g., Lunn et al. (2013) or Gelman et al. (2014). Each of the f functions is a PMF or PDF, and the first argument is suppressed. The i index in the first three rows indicates that the components are mutually independent, and then the f function shows the marginal distribution for each i, which may depend on i. In the third row f does not depend on i, so that the θ_i's are mutually independent and identically distributed, or 'IID'.

At the top (first) level is the measurement model for the sample $(Y_1, \ldots, Y_n)$, where f_{ϵ_i} describes the measurement error and σ^2 would usually be a scale parameter. At the second level is the model for the population $(X_1, \ldots, X_m)$, where $n \leq m$, showing how each element X_i is 'summarised' by its own parameter θ_i. At the third level is the parameter model, in which the parameters are allowed to be different from each other. At the bottom (fourth) level is the 'hyper-parameter' model, which describes how much the parameters can differ, and also provides a PDF for the scale parameter σ^2.

Frequentists would specify their statistical model using just the top two levels, in terms of the parameter $(\sigma^2, \theta_1, \ldots, \theta_m)$, or, if this is too many parameters for the n observables, as it usually is, they will insist that $\theta_1 = \cdots = \theta_m = \theta$, and have just (σ^2, θ). The bottom two levels are the Bayesian's prior distribution. By adding these two levels, Bayesians can allow the θ_i's to vary, but in a limited way that can be controlled by their choices for f_θ and f_0. Usually, f_0 is a 'vague' PDF selected according to some simple rules.

In a Frequentist model we can count the number of parameters, namely $1 + m \cdot \dim \Omega$, or just $1 + \dim \Omega$ if the θ_i's are all the same. We can do that in a Bayesian model too, to give $1 + m \cdot \dim \Omega + \dim \Psi$, if Ψ is the realm of ψ. Bayesian models tend to have many more parameters, which makes them more flex-

ible. But there is a second concept in a Bayesian model, which is the *effective* number of parameters. This can be a lot lower than the actual number of parameters, if it turns out that the observations indicate that the θ_i's are all very similar. So in a Bayesian model the effective number of parameters can depend on the observations. In this sense, a Bayesian model is more adaptive than a Frequentist model.[20]

[20] The issue of how to quantify the effective number of parameters is quite complicated. Spiegelhalter et al. (2002) was a controversial suggestion, and there have been several developments since then, summarised in Spiegelhalter et al. (2014).

1.6 So who's right?

We return to the problem of inference, based on the model $\mathcal{E} = \{\mathcal{Y}, \Omega, f_Y\}$. Here is the pressing question, from the previous two sections: should we concern ourselves with the certificate of the algorithm, or with the choice of the prior distribution?

A Frequentist would say "Don't you want to know that you will be right 'on average' according to some specified rate?" (like 95%). And a Bayesian will reply "Why should my rate 'on average' matter to me right now, when I am thinking only of Θ?"[21] The Bayesian will point out the advantage of being able to construct hierarchical models with richer structure. Then the Frequentist will criticise the 'subjectivity' of the Bayesian's prior distribution. The Bayesian will reply that the model is also subjective, and so 'subjectivity' of itself cannot be used to criticise only Bayesian procedures. And she will go on to point out that there is just as much subjectivity in the Frequentist's choice of algorithm as there is in the Bayesian's choice of prior.

[21] And if she really wants to twist the knife she will also mention the overwhelming evidence that Frequentist statisticians have apparently not been able to achieve their target rates, mentioned at the end of Section 1.4.

There is no clear winner when two paradigms butt heads. However, momentum is now on the side of the Bayesians. Back in the 1920s and 1930s, at the dawn of modern Statistics, the Frequentist paradigm seemed to provide the 'objectivity' that was then prized in science. And computation was so rudimentary that no one thought beyond the simplest possible models, and their natural algorithms. But then the Frequentist paradigm took a couple of hard knocks: from Wald's Complete Class Theorem in 1950 (covered in Chapter 3), and from Birnbaum's Theorem and the Likelihood Principle in the 1960s (covered in Chapter 2). Significance testing was challenged by Lindley's paradox; estimator theory by Stein's paradox and the Neyman-Scott paradox. Bayesian methods were much less troubled by these results, and were developed in the 1950s and 1960s by two very influential champions, L.J. Savage and Dennis Lindey, building on the work of Harold Jeffreys.[22]

[22] With a strong assist from the maverick statistician I.J. Good. The intellectual forebears of the 20th century Bayesian revival included J.M. Keynes, F.P. Ramsey, Bruno de Finetti, and R.T. Cox.

And then in the 1980s, the exponential growth in computer power and new Monte Carlo methods combined to make the Bayesian approach much more practical. Additionally, datasets have got larger and more complicated, favouring the Bayesian approach with its richer model structure, when incorporating the prior distribution. Finally, there is now much more interest in uncertainty in predictions, something that the Bayesian integrate-out approach handles much better than the Frequentist plug-in

approach (Section 1.3).

However, I would not rule out a partial reversal in due course, under pressure from Machine Learning (ML). ML is all about algorithms, which are often developed quite independently of any statistical model. With modern Big Data (BD), the primary concern of an algorithm is that it executes in a reasonable amount of time (see, e.g., Cormen et al., 1990). But it would be natural, when an ML algorithm might be applied by the same agent thousands of times in quite similar situations, to be concerned about its sampling distribution.[23] With BD the certificate can be assessed from a held-out subset of the data, without any need for a statistical model—no need for statisticians at all then! Luckily for us statisticians, there will always be plenty of applications where ML techniques are less effective, because the datasets are smaller, or more complicated. In these applications, I expect Bayesian procedures will come to dominate.[24]

[23] For example, if an algorithm is a binary classifier, to want to know its 'false positive' and 'false negative' rates.

[24] See Harford (2014) for an interesting essay about why big is not always better, and why in many situations we can expect statisticians to outperform 'data analysts'.

2
Principles for Statistical Inference

This chapter will be a lot clearer if you have recently read Chapter 1. An extremely compressed version follows. As a working hypothesis, we accept the truth of a statistical model

$$\mathcal{E} := \{\mathcal{X}, \Omega, f\} \tag{2.1}$$

where $\mathcal{X}$ is the realm of a set of random quantities X, θ is a parameter with domain Ω (the 'parameter space'), and f is a probability mass function for which $f(x;\theta)$ is the probability of $X = x$ under parameter value θ.[1] The true value of the parameter is denoted Θ. Statistical inference is learning about Θ from the value of X, described in terms of an algorithm involving $\mathcal{E}$ and x. Although Statistics is really about prediction, inference is a crucial step in prediction, and therefore often taken as a goal in its own right.

[1] As is my usual convention, I assume, without loss of generality, that $\mathcal{X}$ is countable, and that Ω is uncountable.

Statistical principles guide the way in which we learn about Θ. They are meant to be either self-evident, or logical implications of principles which are self-evident. What is really interesting about Statistics, for both statisticians and philosophers (and real-world decision makers) is that the logical implications of some self-evident principles are not at all self-evident, and have turned out to be inconsistent with prevailing practices. This was a discovery made in the 1960s. Just as interesting, for sociologists (and real-world decision makers) is that the then-prevailing practices have survived the discovery, and continue to be used today.

This chapter is about statistical principles, and their implications for statistical inference. It demonstrates the power of abstract reasoning to shape everyday practice.

2.1 *Reasoning about inferences*

Statistical inferences can be very varied, as a brief look at the 'Results' sections of the papers in an Applied Statistics journal will reveal. In each paper, the authors have decided on a different interpretation of how to represent the 'evidence' from their dataset. On the surface, it does not seem possible to construct and reason about statistical principles when the notion of 'evidence' is so plastic. It was the inspiration of Allan Birnbaum (Birnbaum, 1962) to see—albeit indistinctly at first—that this issue could be side-stepped.

Over the next two decades, his original notion was refined; key papers in this process were Birnbaum (1972), Basu (1975), Dawid (1977), and the book by Berger and Wolpert (1988).

The model $\mathcal{E}$ is accepted as a working hypothesis, and so the existence of the true value Θ is also accepted under the same terms. How the statistician chooses her statements about the true value Θ is entirely down to her and her client: as a point or a set in Ω, as a choice among alternative sets or actions, or maybe as some more complicated, not ruling out visualizations. Dawid (1977) puts this well—his formalism is not excessive, for really understanding this crucial concept. The statistician defines, *a priori*, a set of possible 'inferences about Θ', and her task is to choose an element of this set based on $\mathcal{E}$ and x. Thus the statistician should see herself as a function 'Ev': a mapping from $(\mathcal{E}, x)$ into a predefined set of 'inferences about Θ', or

$$(\mathcal{E}, x) \xmapsto{\text{statistician, Ev}} \text{Inference about } \Theta.$$

Birnbaum called $\mathcal{E}$ the 'experiment', x the 'outcome', and Ev the 'evidence'.

Birnbaum's formalism, of an experiment, an outcome, and an evidence function, helps us to anticipate how we can construct statistical principles. First, there can be different experiments with the same Θ. Second, under some outcomes, we would agree that it is self-evident that these different experiments provide the same evidence about Θ. Finally, as will be shown, these self-evident principles imply other principles. These principles all have the same form: under such and such conditions, the evidence about Θ should be the same. Thus they serve only to rule out inferences that satisfy the conditions but have different evidences. They do not tell us how to do an inference, only what to avoid.

But if you find the idea of 'Ev' too abstract, then replace it in your mind and your notes with a specific instance of 'Ev', such as the ML estimate or a 95% confidence interval. E.g., everywhere you see 'Ev', read it as 'ML estimate of Θ'.

2.2 *The principle of indifference*

Here is our first example of a statistical principle, using the name conferred by Basu (1975). Recollect that once f has been defined, $f(x;\bullet)$ is a function of θ, potentially a different function for each x, and $f(\bullet;\theta)$ is a function of x, potentially a different function for each θ.[2]

[2] I am using '$\bullet$' instead of '$\cdot$' in this chapter and subsequent ones, because I like to use '$\cdot$' to denote scalar multiplication.

Definition 2.1 (Weak Indifference Principle, WIP)**.** Let $\mathcal{E} = \{\mathcal{X}, \Omega, f\}$. If $x, x' \in \mathcal{X}$ satisfy $f(x;\bullet) = f(x';\bullet)$, then $\text{Ev}(\mathcal{E}, x) = \text{Ev}(\mathcal{E}, x')$.

In my opinion, this is not self-evident, although, at the same time, is it not obviously wrong.[3] But we discover that it is the

[3] Birnbaum (1972) thought it was self-evident.

logical implication of two other principles which I accept as self-evident. These other principles are as follows, using the names conferred by Dawid (1977).

Definition 2.2 (Distribution Principle, DP)**.** If $\mathcal{E} = \mathcal{E}'$, then $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}', x)$.

As Dawid (1977) puts it, any information which is not represented in $\mathcal{E}$ is irrelevant. This seems entirely self-evident to me, once we enter the mathematical realm in which we accept the truth of our statistical model.

Definition 2.3 (Transformation Principle, TP)**.** Let $\mathcal{E} = \{\mathcal{X}, \Omega, f\}$. Let $g : \mathcal{X} \to \mathcal{Y}$ be bijective, and let $\mathcal{E}^g$ be the same experiment as $\mathcal{E}$ but expressed in terms of $Y = g(X)$, rather than X. Then $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}^g, g(x))$.

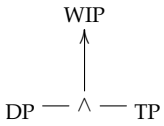
This principle states that inferences should not depend on the way in which the sample space is labelled, which also seems self-evident to me; at least, to violate this principle would be bizarre. But now we have the following result (Basu, 1975; Dawid, 1977).

Theorem 2.4. $(DP \wedge TP) \to WIP$.

Proof. Fix $\mathcal{E}$, and suppose that $x, x' \in \mathcal{X}$ satisfy $f(x; \bullet) = f(x'; \bullet)$, as in the condition of the WIP. Now consider the transformation $g : \mathcal{X} \to \mathcal{X}$ which switches x for x', but leaves all of the other elements of $\mathcal{X}$ unchanged. In this case $\mathcal{E} = \mathcal{E}^g$. Then

$$\begin{aligned} \mathrm{Ev}(\mathcal{E}, x') &= \mathrm{Ev}(\mathcal{E}^g, x') && \text{by the DP} \\ &= \mathrm{Ev}(\mathcal{E}^g, g(x)) \\ &= \mathrm{Ev}(\mathcal{E}, x) && \text{by the TP,} \end{aligned}$$

which is the WIP. □



So I find, as a matter of logic, I must accept the WIP, or else I must decide which of the two principles DP and TP are, contrary to my initial impression, not self-evident at all. This is the pattern of the next two sections, where either I must accept a principle, or, as a matter of logic, I must reject one of the principles that implies it. From now on, I will treat the WIP as self-evident.

2.3 *The Likelihood Principle*

The new concept in this section is a 'mixture' of two experiments. Suppose I have two experiments,

$$\mathcal{E}_1 = \{\mathcal{X}_1, \Omega, f_1\} \quad \text{and} \quad \mathcal{E}_2 = \{\mathcal{X}_2, \Omega, f_2\},$$

which have the same parameter Θ. Rather than do one experiment or the other, I imagine that I can choose between them randomly,

based on known probabilities (p_1, p_2), where $p_2 = 1 - p_1$. The resulting mixture is denoted $\mathcal{E}^* = \{\mathcal{X}^*, \Omega, f^*\}$, where

$$\mathcal{X}^* = (\{1\} \times \mathcal{X}_1) \cup (\{2\} \times \mathcal{X}_2), \tag{2.2a}$$

$$f^*((i, x_i); \theta) = p_i \cdot f_i(x_i; \theta). \tag{2.2b}$$

$\mathcal{E}^*$ is a mixture experiment.

The famous example of a mixture experiment is the 'two instruments' (see Cox and Hinkley, 1974, sec. 2.3). There are two instruments in a laboratory, and one is accurate, the other less so. The accurate one is more in demand, and typically it is busy 80% of the time. The inaccurate one is usually free. So, *a priori*, there is a probability of $p_1 = 0.2$ of getting the accurate instrument, and $p_2 = 0.8$ of getting the inaccurate one. Once a measurement is made, of course, there is no doubt about which of the two instruments was used. The following principle asserts what must be self-evident to everybody, that inferences should be made according to which instrument was used, and not according to the *a priori* uncertainty. Or, to paraphrase, *don't take account of experiments that were not performed*.

Definition 2.5 (Weak Conditionality Principle, WCP)**.** If $\mathcal{E}^*$ is a mixture experiment, as defined above, then

$$\mathrm{Ev}\left(\mathcal{E}^*, (i, x_i)\right) = \mathrm{Ev}(\mathcal{E}_i, x_i).$$

---

* * *

Another principle does not seem, at first glance, to have anything to do with the WCP. This is the Likelihood Principle.[4]

[4] The LP is self-attributed to G. Barnard, see his comment to Birnbaum (1962), p. 308. But it is alluded to in the statistical writings of R.A. Fisher, almost appearing in its modern form in Fisher (1956).

Definition 2.6 (Likelihood Principle, LP)**.** Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two experiments which have the same parameter Θ. If $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy

$$f_1(x_1; \bullet) = c(x_1, x_2) \cdot f_2(x_2; \bullet) \tag{2.3}$$

for some function $c > 0$, then $\mathrm{Ev}(\mathcal{E}_1, x_1) = \mathrm{Ev}(\mathcal{E}_2, x_2)$.

For a given $(\mathcal{E}, x)$, the function $f(x; \bullet)$ is termed the 'likelihood function' for $\theta \in \Omega$. Thus the LP states that if two likelihood functions for the same parameter have the same shape, then the evidence is the same—hence the name. As will be discussed in Section 2.6, Frequentist inferences violate the LP. Therefore the following result was something of the bombshell, when it first emerged in the 1960s. The following form is due to Birnbaum (1972) and Basu (1975).[5]

[5] Birnbaum's original result (Birnbaum, 1962), used a stronger condition than WIP and a slightly weaker condition than WCP. Theorem 2.7 is clearer.

Theorem 2.7 (Birnbaum's Theorem)**.** $(WIP \wedge WCP) \leftrightarrow LP$.

Proof. Both LP $\rightarrow$ WIP and LP $\rightarrow$ WCP are straightforward. The trick is to prove (WIP $\wedge$ WCP) $\rightarrow$ LP. So let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two

experiments which have the same parameter, and suppose that $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy $f_2(x_2; \bullet) = c \cdot f_1(x_1; \bullet)$, where $c > 0$ is some constant which may depend on (x_1, x_2), as in the condition of the LP. The value c is known, so consider the mixture experiment with $p_1 = c/(1+c)$ and $p_2 = 1/(1+c)$. Then

$$\begin{aligned} f^*\big((1, x_1); \bullet\big) &= \frac{c}{1+c} \cdot f_1(x_1; \bullet) \\ &= \frac{1}{1+c} \cdot f_2(x_2; \bullet) \\ &= f^*\big((2, x_2); \bullet\big). \end{aligned}$$

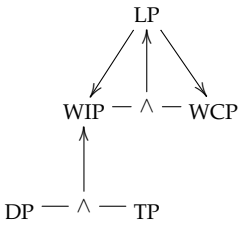
Then the WIP implies that

$$\mathrm{Ev}\left(\mathcal{E}^*, (1, x_1)\right) = \mathrm{Ev}\left(\mathcal{E}^*, (2, x_2)\right).$$

Finally, apply the WCP to each side to infer that

$$\mathrm{Ev}(\mathcal{E}_1, x_1) = \mathrm{Ev}(\mathcal{E}_2, x_2),$$

which is the LP. □



Again, to be clear about the logic: either I accept the LP, or I explain which of the two principles, WIP and WCP, I refute. To me, the WIP is the implication of two principles that are self-evident, and the WCP is itself self-evident, so I must accept the LP, or else invoke and justify an *ad hoc* abandonment of logic.

2.4 *A stronger form of the WCP*

The new concept in this section is 'ancillarity'. This has several different definitions in the statistics literature; mine is close to that of Cox and Hinkley (1974, sec. 2.2).

Definition 2.8 (Ancillary)**.** X is ancillary in experiment

$$\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Omega_1 \times \Omega_2, f_{X,Y}\}$$

exactly when $f_{X,Y}$ factorizes as

$$f_{X,Y}(x, y; \theta) = f_X(x) \cdot f_{Y|X}(y \mid x; \theta).$$

In other words, the marginal distribution of X is completely specified. Not all families of distributions will factorize in this way, but when they do, there are new possibilities for inference, based around stronger forms of the WCP such as the CP immediately below.[6]

When X is ancillary, we can consider the conditional experiment

$$\mathcal{E}^{Y|x} = \{\mathcal{Y}, \Omega, f_{Y|x}\}, \tag{2.4}$$

where $f_{Y|x}(y; \theta) := f_{Y|X}(y \mid x; \theta)$. This is an experiment where we condition on $X = x$, i.e. treat X as known, and treat Y as the only random quantity. This is an attractive idea if we can specify $f_{Y|X}$, because we can then disregard the choice of f_X. Our aspiration is the following principle.

[6] Here I am going to include situations where f_X depends on parameters which are not interesting and which do not appear in $f_{Y|X}$. Technically, an additional self-evident 'sure thing principle' would also be required in this situation.

Definition 2.9 (Conditionality Principle, CP)**.** If X is ancillary in $\mathcal{E}$, then $\text{Ev}\left(\mathcal{E}, (x, y)\right) = \text{Ev}(\mathcal{E}^{Y|x}, y)$.

Here is a example which will be familiar to all statisticians. We have been given a sample $x = (x_1, \ldots, x_n)$ to evaluate. In fact, n itself is likely to be the outcome of a random variable N, because the process of sampling itself is rather uncertain. But we seldom concern ourselves with the distribution of N when we evaluate x; instead we treat N as known. Equivalently, we treat N as ancillary and condition on $N = n$, which would be justified by the CP.

Here is another familiar example. A regression of Y on X appears to make a distinction between the 'dependent variable' Y and the 'covariates' X, with only the former being treated as random. This distinction is insupportable, given that the roles of Y and X are often interchangeable, and determined by the *hypothèse du jour*. What we are actually doing is treating X as ancillary and conditioning on X, which would be justified by the CP.

* * *

Clearly the CP implies the WCP, with the experiment indicator $I \in \{1, 2\}$ being ancillary. But what justification might we have for accepting the CP? Happily the CP comes for free with the LP.

Theorem 2.10. *LP* $\rightarrow$ *CP*.

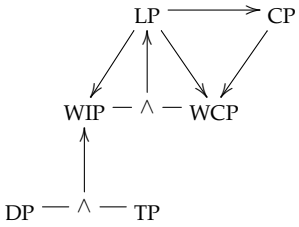
Proof. Suppose that X is ancillary in $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Omega, f_{X,Y}\}$. Thus

$$f_{X,Y}(x, y; \bullet) = f_X(x) \cdot f_{Y|X}(y \mid x; \bullet) = c(x) \cdot f_{Y|x}(y; \bullet),$$

where $c > 0$. Then the LP implies that

$$\text{Ev}\left(\mathcal{E}, (x, y)\right) = \text{Ev}(\mathcal{E}^{Y|x}, y),$$

which is the CP. □



2.5 *Stopping rules*

Consider a sequence of random quantities $X_1, X_2, \ldots$ with marginal PMFs

$$f_n(x_1, \ldots, x_n; \theta) \qquad n = 1, 2, \ldots,$$

where consistency requires that

$$f_n(x_1, \ldots, x_n; \theta) = \sum_{y_1} \cdots \sum_{y_m} f_{n+m}(x_1, \ldots, x_n, y_1, \ldots y_m; \theta)$$

for each $n, m \in 1, 2, \ldots$.[7] In a sequential experiment, the number of X's that are observed is not fixed in advanced but depends on the values seen so far. That is, at time j, the decision to observe X_{j+1} can be modelled by a probability $p_j(x_1, \ldots, x_j)$. We can assume, resources being finite, that the experiment must stop at specified time m, if it has not stopped already, hence $p_m(x_1, \ldots, x_m) = 0$. Denote the stopping rule as $\tau := (p_1, \ldots, p_m)$. Let's aspire to something really striking.

[7] This is Kolmogorov's consistency condition.

Definition 2.11 (Stopping Rule Principle, SRP)**.** In a sequential experiment $\mathcal{E}^\tau$, $\mathrm{Ev}\left(\mathcal{E}^\tau, (x_1, \ldots, x_n)\right)$ does not depend on the stopping rule τ.

The SRP is nothing short of revolutionary, if it is accepted. It implies that that the intentions of the experimenter, represented by τ, are irrelevant for making inferences about Θ, once the observations $(x_1, \ldots, x_n)$ are available. Thus the statistician could proceed as though the simplest possible stopping rule were in effect, which is $p_1 = \cdots = p_{n-1} = 1$ and $p_n = 0$, an experiment with n fixed in advance. Obviously it would be liberating for the statistician to put aside the experimenter's intentions (since they may not be known and could be highly subjective). And in fact when we are given a sample $(x_1, \ldots, x_n)$ we seldom enquire about the experimenter's intentions and try to discover her stopping rule—so no doubt that something like the SRP is ubiquitous in practice. But can it possibly be justified? Indeed it can.[8]

[8] I think this is one of the most beautiful theoretical results in the whole of Statistics.

Theorem 2.12. *LP $\rightarrow$ SRP.*

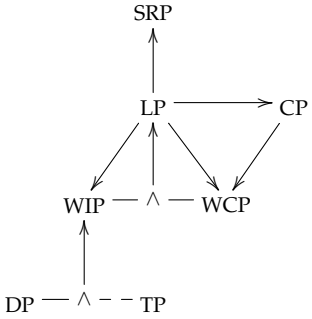
Proof. Let τ be an arbitrary stopping rule, and consider the outcome $(x_1, \ldots, x_n)$, which I will write as $x_{1:n}$ for convenience. The probability of this outcome under τ is

$$\begin{aligned} & f_\tau(x_{1:n}; \theta) \\ &= f_1(x_1; \theta) \cdot \prod_{j=1}^{n-1} p_j(x_{1:j})\, f_{j+1}(x_{j+1} \mid x_{1:j}; \theta) \cdot \big(1 - p_n(x_{1:n})\big) \\ &= \prod_{j=1}^{n-1} p_j(x_{1:j}) \cdot \big(1 - p_n(x_{1:n})\big) \times f_1(x_1; \theta) \prod_{j=2}^{n} f_j(x_j \mid x_{1:(j-1)}; \theta) \\ &= \prod_{j=1}^{n-1} p_j(x_{1:j}) \cdot \big(1 - p_n(x_{1:n})\big) \times f_n(x_{1:n}; \theta). \end{aligned}$$

Now observe that this equation has the form

$$f_\tau(x_{1:n}; \bullet) = c(x_{1:n}) \cdot f_n(x_{1:n}; \bullet) \qquad c > 0. \tag{†}$$

Thus the LP implies that $\mathrm{Ev}(\mathcal{E}^\tau, x_{1:n}) = \mathrm{Ev}(\mathcal{E}^n, x_{1:n})$ where $\mathcal{E}^n := \{\mathcal{X}^n, \Omega, f_n\}$. Since the choice of stopping rule was arbitrary, (†) holds for all stopping rules, showing that the choice of stopping rule is irrelevant. □



To illustrate the SRP, consider the following example from Basu (1975, p. 42). Four different coin-tossing experiments (with some finite limit on the number of tosses) have the same outcome $x =$ (T,H,T,T,H,H,T,H,H,H):

$\mathcal{E}_1$ Toss the coin exactly 10 times;

$\mathcal{E}_2$ Continue tossing until 6 heads appear;

$\mathcal{E}_3$ Continue tossing until 3 consecutive heads appear;

$\mathcal{E}_4$ Continue tossing until the accumulated number of heads exceeds that of tails by exactly 2.

One could easily adduce more sequential experiments which

gave the same outcome. According to the SRP, the evidence for the probability of heads is the same in every case. Once the sequence of heads and tails is known, the intentions of the original experimenter (i.e. the experiment she thought she was doing) are immaterial to inference about the probability of heads, and the simplest experiment $\mathcal{E}_1$ can be used for inference.

* * *

The Stopping Rule Principle has become enshrined in our profession's collective memory due to this iconic comment from L.J. Savage, one of the great statisticians of the 20th century:

> May I digress to say publicly that I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right. (Savage et al., 1962, p. 76)

This comment captures the revolutionary and transformative nature of the SRP.

2.6 *The Likelihood Principle in practice*

Now we should pause for breath, and ask the obvious questions: is the LP vacuous? Or trivial? In other words, Is there any inferential approach which respects it? Or do all inferential approaches respect it? With apologies to the few proponents of likelihood-based inference,[9] I will focus on Frequentist and Bayesian approaches, as outlined in Chapter 1. In brief, the Bayesian approach satisfies the LP, and so the LP is not vacuous. And the Frequentist approach does not satisfy the LP, and so the LP is not trivial. The proof that the Bayesian approach satisfies the LP will be given in Theorem 3.3. Here I concentrate on the Frequentist approach.

[9] Mainly philosophers and physicists.

Theorem 2.13. *Suppose that* $Ev(\mathcal{E}, x)$ *depends on the value of* $f(x'; \bullet)$ *for some* $x' \neq x$. *Then* Ev *does not respect the LP.*

Proof. Let $\mathcal{E} = \{\mathcal{X}, \Omega, f\}$ and let $x'' \neq x, x'$. Define $\mathcal{E}_1 = \{\mathcal{X}, \Omega, f_1\}$, where

$$f_1(x'; \bullet) = f(x''; \bullet)$$
$$f_1(x''; \bullet) = f(x'; \bullet)$$

and $f_1 = f$ elsewhere. Then

$$f(x; \bullet) = f_1(x; \bullet)$$

but $f(x'; \bullet) \neq f_1(x'; \bullet)$ and so

$$\mathrm{Ev}(\mathcal{E}, x) \neq \mathrm{Ev}(\mathcal{E}_1, x)$$

violating the LP. □

In the Frequentist approach, algorithms are certified in terms of their sampling distributions, and selected on the basis of their

certification, as defined in Definition 1.1. Theorem 2.13 shows that Frequentist inference does not respect the LP, because the sampling distribution of the algorithm depends on values for f other than $f(x;\bullet)$.

The two main difficulties with violating the LP are:

1. To reject the LP is to reject at least one of the WIP and the WCP. Yet both of these principles seem self-evident. Therefore violating the LP is either illogical or obtuse.

2. In their everyday practice, statisticians use the CP (treating some variables as ancillary) and the SRP (ignoring the intentions of the experimenter). Neither of these is self-evident, but both are implied by the LP. If the LP is violated, then they both need an alternative justification.

Alternative formal justifications for the CP and the SRP have not been forthcoming.

2.7 *Reflections*

The statistician takes delivery of an outcome x. Her standard practice, as mandated by our profession, is to assumes the truth of a statistical model $\mathcal{E}$, and then turn $(\mathcal{E}, x)$ into an inference about the true value of the parameter Θ. As remarked several times already (see Chapter 1), this is *not* the end of her involvement, but it is a key step, which may be repeated several times, under different notions of the outcome and different statistical models. This chapter concerns this key step: how she turns $(\mathcal{E}, x)$ into an inference about Θ.

Whatever inference is required, we assume that the statistician applies an algorithm to $(\mathcal{E}, x)$. In other words, her inference about Θ is not arbitrary, but transparent and reproducible—this is hardly controversial, because anything else would be non-scientific. Following Birnbaum, the algorithm is denoted 'Ev'. The question now becomes: how does she choose her 'Ev'?

This chapter does not explain how to choose 'Ev'; instead it describes some properties that 'Ev' might have. Some of these properties are self-evident, and to violate them would be hard to justify to an auditor. These properties are the DP (Definition 2.2), TP (Definition 2.3), and WCP (Definition 2.5). Other properties are not at all self-evident; the most important of these are the LP (Definition 2.6), the CP (Definition 2.9), and the SRP (Definition 2.11). These not-self-evident properties would be extremely attractive, were it possible to justify them. And as we have seen, they can all be justified as logical deductions from the properties that are self-evident. This is the essence of Birnbaum's Theorem (Theorem 2.7).

For over a century, statisticians have been proposing methods for selecting algorithms for 'Ev', independently of this strand of research concerning the properties that such algorithms ought

to have (remember that Birbaum's Theorem was published in 1962). Bayesian inference, which turns out to respect the LP, is compatible with all of the properties given above, but Frequentist inference, which turns out to violate the LP, is not. The two main consequences of this violation are described in Section 2.6.

Now it is important to be clear about one thing. Ultimately, an inference is a single element in the space of 'possible inferences about Θ'. An inference cannot be evaluated according to whether or not it satsfies the LP. What is being evaluated in this chapter is the algorithm, the mechanism by which $\mathcal{E}$ and x are turned into an inference. It is quite possible that statisticians of quite different persuasions will produce effectively identical inferences from different algorithms. For example, if asked for a set estimate of Θ, a Bayesian statistician might produce a 95% High Density Region, and a Frequentist statistician a 95% confidence set, but they might be effectively the same set. But it is not the inference that is the primary concern of the auditor: it is the justification for the inference, among the uncountable other inferences that might have been made but weren't. The auditor checks the 'why', before passing the 'what' on to the client.

So the auditor will ask: why do you choose algorithm 'Ev'? The Frequentist statistician will reply, "Because it is a 95% confidence procedure for Θ, and, among the uncountable number of such procedures, this is a good choice [for some reasons that are then given]." The Bayesian statistician will reply "Because it is a 95% High Posterior Density region for Θ for prior distribution π, and among the uncountable number of prior distributions, π is a good choice [for some reasons that are then given]." Let's assume that the reasons are compelling, in both cases. The auditor has a follow-up question for the Frequentist but not for the Bayesian: "Why are you not concerned about violating the Likelihood Principle?" A well-informed auditor will know the theory of the previous sections, and the consequences of violating the LP that are given in Section 2.6. For example, violating the LP is either illogical or obtuse—neither of these properties are desirable in an applied statistician.

To be frank I do not have a good answer to this question, which is why I would choose *not* to violate the LP, in the way that I choose 'Ev'. However, in the spirit of fair play I will suggest two possibilities.[10]

First, the Frequentist might reply, "Because this is how we do things in (say) Social Psychology", i.e. an appeal to current practice. This answer is contrary to the scientific norm of scepticism, and may upset the client, who thought he was paying for a scientist. The counter-argument is that 'science is what scientists do', which is a naturalistic as opposed to normative view of science (see, e.g., Ziman, 2000). Under the naturalistic view, violating the LP is scientific as long as it is the standard practice among the *soi-disant* scientists in Social Psychology. Personally, I don't think this excuses

[10] Another possibility to add to these two might be "I'm not interested in principles, I let the data speak for itself." This person would suit a client who wanted an illogical and unprincipled data analyst; or "reckless and treacherous", according to Alfred Marshall, writing in 1885 (Stigler, 2016, p. 202). If you are this person, you can probably charge a lot of money.

these scientists from having a compelling reason for violating the LP (e.g., explaining why they are neither illogical nor obtuse). But apparently most Social Psychologists disagree with me, or else they are ignorant of the LP and its implications.

Second, the Frequentist might reply "Because it is important to me that I control my error rate over the course of my career", which is incompatible with the LP. In other words, the statistician ensures that, by always using a 95% confidence procedure, the true value of Θ will be inside at least 95% of her confidence sets, over her career. This is a very interesting answer, revealing the statistician's egocentricity in putting her career error rate before the needs of her current client. I can just about imagine a client demanding "I want a statistician who is right at least 95% of the time". Personally, though, I would advise a client against this, and favour instead a statistician who is concerned not with her career error rate, but rather with the client's particular problem.

3
Statistical Decision Theory

3.1 Introduction

The basic premise of Statistical Decision Theory is that we want to make inferences about the parameter of a family of distributions (see section 1.3). So the starting point of this chapter is a model for the observables $Y \in \mathcal{Y}$ of the general form

$$\mathcal{E} = \{\mathcal{Y}, \Omega, f\},$$

just as in chapter 1 and chapter 2. The value $f(y;\theta)$ denotes the probability that $Y = y$ under family member $\theta \in \Omega$, where θ is the parameter, and Ω is the parameter space. I will stick with my convention that $\mathcal{Y}$ is countable and Ω is uncountably infinite. I will assume throughout this chapter that $f(y;\theta)$ is easy to evaluate (see section 1.2).

We accept as our working hypothesis that $\mathcal{E}$ is true (see section 1.1), so that inference is learning about Θ, the true value of the parameter. More precisely, we would like to understand how to construct the 'Ev' function from chapter 2, in such a way that it reflects our needs, which will vary from application to application. Statistical Decision Theory allows us to select an 'Ev' which is suitable for the type of inference we want to make, and which reflects the consequence of making a poor inference.

The set of possible inferences is termed the *action set*, $\mathcal{A}$, with typical element a. The consequence of making a poor inference is specified as the *loss function* $L : \mathcal{A} \times \Omega \to \mathbb{R}$, with larger values indicating worse consequences. Thus $L(a,\theta)$ is the loss incurred by the statistician (or her client) if action a is taken and Θ turns out to be θ. I will assume, as is natural, that L is bounded, but many results below also hold in the more general case.

Before making her choice of action, the statistician will observe y, a value for Y. Her choice should be some function of the value y, and this is represented as a *decision rule*, $\delta : \mathcal{Y} \to \mathcal{A}$. As we are taking the model $\mathcal{E}$ as given, $\delta(y)$ in this chapter is the analogue of $\mathrm{Ev}(\mathcal{E}, y)$ from chapter 2.

The three main types of inference about Θ are (i) point estimation, (ii) set estimation, and (iii) hypothesis testing. It is a great conceptual and practical simplification that Statistical Decision

Theory distinguishes between these three types simply according to their action sets, which are:

Type of inference	Action set $\mathcal{A}$
Point estimation	The parameter space, Ω. See section 3.5.
Set estimation	The set of all subsets of Ω, denoted 2^Ω. See section 3.6.
Hypothesis testing	A specified partition of Ω, denoted $\mathcal{H}$ below. See section 3.7.

All three of these types of inference are easily adapted to specified functions of Θ, say $g(\Theta)$. Thus point estimation would have $\mathcal{A} = g\Omega$; set estimation would have $\mathcal{A} = 2^{g\Omega}$, and hypothesis testing would have $\mathcal{A} =$ some partition of $g\Omega$. For example, if $\theta = (\theta_1, \theta_2)$ but θ_2 is nuisance parameter, then $g(\theta) = \theta_1$. In point estimation, $\mathcal{A} = \Omega_1$, and $L(a, \theta) = L_1(a, \theta_1)$, where θ_1 is the value of Θ_1, and $a \in \Omega_1$ is the point estimate of Θ_1.

The next three sections develop some general results for Statistical Decision Theory, applicable to all types of inference, and then the later sections consider each of the three types in more detail.

3.2 *Bayes rules*

In a Bayesian approach, Θ is treated as a random variable, and the model $\mathcal{E}$ is augmented by a prior probability density function (PDF) π, for which $\mathbb{P}(\Theta \in S) = \int_{\theta \in S} \pi(\theta)\, \mathrm{d}\theta$ for any well-behaved $S \subset \Omega$; see section 1.5. I will write the joint distribution of (Y, Θ) as

$$\mathrm{p}(y, \theta) = f(y; \theta)\, \pi(\theta).$$

From this joint distribution, we can also calculate, as needed, the marginal distribution $\mathrm{p}(y)$ and the posterior distribution $\mathrm{p}(\theta \mid y)$; the latter using Bayes's theorem.

Definition 3.1 (Bayes rule)**.** Let $\mathcal{D}$ be the set of all possible decision rules. The decision rule δ^* is a *Bayes rule* exactly when

$$\mathbb{E}\{L(\delta^*(Y), \Theta)\} \leq \mathbb{E}\{L(\delta(Y), \Theta)\}$$

for all $\delta \in \mathcal{D}$.

The value $\mathbb{E}\{L(\delta(Y), \Theta)\}$ is termed the *Bayes risk* of decision rule δ, and is always well-defined under the condition that L is bounded. Therefore a Bayes rule is any decision rule which minimizes the Bayes risk, for some action set, loss function, model, and prior distribution. There is a justly famous result which gives the explicit form for a Bayes rule.

Theorem 3.2 (Bayes Rule Theorem, BRT)**.** *If $\mathcal{A}$ is finite, then a Bayes rule exists*[1] *and satisfies $\delta^* = \tilde{\delta}$, where*

$$\tilde{\delta}(y) := \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}\{L(a, \Theta) \mid Y = y\}. \tag{3.1}$$

[1] Finiteness of $\mathcal{A}$ ensures existence. Similar but more general results are possible, but they require tedious and distracting topological conditions to ensure that a minimum obtains within $\mathcal{D}$.

Proof. I will show that $\mathbb{E}\{L(\delta(Y),\Theta)\} \geq \mathbb{E}\{L(\tilde{\delta}(Y),\Theta)\}$ for all $\delta \in \mathcal{D}$; i.e. that $\tilde{\delta}$ minimises the Bayes risk. Let δ be arbitrary. Then

$$\begin{aligned}
\mathbb{E}\{L(\delta(Y),\Theta)\} &= \int \sum_y L(\delta(y),\theta)\ \mathrm{p}(y,\theta)\,\mathrm{d}\theta \\
&= \sum_y \int L(\delta(y),\theta)\,\mathrm{p}(\theta \mid y)\,\mathrm{d}\theta \cdot \mathrm{p}(y) \\
&\geq \sum_y \min_a \left\{ \int L(a,\theta)\,\mathrm{p}(\theta \mid y)\,\mathrm{d}\theta \right\} \cdot \mathrm{p}(y) \quad \text{as } \mathrm{p}(y) \geq 0 \\
&= \sum_y \int L(\tilde{\delta}(y),\theta)\,\mathrm{p}(\theta \mid y)\,\mathrm{d}\theta \cdot \mathrm{p}(y) \qquad \text{from (3.1)} \\
&= \int \sum_y L(\tilde{\delta}(y),\theta)\ \mathrm{p}(y,\theta)\,\mathrm{d}\theta \\
&= \mathbb{E}\{L(\tilde{\delta}(Y),\Theta)\}. \qquad \square
\end{aligned}$$

This astounding result indicates that the minimization of expected loss over the space of all functions from $\mathcal{Y}$ to $\mathcal{A}$ can be achieved by the pointwise minimization over $\mathcal{A}$ of the expected loss conditional on $Y = y$. It converts an apparently intractable problem into a simple one.

The next result will not be a surprise for those who have read chapter 2.

Theorem 3.3. *Bayes rules respect the Likelihood Principle (LP, see Theorem 2.6).*

Proof. Let $\mathcal{E}_1 = \{\mathcal{Y}_1, \Omega, f_1\}$ and $\mathcal{E}_2 = \{\mathcal{Y}_2, \Omega, f_2\}$ be different models with the same parameter Θ. Because they have the same parameter, they have the same prior distribution π. By Bayes's theorem,

$$\begin{aligned}
\mathrm{p}_1(\theta \mid y_1) &\propto f_1(y_1;\theta)\,\pi(\theta) \\
\mathrm{p}_2(\theta \mid y_2) &\propto f_2(y_2;\theta)\,\pi(\theta)
\end{aligned}$$

where the missing multiplicative constants are $\mathrm{p}_1(y_1)^{-1}$ and $\mathrm{p}_2(y_2)^{-1}$, respectively. Now suppose that y_1, y_2 satisfy

$$f_1(y_1;\bullet) = c(y_1,y_2) \cdot f_2(y_2;\bullet),$$

as in the condition for the LP. I will show that this implies $\delta_1^*(y_1) = \delta_2^*(y_2)$, as required by the LP. By the Bayes Rule Theorem (Theorem 3.2),

$$\begin{aligned}
\delta_1^*(y_1) &= \operatorname*{argmin}_a \mathbb{E}_1\{L(a,\Theta) \mid Y_1 = y_1\} \\
&= \operatorname*{argmin}_a \int L(a,\theta) \cdot f_1(y_1;\theta)\,\pi(\theta)\,\mathrm{d}\theta \\
&= \operatorname*{argmin}_a \int L(a,\theta) \cdot c(y_1,y_2)\,f_2(y_2;\theta)\,\pi(\theta)\,\mathrm{d}\theta \\
&= \operatorname*{argmin}_a \int L(a,\theta) \cdot f_2(y_2;\theta)\,\pi(\theta)\,\mathrm{d}\theta \\
&= \operatorname*{argmin}_a \mathbb{E}_2\{L(a,\Theta) \mid Y_2 = y_2\} \\
&= \delta_2^*(y_2). \qquad \square
\end{aligned}$$

To hark back to the analysis in chapter 2, if your inference (i.e. your decision rule) does not respect the LP then you are either illogical or obtuse—please excuse me for being blunt. So Theorem 3.3 is

a good reason for selecting a Bayes rule as your decision rule. You can also be sure that your decision rule respects the Conditionality Principle (CP, Theorem 2.9) and the Stopping Rule Principle (SRP, Theorem 2.11). To assert the contrapositive, if your decision rule does not respect the LP, CP, and SRP, then it cannot be a Bayes rule.

3.3 *Admissible rules*

As discussed in section 1.4, Frequentist statisticians are averse to prior distributions. But it is not possible to construct Bayes rules without them, and so Frequentist statisticians need another approach to selecting their decision rule for some action set, loss function, and model.

The accepted approach is to narrow the set of possible decision rules by ruling out those that are obviously bad. Define the *risk function* for rule δ as

$$R(\delta, \theta) := \mathbb{E}\{L(\delta(Y), \theta); \theta\} = \sum_y L(\delta(y), \theta)\, f(y; \theta). \tag{3.2}$$

That is, $R(\delta, \theta)$ is the expected loss from rule δ in family member θ. A decision rule δ *dominates* another rule δ' exactly when

$$R(\delta, \theta) \leq R(\delta', \theta) \quad \text{for all } \theta \in \Omega,$$

with a strict inequality for at least one $\theta \in \Omega$. If you had both δ and δ', you would never want to use δ'.[2] A decison rule is *admissible* exactly when it is not dominated by any other rule; otherwise it is *inadmissible*. So the accepted approach is to reduce the set of possible decision rules under consideration by only using admissible rules.

[2] Here I am assuming that all other considerations are the same in the two cases: e.g. $\delta(y)$ and $\delta'(y)$ take about the same amount of resource to compute.

It is hard to disagree with this approach, although one wonders how big the set of admissible rules will be, and how easy it is to enumerate the set of admissible rules in order to choose between them. It turns out that this issue has a clear-cut answer.

Theorem 3.4 (Wald's Complete Class Theorem, CCT)**.** *Let $\mathcal{E} = \{\mathcal{Y}, \Omega, f\}$, $\mathcal{A}$, and L be given. In the case where Ω is finite, a decision rule δ is admissible* if and only if *it is a Bayes rule for some positive prior distribution π.*

The proof is given in section 3.4. There are generalisations of this theorem to non-finite and uncountable Ω; however, the results are highly technical. See Ferguson (1967, ch. 2), Schervish (1995, ch. 3), Berger (1985, chs 4, 8), and Ghosh and Meeden (1997, ch. 2) for more details and references to the original literature. In the rest of this section, I will assume the more general result, which is that a decision rule is admissible if and only if it is a Bayes rule, which holds for practical purposes.

So what does the CCT say? First of all, admissible decision rules respect the LP. This follows from the fact that admissible rules are Bayes rules, and Bayes rules respect the LP, by Theorem 3.3.

Insofar as we think respecting the LP is a good thing, this provides support for using admissible decision rules, because we cannot be certain that inadmissible rules respect the LP. Second, if you select a Bayes rule according to some positive prior distribution π then you cannot ever choose an inadmissible decision rule. So the CCT states that there is a very simple way to protect yourself from choosing an inadmissible decision rule.

But here is where you must pay close attention to logic. Suppose that δ' is inadmissible and δ is admissible. It does not follow that δ dominates δ'. So just knowing of an admissible rule does not mean that you should abandon your inadmissible rule δ'. You can argue that although you know that δ' is inadmissible, you do not know of a rule which dominates it. All you know, from the CCT, is the family of rules within which the dominating rule must live: it will be a Bayes rule for some positive π. Statisticians sometimes use inadmissible rules. They can argue that yes, their rule δ is or may be inadmissible, which is unfortunate, but since the identity of the dominating rule is not known, it is not wrong to go on using δ. Do not attempt to explore this rather arcane line of reasoning with your client!

3.4 *The Complete Class Theorem*

This section can be skipped once the previous section has been read. It proves a very powerful result, Theorem 3.4 above, originally due to an iconic figure in Statistics, Abraham Wald.[3] The parameter space is assumed to be finite, so write it as

$$\Omega = \{\theta_1, \ldots, \theta_k\}.$$

[3] For his tragic story, see https://en.wikipedia.org/wiki/Abraham_Wald.

Denote the available decision rules as δ_i, for $i = 1, 2, \ldots$; I am assuming that the set of rules is countable, but this is without loss of generality (we will shortly create an uncountable number of decision rules). For each decision rule, define the risk function as

$$R_{ij} := \mathbb{E}\{L(\delta_i(Y), \theta_j); \theta_j\} \quad \begin{cases} i = 1, 2, \ldots \\ j = 1, \ldots, k. \end{cases}$$

Thus R_{ij} is the expected loss for rule δ_i under parameter value θ_j.

I will give a blackboard proof for $k = 2$ which generalises to any finite k. Call $\delta_1, \delta_2, \ldots$ the 'pure' rules, and $R_1, R_2, \ldots$ the pure risks, where $R_i = (R_{i1}, \ldots, R_{ik})$. Panel (a) in Figure 3.4 shows a set of pure risks when $k = 2$.

We must widen the set of available decision rules, to include rules selected randomly from the pure rules according to probabilities $w = (w_1, w_2, \ldots)$. This is because a rule δ_i might not be dominated by a pure rule but it might be dominated by a randomised rule; see Figure 3.1. Let $\mathbb{P}(I = i) = w_i$. Then the risk of randomised rule w is

$$R_{wj} = \mathbb{E}\left\{L(\delta_I(Y), \theta_j); \theta_j\right\} = \sum_i R_{ij} \cdot w_i,$$

by the Law of Iterated Expectation (LIE). The set of all rules, pure and randomised, is termed the *risk set*, and it is the convex hull of $\{R_1, R_2, \dots\}$. Every point in the risk set is an attainable risk, for a suitable choice of w. See Panel (b) of Figure 3.4. From now on, we can refer to 'risks' rather than 'rules'.

Now consider the subset of the risk set which is admissible. A risk is dominated if there is another risk in its 'southwest' quadrant. So the only admissible risks in the risk set are on the southwest boundary, shown in Panel (c) of Figure 3.4. We have identified the set of admissible risks: the pure risks on the southwest boundary, and the randomised risks which lie on the facets between the pure risks.

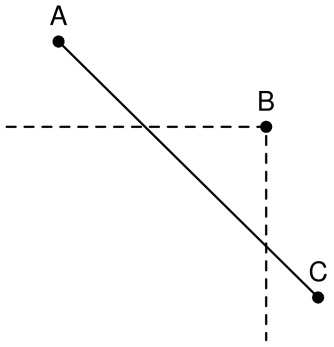


Figure 3.1: Rule B is not dominated by either A or δ, but it is dominated by some randomised rules based on A and δ, notably those with risks that lie in the facet between A and δ within the dashed lines.

Now I show that this set of admissible risks is identical to the set of risks for Bayes rules for some positive prior probability. Fix $\pi = (\pi_1, 1-\pi_1)$ with $0 < \pi_1 < 1$, and consider the set of risks with a specified Bayes risk a, i.e. the values (r_1, r_2) for which

$$\begin{aligned}
a &= \mathbb{E}\{L(\delta(Y), \Theta)\} && \text{defn of Bayes risk} \\
&= \mathbb{E}\left[\mathbb{E}\{L(\delta(Y), \Theta) \mid \Theta\}\right] && \text{by the LIE} \\
&= \mathbb{E}\{R(\delta, \Theta)\} && \text{defn of risk function} \\
&= \sum_{j=1}^{k} R(\delta, \theta_j) \cdot \pi_j && \Omega \text{ finite} \\
&= r_1 \cdot \pi_1 + r_2 \cdot (1-\pi_1) && \text{for } k = 2.
\end{aligned}$$

On the panels in Figure 3.4, this is a straight line with equation

$$r_2 = \frac{a}{1-\pi_1} + \frac{-\pi_1}{1-\pi_1} r_1.$$

This line may pass below the risk set, in which case there is no attainable risk which has Bayes risk of a. So increase a until the line just touches the risk set, at risk $B(\pi)$ with Bayes risk b; see Panel (d) in Figure 3.4. $B(\pi)$ is the attainable risk which achieves the minimum Bayes risk for π, i.e. it is the risk of the Bayes rule for π. Varying π in the open interval $(0, 1)$ and repeating the exercise shows that the set of admissible risks and the set of risks for Bayes rules with positive prior probability are identical.

This proof generalises to any finite k according to the Supporting Hyperplane Theorem; see, e.g., Ferguson (1967, ch. 2) or Schervish (1995, ch. 3).

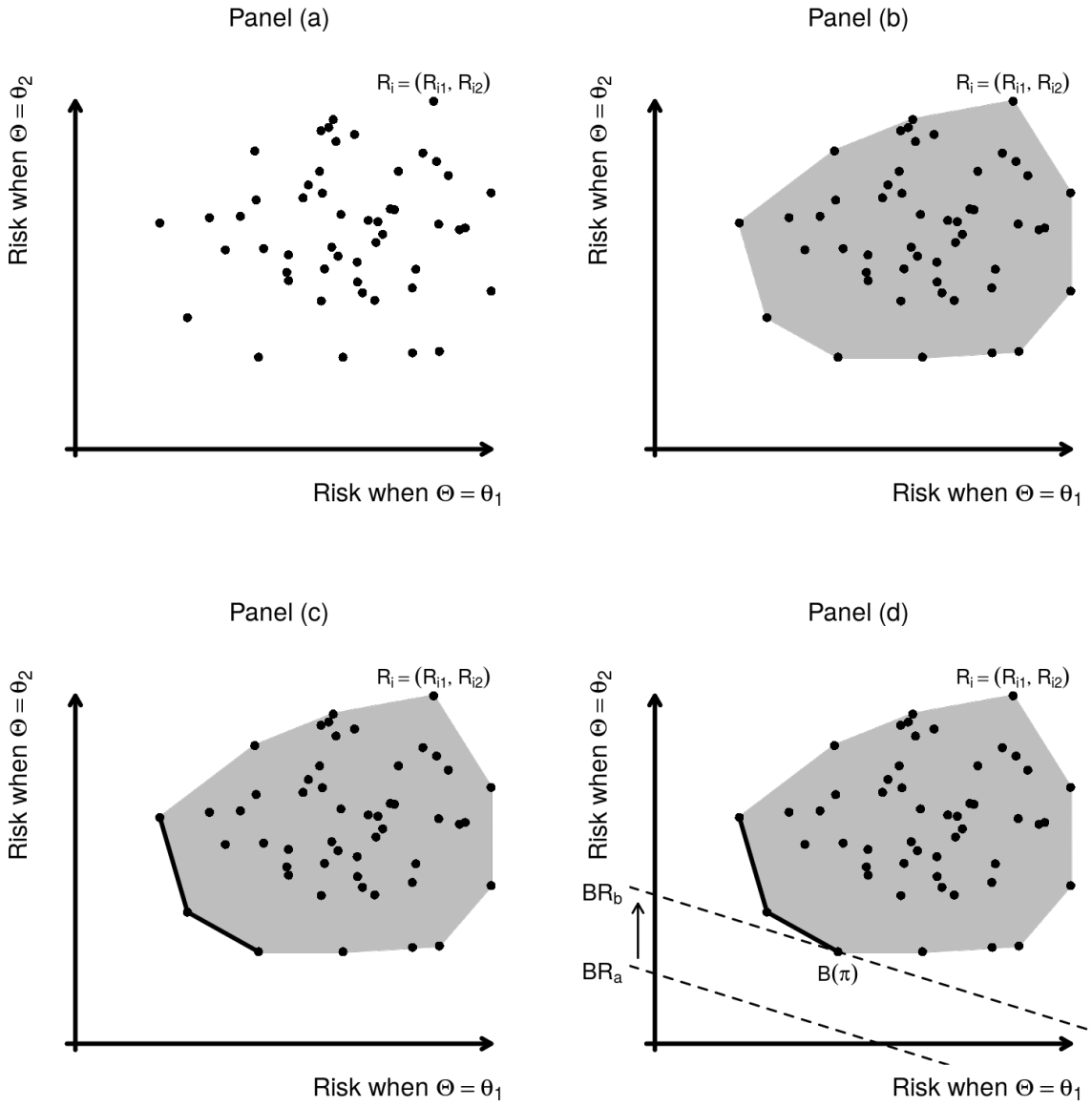


Figure 3.4. Blackboard proof of Theorem 3.4, with $\Omega = \{\theta_1, \theta_2\}$. Panel (a). The risks for a set of pure decision rules. Panel (b). The risk set: the convex hull of the pure risks, showing all risks that are attainable using randomised rules. Panel (c). The set of admissible risks is shown with a thick line. Panel (d). The dashed line 'BR_a' shows the set of risks which have Bayes risk a, for fixed probabilities $\pi = (\pi_1, 1 - \pi_1)$, where $0 < \pi_1 < 1$. None of the risks on BR_a are attainable. By increasing the Bayes risk to b, admissible pure risk $B(\pi)$ becomes attainable. $B(\pi)$ is the Bayes rule for π. Changing π changes the gradient of the dashed line, but it always just touches the set of attainable risks on the set of admissible risks.

3.5 *Point estimation*

For point estimation the action space is $\mathcal{A} = \Omega$, and the loss function $L(a, \theta)$ represents the (negative) consequence of choosing a as a point estimate of Θ, when in fact $\Theta = \theta$. A point estimate of Θ is often termed a *point prediction*, or just 'prediction'.

There will be situations where a function $L : \Omega \times \Omega \to \mathbb{R}$ is fairly easy to specify. Fox example, consider the Netflix challenge.[4] Netflix wants to make a prediction $a \in \Omega = \{1, 2, 3, 4, 5\}$ for a film that a client has not seen yet, but who will rate the film as Θ. Netflix suffers a reputational loss (which may lead to revenue loss) when a recommended film is rated below 5 by the client. But in fact Netflix will only recommend films that it predicts will be 5's, and so its loss function is something like

$$L(a, \theta) = \begin{cases} \epsilon \cdot (5 - a) & a = 1, 2, 3, 4 \\ a - \theta & a = 5 \end{cases}$$

where ϵ, which is a small positive value, is there to reflect that Netflix wants to make recommendations. In the Netflix challenge, the actual loss function was $L(a, \theta) = (a - \theta)^2$, which either goes to show that the people at Netflix are not very bright or, perhaps more likely, that the entire challenge was in fact a marketing exercise.

In many cases, however, specifying the loss function presents a challenge. Hence the need for a generic loss function which is acceptable over a wide range of situations. A natural choice in the very common case where Ω is a convex subset of $\mathbb{R}^d$ is a *convex loss function*,

$$L(a, \theta) = h(a - \theta) \tag{3.3}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is a smooth non-negative convex function with $h(\mathbf{0}) = 0$. This type of loss function asserts that small errors are much more tolerable than large ones. One possible further restriction would be that h is an even function.[5] This would assert that under-prediction incurs the same loss as over-prediction. There are many situations where an even function is *not* appropriate, but in these cases a generic loss function should be replaced by a more specific one.[6]

Proceeding further along the same lines, an even, differentiable and strictly convex loss function can be approximated by a *quadratic loss function*,

$$h(x) \propto x^T Q\, x \tag{3.4}$$

where Q is a symmetric positive-definite $d \times d$ matrix. This follows directly from a Taylor series expansion of h around $\mathbf{0}$:

$$h(x) = 0 + 0 + \tfrac{1}{2} x^T \nabla^2 h(\mathbf{0})\, x + 0 + O(\|x\|^4)$$

where the first 0 is because $h(\mathbf{0}) = 0$, the second 0 is because $\nabla h(\mathbf{0}) = 0$ since h is minimized at $x = \mathbf{0}$, and the third 0 is because

[4] See https://en.wikipedia.org/wiki/Netflix_Prize.

[5] I.e. $h(x) = h(-x)$.

[6] See, e.g., Milner and Rougier (2014), on predicting the weights of donkeys.

h is an even function. $\nabla^2 h$ is the *hessian matrix* of second derivatives, and it is symmetric by construction, and positive definite at $x = \mathbf{0}$, if h is strictly convex and minimized at $\mathbf{0}$.

In the absence of anything more specific the quadratic loss function is the generic loss function for point estimation. Hence the following result is widely applicable.

Theorem 3.5. *Under a quadratic loss function, the Bayes rule for point estimation is the conditional expectation*

$$\delta^*(y) = \mathbb{E}(\Theta \mid Y = y).$$

A Bayes rule for a point estimation is known as a *Bayes estimator*. Note that although the matrix Q is involved in defining the quadratic loss function in (3.4), it does not influence the Bayes estimator. Thus the Bayes estimator is the same for an uncountably large class of loss functions. Depending on your point of view, this is either its most attractive or its most disturbing feature.

Proof of Theorem 3.5. Here is a proof that does not involve differentiation. The BRT (Theorem 3.2) asserts that

$$\delta^*(y) = \operatorname*{argmin}_{a \in \Omega} \mathbb{E}\{L(a, \Theta) \mid Y = y\}. \tag{3.5}$$

So let $\psi(y) := \mathbb{E}(\Theta \mid Y = y)$. For simplicity, treat θ as a scalar. Then

$$\begin{aligned} L(a, \theta) &\propto (a - \theta)^2 \\ &= (a - \psi(y) + \psi(y) - \theta)^2 \\ &= (a - \psi(y))^2 + 2(a - \psi(y))(\psi(y) - \theta) + (\psi(y) - \theta)^2. \end{aligned}$$

Take expectations conditional on $Y = y$ to get

$$\mathbb{E}\{L(a, \Theta) \mid Y = y\} \propto (a - \psi(y))^2 + \mathbb{E}\{(\psi(y) - \theta)^2 \mid Y = y\}, \qquad (\dagger)$$

where the cross-product term is zero. Only the first term contains a, and this term is minimized over a by setting $a = \psi(y)$, as was to be shown.

The extension to vector θ with loss function (3.4) is straightforward, but involves more ink. It is crucial that Q in (3.4) is positive definite, because otherwise the first term in (†), which becomes $(a - \psi(y))^T Q\,(a - \psi(y))$, is not minimized if and only if $a = \psi(y)$. □

Now apply the CCT (Theorem 3.4) to this result. For quadratic loss, a point estimator for θ is admissible if and only if it is the conditional expectation with respect to some positive prior distribution π.[7] Among the casualties of this conclusion is the Maximum Likelihood Estimator (MLE),

$$\hat{\theta}(y) := \operatorname*{arg\,max}_{\theta \in \Omega} f(y; \theta).$$

Stein's paradox showed that under quadratic loss, the MLE is not always admissible in the case of a Multinormal distribution with

[7] This is under the conditions of Theorem 3.4, or with appropriate extensions of them in the non-finite cases.

known variance, by producing an estimator which dominated it. This result caused such consternation when first published that it might be termed 'Stein's bombshell'. See Efron and Morris (1977) for more details, Samworth (2012) for an accessible proof, and Efron and Hastie (2016) for the consequences. Persi Diaconis thought this was such a powerful result that he focused on it for his brief article on Mathematical Statistics in the *The Princeton Companion to Mathematics* (Ed. T. Gowers, 2008, 1056 pages). Nevertheless, the MLE is still the dominant point estimator in large areas of applied statistics, even though its admissibility under quadratic loss is questionable.

3.6 Set estimation

For set estimation the action space is $\mathcal{A} = 2^\Omega$, and the loss function $L(a, \theta)$ represents the (negative) consequences of choosing $a \subset \Omega$ as a set estimate of Θ, when the true value of Θ is θ.

There are two contradictory requirements for set estimators of Θ. We want the sets to be small, but we also want them to contain Θ. There is a simple way to represent these two requirements as a loss function, which is to use

$$L(a, \theta) = |a| + \kappa \cdot (1 - \mathbb{1}_{\theta \in a}) \quad \text{for some } \kappa > 0 \tag{3.6a}$$

where $|a|$ is the volume of a.[8] The value of κ controls the trade-off between the two requirements. If $\kappa \downarrow 0$ then the Bayes rule is the empty set, for all y. If $\kappa \uparrow \infty$ then the Bayes rule is Ω, for all y. For κ in-between, the Bayes rule will depend on beliefs about Y and the value y. Theorem 3.6 below continues to hold for the much more general set of loss functions

$$L(a, \theta) = g(|a|) + h(1 - \mathbb{1}_{\theta \in a}) \tag{3.6b}$$

where g is non-decreasing and h is strictly increasing. This is a large set of loss functions, which should satisfy most clients who do not have a specific loss function already in mind.

For point estimators there was a simple characterisation of the Bayes rule for quadratic loss functions (Theorem 3.5). For set estimators the situation is not so simple. However, for loss functions of the form (3.6) there is a simple necessary condition for a rule to be a Bayes rule. A set $a \subset \Omega$ is a *level set* of the posterior distribution exactly when $a = \{\theta : \mathrm{p}(\theta \mid y) \geq k\}$ for some k.

Theorem 3.6. *If $\delta^* : \mathcal{Y} \to 2^\Omega$ is a Bayes rule for the loss function in* (3.6a)*, then it is a level set of the posterior distribution.*

Proof. For fixed y, I show that if a is not a level set of the posterior distribution, then there is an $a' \neq a$ which has a smaller expected loss; hence $\delta^*(y) \neq a$ according to the Bayes Rule theorem (BRT, Theorem 3.2).

First, note that

$$\mathbb{E}\{L(a, \Theta) \mid Y = y\} = |a| + \kappa \cdot \mathbb{P}(\Theta \notin a \mid Y = y). \tag{†}$$

[8] Technically, Lebesgue measure, if Ω is a convex subset of $\mathbb{R}^d$.

Now suppose that a is not a level set of $\mathrm{p}(\theta \mid y)$. In that case there is a $\theta \in a$ and a $\theta' \notin a$ for which $\mathrm{p}(\theta' \mid y) > \mathrm{p}(\theta \mid y)$. Let $a' = a \cup \mathrm{d}\theta' \setminus \mathrm{d}\theta$.[9] Then $|a'| = |a|$, but

$$\mathbb{P}(\Theta \notin a' \mid Y = y) < \mathbb{P}(\Theta \notin a \mid Y = y).$$

[9] Here, $\mathrm{d}\theta$ is the tiny region of Ω around θ, and $\mathrm{d}\theta'$ is the tiny region of Ω around θ', for which $|\mathrm{d}\theta| = |\mathrm{d}\theta'|$.

Thus

$$\mathbb{E}\{L(a', \Theta) \mid Y = y\} < \mathbb{E}\{L(a, \Theta) \mid Y = y\}$$

from (†), showing that $\delta^*(y) \neq a$. □

Now relate this result to the CCT (Theorem 3.4). First, Theorem 3.6 asserts that δ being a level set of the posterior distribution is necessary (but not sufficient) for δ to be a Bayes rule for loss functions of the form (3.6). Second, the CCT asserts that being a Bayes rule is necessary (but not sufficient) for δ to be admissible.[10] So being a level set of a posterior distribution for some prior distribution π (which is *not* allowed to depend on y) is a necessary condition for being admissible under (3.6).

[10] Necessary but not sufficient because being a Bayes rule AND having a positive prior distribution is equivalent to being admissible by the CCT, so being a Bayes rule without a condition on the prior distribution is necessary but not sufficient. As before, terms and conditions apply in the non-finite cases.

Now no one actually has (3.6) as their loss function; κ is a very inaccessible quantity. Eq. (3.6) is a generic loss function designed to help understand the features of a useful set estimator. Bayesian set estimators are usually *level* 95% *high posterior density (HPD) regions*. This is the level set of the posterior distribution which contains 95% of the posterior probability; other levels are also used.[11] So HPD regions satisfy the necessary condition for being a set estimator for the generic loss function (3.6).

[11] HPD regions have the useful property of being nested for different levels.

Frequentist set estimators achieve a similar outcome if they are level sets of the likelihood function $f(y; \bullet)$, because the posterior distribution is proportional to the likelihood function under a uniform prior distribution.[12] Frequentists do not need to actually adopt a unform prior distribution: they only need to point out that the uniform prior distribution ensures the admissibility of their 'level-sets of the likelihood function' estimator for the generic loss function (3.6), via the CCT.

[12] Or an almost-uniform prior distribution, in the case where Ω is unbounded, because the prior distribution will have to taper or be truncated in order to integrate to 1 over Ω.

3.7 *Hypothesis tests*

For hypothesis tests, the action space is a partition of Ω, denoted

$$\mathcal{H} := \{H_0, H_1, \ldots, H_d\}.$$

Each element of $\mathcal{H}$ is termed a *hypothesis*; it is traditional to number the hypotheses from zero, where H_0 is termed the *null hypothesis*. The loss function $L(H_i, \theta)$ represents the (negative) consequences of choosing element H_i, when the true value of Θ is θ. It would be usual for the loss function to satisfy

$$\theta \in H_i \implies L(H_i, \theta) = \min_{i'} L(H_{i'}, \theta)$$

on the grounds that an incorrect choice of element should never incur a smaller loss than the correct choice.

There is one case where we have a complete theory of Bayes/admissible rules. Let $\Omega = \{\theta_0, \theta_1\}$, with $H_i = \{\theta_i\}$, for which the loss function will have the form

L	θ_0	θ_1
H_0	0	ℓ_1
H_1	ℓ_0	0

with $\ell_0, \ell_1 > 0$. Then it is straightforward to show that the Bayes rule for choosing between H_0 and H_1 has the form

$$\frac{f(y;\theta_0)}{f(y;\theta_1)} \begin{cases} < c & \text{choose } H_1 \\ = c & \text{toss a coin} \\ > c & \text{choose } H_0 \end{cases} \tag{3.7}$$

where $c = (\pi_1/\pi_0) \cdot (\ell_1/\ell_0)$. Thus the CCT states that a decision rule is admissible if and only if it has the form in (3.7) for some $c > 0$.

In situations more complicated than this, it is extremely challenging and time-consuming to specify a loss function. And yet statisticians would still like to choose between hypotheses, in decision problems whose outcome does not seem to justify the effort required to specify the loss function.[13]

[13] Just to be clear, *important* decisions should not be based on cut-price procedures: an important decision warrants the effort required to specify a loss function.

There is a generic loss function for hypothesis tests, but it is hardly defensible. The *o-1 ('zero-one') loss function* is

$$L(H_i, \theta) = 1 - \mathbb{1}_{\theta \in H_i},$$

i.e., zero if θ is in H_i, and one if it is not (see Fig. 3.2). Its Bayes rule is to select the hypothesis with the largest posterior probability. It is hard to think of a reason why the 0-1 loss function would approximate a wide range of actual loss functions, unlike in the cases of generic loss functions for point estimation and set estimation. This is not to say that it is wrong to select the hypothesis with the largest posterior probability; only that the 0-1 loss function does not provide a very compelling reason.

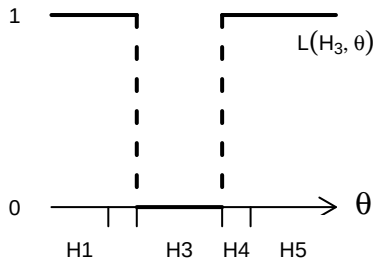


Figure 3.2: Zero-one loss function.

* * *

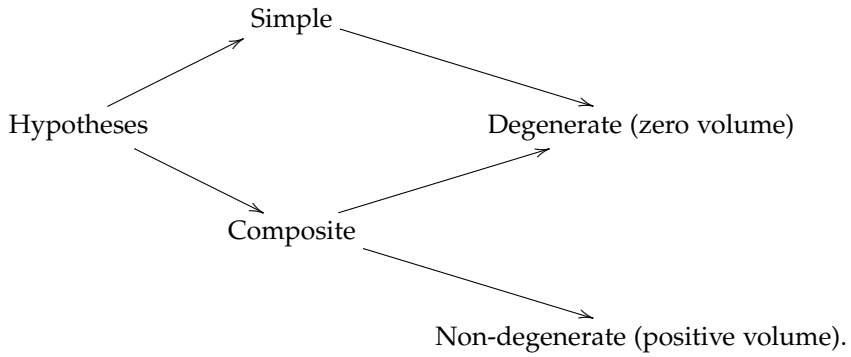
There is another approach which has proved much more popular. In fact, it is the dominant approach to hypothesis testing. This is to co-opt the theory of set estimators, for which there *is* a defensible generic loss function (see section 3.6). The statistician can use her set estimator $\delta : \mathcal{Y} \to 2^\Omega$ to make at least some distinctions between the members of $\mathcal{H}$, on the basis of the value of the observable, y^{obs}:

- 'Accept' H_i exactly when $\delta(y^{\text{obs}}) \subset H_i$,
- 'Reject' H_i exactly when $\delta(y^{\text{obs}}) \cap H_i = \emptyset$,
- 'Undecided' about H_i otherwise.

Note that these three terms are given in scare quotes, to indicate that they acquire a technical meaning in this context. We do not use

the scare quotes in practice, but we always bear in mind that we are not "accepting H_i" in the vernacular sense, but simply asserting that $\delta(y^{\text{obs}}) \subset H_i$ for our particular choice of δ.

In order to see how this approach plays out, we need to distinguish three types of hypothesis. The traditional distinction is between *simple hypotheses*, where $H_i = \{\theta_i\}$, a single element of Ω, and *composite hypotheses*, where H_i comprises more than a single element of Ω. Within composite hypotheses, though, we have *degenerate hypotheses*, which have zero volume in Ω, and *non-degenerate hypotheses*, which have positive volume; simple hypotheses always have zero volume. So here is the picture:



Obviously, it is effectively impossible to put a set inside a degenerate hypothesis, and so it is effectively impossible to accept a degenerate hypothesis using a set estimator—it is only possible reject it, or to be undecided.

To illustrate, suppose that the model is

$$\mathcal{E} = \left\{\mathbb{R}, (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++}, f\right\}$$

where f is the Normal probability density function (see Fig. 3.3). $H_1 : \{\mu = 0, \sigma^2 = 1\}$ would be a simple hypothesis; $H_2 : \{\mu = 0\}$ would be a composite degenerate hypothesis, and $H_3 : \{\mu > 0\}$ would be a composite non-degenerate hypothesis. It is possible to reject or be undecided about all three hypotheses, but it is only possible to accept H_3. Some statistics teachers seem to be confused about this, asserting that "it is never possible to accept the null hypothesis", or similar. This is not true in general, but it is true in the special case where the null hypothesis is degenerate (as is often the case in practice).

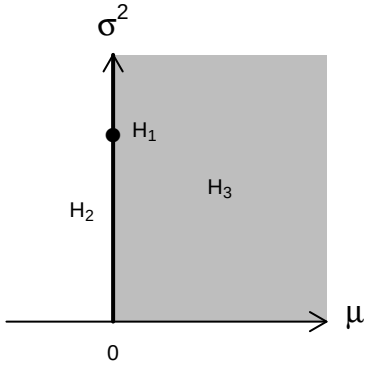


Figure 3.3: Different types of hypothesis. $H_1 : \mu = 0, \sigma^2 = 1$ is simple; $H_2 : \mu = 0$ is composite but degenerate; $H_3 : \mu > 0$ is composite and non-degenerate.

This set-estimator approach to hypothesis testing seems quite clear-cut, but we must end on two cautions. First, the statistician has not solved the decision problem of choosing an element of $\mathcal{H}$. She has solved a different problem. Based on a set estimator, she may reject H_0 on the basis of y^{obs}, but that does not mean she should proceed as though H_0 is false. This would require her to solve the correct decision problem, for which she would have to supply a loss function.

Second, in many situations, a hypothesis test is only superficially the right approach: attractive because of its simplicity, but limited

for the same reason. For example, suppose that $H_0 : \{\mu \leq 0\}$ and $H_1 : \{\mu > 0\}$, where a positive value of μ indicates that a new type of drug does more good than harm. One could accept H_1 and yet the set estimate could be pressed close up against the line $\mu = 0$ without touching it, or one could be undecided about H_1 and yet most of the set estimate could be much larger than $\mu = 0$, with only a small tail crossing over. It is excessively crude to reduce a set estimate to a discrete choice between elements of $\mathcal{H}$, and for this reason many statisticians have never done a hypothesis test.[14] This is not a new revelation. Over fifty years ago, Edwards et al. (1963, p. 213) wrote

> No aspect of classical statistics has been so popular with psychologists and other scientists as hypothesis testing, though some classical statisticians agree with us that the topic has been overemphasized. A statistician of great experience told us, "I don't know much about tests, because I have never had occasion to use one."

Plus ça change, as they say.

[14] Including me, since I became a proper statistician.

4
Confidence sets and p-values

This chapter is a continuation of Chapter 3, and the same conditions hold; re-read the introduction to Chapter 3 if necessary. As usual, the model is $\{\mathcal{Y}, \Omega, f\}$.

4.1 Confidence procedures and confidence sets

Definition 4.1 (Confidence procedure)**.** $C : \mathcal{Y} \to 2^\Omega$ is a level-$(1-\alpha)$ confidence procedure exactly when

$$\mathbb{P}\{\theta \in C(Y); \theta\} \geq 1-\alpha \quad \text{for all } \theta \in \Omega.$$

If the probability equals $(1-\alpha)$ for all θ, then C is an *exact* level-$(1-\alpha)$ confidence procedure.[1]

[1] Exact is a special case. But when it necessary to emphasize that C is not exact, the term 'conservative' is used.

The value $\mathbb{P}\{\theta \in C(Y); \theta\}$ is termed the *coverage* of C at θ. Thus a 95% confidence procedure has coverage of at least 95% for all θ, and an exact 95% confidence procedure has coverage of exactly 95% for all θ.

It is helpful to distinguish between the confidence procedure C, which is a function of y, and the result when C is evaluated at the observations y^{obs}, which is a set in Ω. I like the terms used in Morey et al. (2016), which I will also adapt to p-values in Definition 4.5.

Definition 4.2 (Confidence set)**.** $C(y^{\text{obs}})$ is a level-$(1-\alpha)$ confidence set exactly when C is a level-$(1-\alpha)$ confidence procedure.

So a confidence procedure is a function, and a confidence set is a set. If $\Omega \subset \mathbb{R}$ and $C(y^{\text{obs}})$ is convex, i.e. an interval, then a confidence set (interval) is represented by a lower and upper value. We should write, for example, "using procedure C, the 95% confidence interval for θ is $[0.78, 0.85]$", inserting "exact" if the confidence procedure C is exact.

* * *

The challenge with confidence procedures is to construct one with a specified level (look back to Section 1.4). You could propose an arbitrary $C : \mathcal{Y} \to 2^\Omega$, and then laboriously compute the coverage for every $\theta \in \Omega$. At that point you would know the level of C as a

confidence procedure, but it is unlikely to be 95%; adjusting C and iterating this procedure many times until the minimum coverage was equal to 95% would be exceedingly tedious. So we need to go backwards: start with the level, e.g. 95%, then construct a C designed to have this level.

Definition 4.3 (Family of confidence procedures)**.** $C : \mathcal{Y} \times [0,1] \to 2^{\Omega}$ is a family of confidence procedures exactly when $C(\bullet; \alpha)$ is a level-$(1-\alpha)$ confidence procedure for every $\alpha \in [0,1]$. C is a nesting family exactly when $\alpha < \alpha'$ implies that $C(y;\alpha) \supset C(y;\alpha')$.

4.2 *Significance procedures, and duality*

Before defining a significance procedure, we need some additional concepts. Let X and Y be two scalar random quantities. Then X stochastically dominates Y exactly when

$$\mathbb{P}(X \leq v) \leq \mathbb{P}(Y \leq v) \quad \text{for all } v \in \mathbb{R}.$$

Visually, the distribution function for X is never to the left of the distribution function for Y.[2] Although it is not in general use, I define the following term.

[2] Recollect that the distribution function of X has the form $F_X(x) := \mathbb{P}(X \leq x)$ for $x \in \mathbb{R}$.

Definition 4.4 (Super-uniform)**.** The random quantity X is *super-uniform* exactly when it stochastically dominates a standard uniform random quantity, i.e.

$$\mathbb{P}(X \leq u) \leq u \quad \text{for all } u \in [0,1].$$

Now we can define a significance procedure: note the similarities with the definition of a confidence procedure, which are not coincidental.

Definition 4.5 (Significance procedure)**.**

1. $p : \mathcal{Y} \to \mathbb{R}$ is a *significance procedure* for $\theta_0 \in \Omega$ exactly when $p(Y)$ is super-uniform under θ_0. If $p(Y)$ is uniform under θ_0, then p is an *exact* significance procedure for θ_0.

2. $p(y^{\text{obs}})$ is a significance level or *p-value* for θ_0 exactly when p is a significance procedure for θ_0.

3. $p : \mathcal{Y} \times \Omega \to \mathbb{R}$ is a family of significance procedures exactly when $p(\bullet; \theta_0)$ is a significance procedure for θ_0, for every $\theta_0 \in \Omega$.

With this definition, the punchline is easy to anticipate.

Theorem 4.6 (Duality theorem)**.**

1. *Let p be a family of significance procedures. Then*

$$C(y;\alpha) := \{\theta \in \Omega : p(y;\theta) > \alpha\}$$

   is a nesting family of confidence procedures.

2. *Conversely, let C be a nesting family of confidence procedures. Then*

$$p(y;\theta_0) := \inf\{\alpha : \theta_0 \notin C(y;\alpha)\}$$

   is a family of significance procedures.

If either is exact, then the other is exact as well.

Proof.

Proof of (1). It is clear that C is nesting. So we need to show that $\mathbb{P}\{\theta \in C(Y;\alpha);\theta\} \geq 1-\alpha$ for all $\theta \in \Omega$. So let θ be arbitrary, and then:

$$\begin{aligned}
\mathbb{P}\{\theta \in C(Y;\alpha);\theta\} &= \mathbb{P}\{p(Y;\theta) > \alpha;\theta\} \\
&= 1 - \mathbb{P}\{p(Y;\theta) \leq \alpha;\theta\} \\
&= 1 - (\leq \alpha) && p \text{ super-uniform for all } \theta \\
&\geq 1-\alpha
\end{aligned}$$

as needed to be shown. If p is exact, then the inequality is replaced by an equality, and hence C is exact as well.

Proof of (2). We need to show that $p(Y;\theta_0)$ is super-uniform for all $\theta_0 \in \Omega$. The crucial insight is that

$$\inf\{\alpha : \theta_0 \notin C(y;\alpha)\} \leq u \iff \theta_0 \notin C(y;u).$$

This follows by the nesting property. Here I am finessing the issue of the boundary of C by assuming that if $\alpha^* := \inf\{\alpha : \theta_0 \notin C(y;\alpha)\}$, then $\theta_0 \notin C(y;\alpha^*)$. So let θ_0 and $u \in [0,1]$ be arbitrary, and then

$$\mathbb{P}\{p(Y;\theta_0) \leq u;\theta_0\} = \mathbb{P}\{\theta_0 \notin C(Y;u);\theta_0\} \leq u$$

because $C(\bullet;u)$ is a level-$(1-u)$ confidence procedure. Hence p is super-uniform, as needed to be shown. If C is exact, then the inequality is replaced by an equality, and hence p is exact as well. □

Theorem 4.6 shows that confidence procedures and significance procedures are two sides of the same coin. If we have a way of constructing families of confidence procedures then we have a way of contructing families significance procedures, and *vice versa*. If we have a *good* way of constructing confidence procedures then (presumably, and in principle) we have a good way of constructing significance procedures. This is helpful because, as Section 4.3 will show, there are an uncountable number of families of significance

procedures, and so there are an uncountable number of families of confidence procedures. Naturally, in both these cases, almost all of the possible procedures are useless for our inference. So just being a confidence procedure, or just being a signifiance procedure, is never enough. We need to know how to make good choices.

4.3 *Families of significance procedures*

Here is a very general way to construct a family of significance procedures. Below I will show that this family is also easy to compute, by simulation (after Theorem 4.8).

Theorem 4.7. *This proof comes from Casella and Berger (2002, sec. 8.3.4). Let $t : \mathcal{Y} \to \mathbb{R}$. Define*

$$p_t(y;\theta_0) := \mathbb{P}\left\{t(Y) \geq t(y);\theta_0\right\}.$$

Then p_t is a family of significance procedures. If the distribution function of $t(Y)$ is continuous, then p_t is exact.

Proof. Directly from the definition,

$$\begin{aligned} p_t(y;\theta_0) &= \mathbb{P}\{t(Y) \geq t(y);\theta_0\} \\ &= \mathbb{P}\{-t(Y) \leq -t(y);\theta_0\} =: G(-t(y)) \end{aligned}$$

where G is the distribution function of $-t(Y)$ under θ_0. Then

$$p_t(Y;\theta_0) = G(-t(Y))$$

which is super-uniform under θ_0 according to the Probability Integral Transform (see Section 4.7, notably Theorem 4.14). The PIT also covers the case where the distribution function of $t(Y)$ is continuous, in which case $p_t(Y;\theta_0)$ is uniform under θ_0. □

So there is a family of significance procedures for each possible function $t : \mathcal{Y} \to \mathbb{R}$. Clearly only a tiny fraction of these can be useful functions, and the rest must be useless. Some, like $t(y) = c$ for some constant c, are always useless. Others, like $t(y) = \sin(y_1)$ might sometimes be a little bit useful, while others, like $t(y) = \sum_i y_i$ might be quite useful—but it all depends on the circumstances.

Some additional criteria are required to separate out good from poor choices of the test statistic t, when using the contruction in Theorem 4.7. The most pertinent criterion is:

- Select a test statistic for which $t(Y)$ which will tend to be larger for decision-relevant departures from θ_0.

This will ensure that $p_t(Y;\theta_0)$ will tend to be smaller under decision-relevant departures from θ_0; small p-values are more interesting, precisely because significance procedures are super-uniform under θ_0.

4.3.1 Computing p-values

Only in very special cases will it be possible to find a closed-form expression for p_t from which we can compute the p-value $p_t(y^{\text{obs}};\theta_0)$. Instead, we can use simulation, according to the following result (adapted from Besag and Clifford, 1989).

Theorem 4.8. *For any finite sequence of scalar random quantities $X^0, X^1, \ldots, X^m$, define the rank of X^0 in the sequence as*

$$R := \sum_{i=1}^{m} \mathbb{1}_{X^i \leq X^0}.$$

If $X^0, X^1, \ldots, X^m$ are exchangeable then R has a uniform distribution on the integers $0, 1, \ldots, m$, and $(R+1)/(m+1)$ has a super-uniform distribution (see Definition 4.4).

Proof. By exchangeability, X^0 has the same probability of having rank r as any of the other X's, for any r, and therefore

$$\mathbb{P}(R = r) = \frac{1}{m+1} \quad \text{for } r = 0, 1, \ldots, m \qquad (\dagger)$$

and zero otherwise, proving the first claim.

To prove the second claim,[3]

$$\begin{aligned}
\mathbb{P}\left\{\frac{R+1}{m+1} \leq u\right\} &= \mathbb{P}\left\{R+1 \leq u(m+1)\right\} \\
&= \mathbb{P}\left\{R+1 \leq \lfloor u(m+1) \rfloor\right\} \quad \text{as } R \text{ is an integer} \\
&= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \mathbb{P}(R = r) \\
&= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \frac{1}{m+1} \qquad \text{from } (\dagger) \\
&= \frac{\lfloor u(m+1) \rfloor}{m+1} \leq u,
\end{aligned}$$

as required. □

[3] Notation: $\lfloor x \rfloor$ is the largest integer no larger than x, termed the 'floor' of x.

To use this result, fix the test statistic t and define $T^i := t(Y^i)$ where $Y^1, \ldots, Y^m \overset{\text{iid}}{\sim} f(\bullet;\theta_0)$. Define

$$R_t(y;\theta_0) := \sum_{i=1}^{m} \mathbb{1}_{-T^i \leq -t(y)} = \sum_{i=1}^{m} \mathbb{1}_{T^i \geq t(y)},$$

where θ_0 is an argument to R because θ_0 needs to be specified in order to simulate $T^1, \ldots, T^m$. Then Theorem 4.8 implies that

$$P_t(y;\theta_0) := \frac{R_t(y;\theta_0) + 1}{m+1}$$

has a super-uniform distribution under $Y \sim f(\bullet;\theta_0)$, because in this case $t(Y), T^1, \ldots, T^m$ are exchangeable. Furthermore, the Weak Law

of Large Numbers (WLLN) implies that

$$\begin{aligned}\lim_{m\to\infty} P_t(y;\theta_0) &= \lim \frac{R_t(y;\theta_0)+1}{m+1} \\ &= \lim \frac{m^{-1}\{R_t(y;\theta_0)+1\}}{m^{-1}\{m+1\}} \\ &= \lim m^{-1} R_t(y;\theta_0) \\ &= \mathbb{P}\{T \geq t(y);\theta_0\} = p_t(y;\theta_0).\end{aligned}$$

Therefore, not only is $P_t(Y;\theta_0)$ super-uniform under θ_0, so that P_t is a family of significance procedures for every m, but the limiting value of $P_t(y;\theta_0)$ as m becomes large is $p_t(y;\theta_0)$.

In summary, if you can simulate from your model under θ_0 then you can produce a p-value for any test statistic t, namely $P_t(y^{\text{obs}};\theta_0)$, and if you can simulate cheaply, so that the number of simulations m is large, then $P_t(y^{\text{obs}};\theta_0) \approx p_t(y^{\text{obs}};\theta_0)$.

The less-encouraging news is that this simulation-based approach is not well-adapted to constructing confidence sets. Let C_t be the family of confidence procedures induced by p_t using Duality (Theorem 4.6). We can answer the question 'Is $\theta_0 \in C_t(y^{\text{obs}};\alpha)$?' with one set of m simulations. These simulations give a value $P_t(y^{\text{obs}};\theta_0)$ which is either larger or not-larger than α. If $P_t(y^{\text{obs}};\theta_0) > \alpha$, then $\theta_0 \in C_t(y^{\text{obs}};\alpha)$, and otherwise it is not. Clearly, though, this is not an effective way to enumerate all of the points in $C_t(y^{\text{obs}};\alpha)$, because we would need to do m simulations for each point in Ω.

4.4 *Good choices of confidence procedures*

Here is a property that a confidence procedure may or may not have.

Definition 4.9 (Level set property, LSP)**.** A confidence procedure C has the level set property exactly when

$$C(y) = \{\theta : f(y;\theta) > g(y)\}$$

for some $g : \mathcal{Y} \to \mathbb{R}$.

Recollect from Section 3.6 that the LSP is akin to a necessary condition for C to be an admissible set estimator under the loss function in (3.6), by Theorem 3.6. So under these terms, which seem reasonable, confidence procedures without the LSP would be bad choices.

As usual, we must ask whether the LSP is vacuous: can we construct a family of confidence procedures with the LSP? Indeed we can. Here is a result that has pedagogic value,[4] because it can be used to generate an uncountable number of families of confidence procedures, each with the LSP.

[4] This means that you may not want to use these confidence procedures in practice!

Theorem 4.10. *Let h be any PMF for Y. Then*

$$C(y;\alpha) := \{\theta \in \Omega : f(y;\theta) > \alpha \cdot h(y)\} \tag{4.1}$$

is a family of confidence procedures, with the LSP.

Proof. Define $g(y, \theta) := f(y;\theta)/h(y)$, which may equal ∞ if $h(y) = 0$. Then the result follows immediately from Theorem 4.6 part (1), because $g(Y, \theta)$ is super-uniform for each θ:

$$\begin{aligned}\mathbb{P}\{f(Y;\theta)/h(Y) \leq u;\theta\} &= \mathbb{P}\{h(Y)/f(Y;\theta) \geq 1/u;\theta\} \\ &\leq \frac{\mathbb{E}\{h(Y)/f(Y;\theta);\theta\}}{1/u} \qquad \text{Markov's inequality} \\ &\leq \frac{1}{1/u} = u.\end{aligned}$$

For the final inequality, let $\mathcal{Y}(\theta) := \{y \in \mathcal{Y} : f(y;\theta) > 0\}$. Then

$$\begin{aligned}\mathbb{E}\{h(Y)/f(Y;\theta);\theta\} &= \sum_{y \in \mathcal{Y}(\theta)} \frac{h(y)}{f(y;\theta)} \cdot f(y;\theta) \\ &= \sum_{y \in \mathcal{Y}(\theta)} h(y) \leq 1,\end{aligned}$$

because h is a probability mass function. □

Among the interesting choices for h, one possibility is $h = f(\bullet;\theta_0)$, for some $\theta_0 \in \Omega$. Note that with this choice, the confidence set of (4.1) always contains θ_0. So we know that we can construct a level-$(1-\alpha)$ LSP confidence procedure whose confidence sets will always contain θ_0, for any $\theta_0 \in \Omega$. Two statisticians can both construct 95% confidence sets for θ which satisfy the LSP, using different families of confidence procedures. Yet using the approach in Section 3.7, the first statistician may reject the null hypothesis that $H_0 : \Theta = \theta_0$, and the second statistician may fail to reject it, for any $\theta_0 \in \Omega$. This does not fill one with confidence about using confidence procedures for hypothesis tests.

Actually, the situation is not as grim as it seems. Markov's inequality is very slack, and so the coverage of the family of confidence procedures defined in Theorem 4.10 is likely to be much larger than $(1-\alpha)$, e.g. much larger than 95%.

For any confidence procedure, the diameter of $C(y)$ can grow rapidly with its coverage.[5] In fact, the relation must be extrememly convex when coverage is nearly one, because, in the case where $\Omega = \mathbb{R}$, the diameter at coverage $= 1$ is unbounded. So an increase in the coverage from, say 95% to 99%, could easily correspond to a doubling or more of the diameter of the confidence procedure.

[5] The diameter of a set in a metric space such as Euclidean space is the maximum of the distance between two points in the set.

A more likely outcome in the two statisticians situation is that $C_h(y; 0.05)$ is large for many different choices of h, in which case no one rejects the null hypothesis, which is not a useful outcome for a hypothesis test. But perhaps it is a useful antidote to the current 'crisis of reproducibility', in which far too many null hypotheses are being rejected in published papers.

All in all, it would be much better to use an exact family of confidence procedures which satisfy the LSP, if one existed. And, for perhaps the most popular model in the whole of Statistics, this is the case. This is the Linear Model with a Normal error. I do not cover it here; see, e.g., Wood (2017, ch. 1). This model is a *very*

special case, and it is unfortunate that so many people who are learning statistics have their intuition shaped by it.

4.5 Generalizations

So far, confidence procedures and significance procedures have been defined with respect to a point $\theta_0 \in \Omega$. Often, though, we require a more general treatment, where a confidence procedure is defined for some $g : \theta \mapsto \phi$, where g may not be bijective; or where a significance procedure is defined for some $\Omega_0 \subset \Omega$, where Ω_0 may not be a single point. These general treatments are always possible, but the result is often very conservative. As discussed at the end of Section 4.4, conservative procedures are formally correct but they can be practically useless.

4.5.1 Marginalization of confidence procedures

Suppose that $g : \theta \mapsto \phi$ is some specified function, and we would like a confidence procedure for ϕ. If C is a level-$(1-\alpha)$ confidence procedure for ϕ then it must have ϕ-coverage of at least $(1-\alpha)$ for all $\theta \in \Omega$. The most common situation is where $\Omega \subset \mathbb{R}^p$, and g extracts a single component of θ: for example, $\theta = (\mu, \sigma^2)$ and $g(\theta) = \mu$.

Theorem 4.11. *Suppose that $g : \theta \mapsto \phi$, and that C is a level-$(1-\alpha)$ procedure for θ. Then gC is a level-$(1-\alpha)$ confidence procedure for ϕ.*[6]

[6] $gC := \big\{\phi : \phi = g(\theta) \text{ for some } \theta \in C\big\}$.

Proof. Follows immediately from the fact that $\theta \in C(y)$ implies that $\phi \in gC(y)$ for all y, and hence

$$\mathbb{P}\{\theta \in C(Y); \theta\} \leq \mathbb{P}\{\phi \in gC(Y); \theta\}$$

for all $\theta \in \Omega$. So if C has θ-coverage of at least $(1-\alpha)$, then gC has ϕ-coverage of at least $(1-\alpha)$ as well. □

This result shows that we can derive level-$(1-\alpha)$ confidence procedures for functions of θ directly from level-$(1-\alpha)$ confidence procedures for θ. Furthermore, if the confidence procedure for θ is easy to enumerate, then the confidence procedure for ϕ is easy to enumerate too—just by transforming each element. But it also shows that the coverage of such derived procedures will typically be more than $(1-\alpha)$, even if the original confidence procedure is exact: thus gC is a conservative confidence procedure. As already noted, conservative confidence procedures can often be far larger than they need to be: sometimes too large to be useful.

4.5.2 Generalization of significance procedures

There is a simple result which extends a family of significance procedures over a set in Ω.

Theorem 4.12. *Let $\Omega_0 \subset \Omega$. If p is a family of significance procedures, then*

$$P(y;\Omega_0) := \sup_{\theta_0 \in \Omega_0} p(y;\theta_0)$$

is super-uniform for all $\theta \in \Omega_0$.

Proof. $P(y;\Omega_0) \leq u$ implies that $p(y;\theta_0) \leq u$ for all $\theta_0 \in \Omega$. Let $\theta \in \Omega_0$ be arbitrary; then, for any $u \geq 0$,

$$\mathbb{P}\{P(Y;\Omega_0) \leq u;\theta\} \leq \mathbb{P}\{p(Y;\theta) \leq u;\theta\} \leq u, \quad \theta \in \Omega_0,$$

showing that $P(y;\Omega_0)$ is super-uniform for all $\theta \in \Omega_0$. □

As with the marginalization of confidence procedures, this result shows that we can derive a significance procedure for an arbitrary $\Omega_0 \subset \Omega$. The difference, though, is that this is rather impractical, because of the need—in general—to maximize over a possibly unbounded set Ω_0. As a result, this type of p-value is not much used in practice. It is sometimes replaced by simple approximations. For example, if the parameter is (ν, θ) then a p-value for ν_0 could be approximated by plugging-in a specific value for θ, such as the maximum liklehood value, and treating the model as though it were parameterized by ν alone. But this does not give rise to a well-defined significance procedure for ν_0 on the basis of the original model. Adopting this type of approach is something of an act of desperation, for when Theorem 4.12 is intractable. The difficulty is that you get a number, but you do not know what it signifies.

4.6 Reflections

4.6.1 On the definitions

The first thing to note is the abundance of families of confidence procedures and significance procedures, most of which are useless. For example, let U be a uniform random quantity. Based on the definition alone,

$$C(y;\alpha) = \begin{cases} \{0\} & U < \alpha \\ \Omega & U \geq \alpha, \end{cases}$$

is a perfectly acceptable family of exact confidence procedures, and

$$p(y;\theta_0) = U$$

is a perfectly acceptable family of exact significance procedures. They are both useless. This absurdity has been captured in a cartoon, see Fig. 4.1. You cannot object that these examples are pathological because they contain the auxilliary random quantity U, because the most accessible method for computing p-values also contains auxilliary random quantities (see Section 4.3.1). You could object that the family of significance procedures does not have the LSP property (Definition 4.9), which is a valid objection if you intend to apply the LSP rigorously. But would you then have to insist

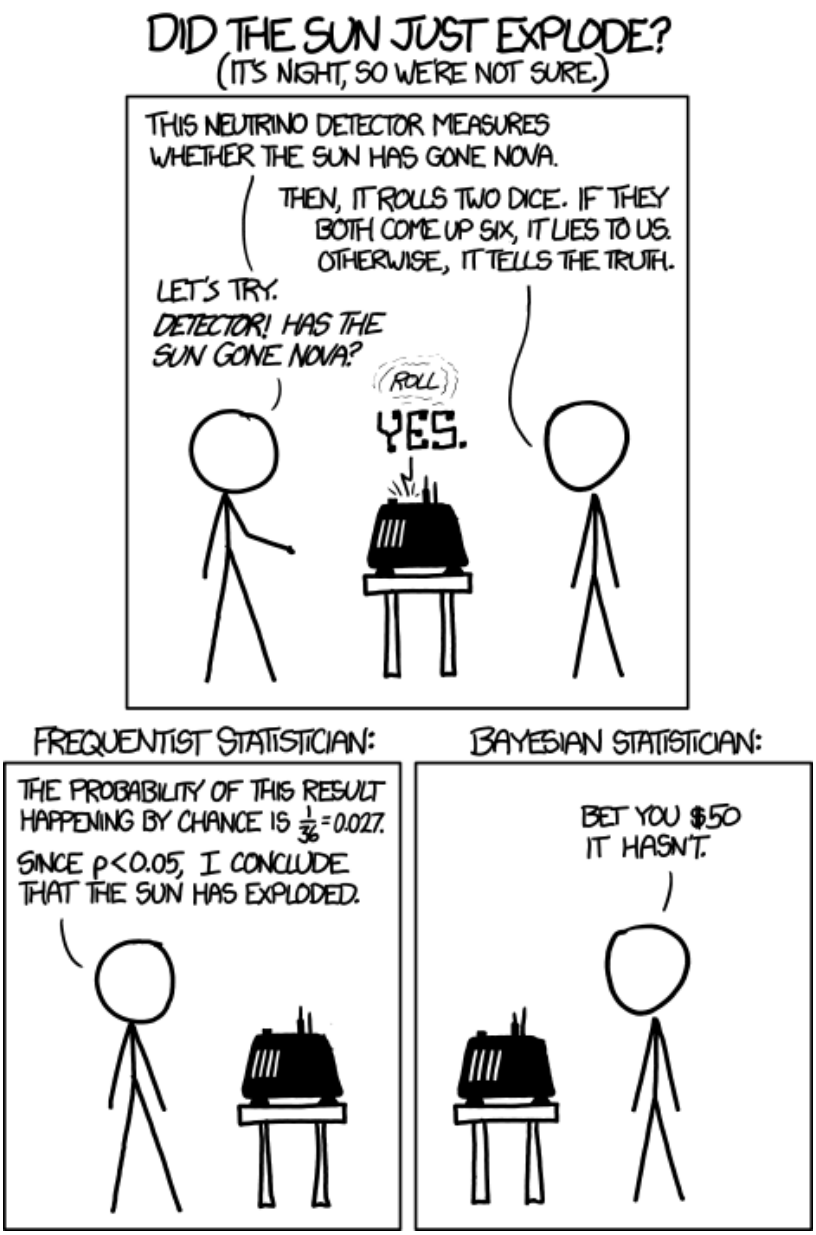


Figure 4.1: From https://xkcd.com/1132/.

that every significance procedure's dual confidence procedure (see Theorem 4.6) should also have the LSP?

The second thing to note is how often confidence procedures and significance procedures will be conservative. This means that there is some region of the parameter space where the actual coverage of the confidence procedure is more than the nominal coverage of $(1 - \alpha)$. Or where the significance procedure has a super-uniform but not uniform distribution under θ_0. As shown in this chapter:

- A generic method for constructing families of confidence procedures with the LSP (see Theorem 4.10) is always conservative.
- Confidence procedures for non-bijective functions of the parameters are always conservative (see Theorem 4.11).
- Significance procedures based on test statistics where $t(Y)$ is

discrete are always conservative (see Theorem 4.7).

- Significance procedures for composite hypotheses are always conservative (see Theorem 4.12).

A conservative procedure is hard to interpret. For example, you have set $\alpha = 0.05$ and the conservative confidence procedure is delivering "at least 95% coverage". You don't know how much more than 95% you are getting in the region where the true value of the parameter might be, but you do know that the set diameter is a strongly convex function of coverage, so your quoted set may be far larger than one with exactly 95% coverage, if it could be found.

4.6.2 On the intepretations

The technical definition of confidence procedures and significance procedures is daunting for the non-specialist. Here is a typical dialogue:

Statistician: The 95% confidence interval for your parameter is $(0.78, 0.85)$.

Client: So that means the probability that the true value of the parameter lies in the interval $(0.78, 0.85)$ is 95%.

Statistician: Err, no. $(0.78, 0.85)$ is one realization of a random interval with the property that it will contain the true value of the parameter at least 95% of the time, no matter what the true value happens to be.

Client: I'm not sure that's what I wanted.

Here is a similar dialogue for a p-value:

Statistician: The p-value for your hypothesis is 0.07.

Client: So that means the probability that the hypothesis is true is only 7%.

Statistician: Err, no. 7% is one realization of a random quantity with the property that its probability of being not more than 7% under your hypothesis is not more than 7%.

Client: I'm not sure that's what I wanted.

This second dialogue could have gone differently, if the statistician had focused on Theorem 4.7 as the functional definition of a p-value.

Statistician: The p-value for your hypothesis is 0.07.

Client: So that means the probability that the hypothesis is true is only 7%.

Statistician: Err, no. It means that the probability of a value of the test statistic [here she names the statistic] being at least as large as the observed value under your hypothesis is not more than 7%.

Client: I'm not sure that's what I wanted.

* * *

It is a very common observation, made repeatedly over the last 50 years, that clients think more like Bayesians than Frequentists, as represented by these dialogues (see, e.g., Rubin, 1984). Frequentist statisticians have to wrestle with the issue that their clients will likely misinterpret their results. This is bad enough for confidence sets (see, e.g., Morey et al., 2016), but potentially disastrous for p-values. A p-value $p(y;\theta_0)$ refers only to θ_0,[7] making no reference at all to other hypotheses about Θ. But a posterior probability $\mathrm{p}(\theta_0 \mid y^{\mathrm{obs}})$ contrasts θ_0 with other values in Ω which Θ might have taken. The two outcomes can be radically different, as first captured in Lindley's paradox (Lindley, 1957). To leave your client thinking that a small value for $p(y^{\mathrm{obs}};\theta_0)$ has rejected $\Theta = \theta_0$ on the basis of the data y^{obs} is irresponsible, and potentially dangerous.

[7] Or Ω_0 in the more general case.

4.7 *The Probability Integral Transform*

Here is a very elegant and useful piece of probability theory. Let X be a scalar random quantity with realm $\mathcal{X}$ and distribution function $F(x) := \mathbb{P}(X \le x)$. By convention, F is defined for all $x \in \mathbb{R}$. By construction, $\lim_{x\downarrow-\infty} F(x) = 0$, $\lim_{x\uparrow\infty} F(x) = 1$, F is non-decreasing, and F is continuous from the right, i.e.

$$\lim_{x'\downarrow x} F(x') = F(x).$$

Define the *quantile function*

$$F^-(u) := \inf\{x \in \mathbb{R} : F(x) \ge u\}. \tag{4.2}$$

The following result is a cornerstone of generating random quantities with easy-to-evaluate quantile functions.

Theorem 4.13 (Probability Integral Transform, PIT)**.** *Let U have a standard uniform distribution. If F^- is the quantile function of X, then $F^-(U)$ and X have the same distribution.*

Proof. Let F be the distribution function of X. We must show that

$$F^-(u) \le x \iff u \le F(x) \tag{†}$$

because then

$$\mathbb{P}\{F^-(U) \le x\} = \mathbb{P}\{U \le F(x)\} = F(x)$$

as required. So stare at Figure 4.2 for a while.

It is easy to check that

$$u \le F(x) \implies F^-(u) \le x,$$

which is one half of (†). It is also easy to check that

$$u' > F(x) \implies F^-(u') > x.$$

Taking the contrapositive of this second implication gives

$$F^-(u') \le x \implies u' \le F(x),$$

which is the other half of (†). □

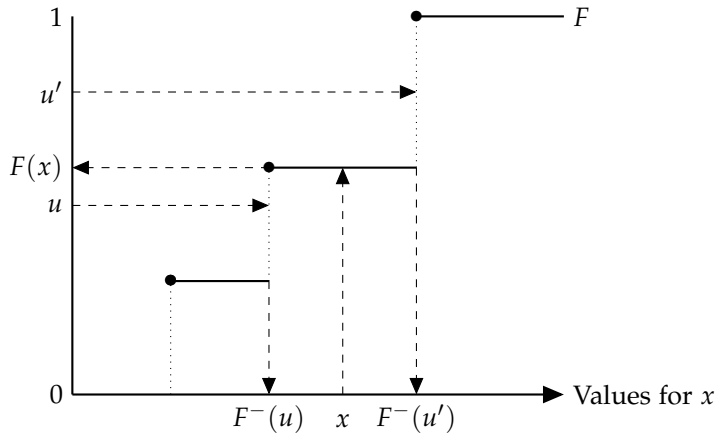


Figure 4.2: Figure for the proof of Theorem 4.13. The distribution function F is non-decreasing and continuous from the right. The quantile function F^- is defined in (4.2).

Theorem 4.13 is the basis for the following result; recollect the definition of a super-uniform random quantity from Definition 4.4. This result is used in Theorem 4.7.

Theorem 4.14. *If F is the distribution function of X, then $F(X)$ has a super-uniform distribution. If F is continuous then $F(X)$ has a uniform distribution.*

Proof. Check from Figure 4.2 that $F(F^-(u)) \geq u$. Then

$$\begin{aligned} \mathbb{P}\{F(X) \leq u\} &= \mathbb{P}\{F(F^-(U)) \leq u\} \qquad \text{from Theorem 4.13} \\ &\leq \mathbb{P}\{U \leq u\} \\ &= u. \end{aligned}$$

In the case where F is continuous, it is strictly increasing except on sets which have probability zero. Then

$$\mathbb{P}\{F(X) \leq u\} = \mathbb{P}\{F(F^-(U)) \leq u\} = \mathbb{P}\{U \leq u\} = u,$$

as required. □

5
Bibliography

Bartlett, M. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, 44:533–534. 56

Basu, D. (1975). Statistical information and likelihood. *Sankhyā*, 37(1):1–71. With discussion. 16, 17, 18, 21

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York, Inc., NY, USA, second edition. 30

Berger, J. and Wolpert, R. (1988). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward CA, USA, second edition. Available online, `http://projecteuclid.org/euclid.lnms/1215466210`. 16

Besag, J. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika*, 76(4):633–642. 45

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57:269–306. 15, 18

Birnbaum, A. (1972). More concepts of statistical evidence. *Journal of the American Statistical Association*, 67:858–861. 16, 18

Casella, G. and Berger, R. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury, 2nd edition. 3, 5, 44

Cormen, T., Leiserson, C., and Rivest, R. (1990). *Introduction to Algorithms*. The MIT Press, Cambridge, MA. 13

Cox, D. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge, UK. 3

Cox, D. and Donnelly, C. (2011). *Principles of Applied Statistics*. Cambridge University Press, Cambridge, UK. 3

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London, UK. 18, 19

Davison, A. (2003). *Statistical Models*. Cambridge University Press, Cambridge, UK. 5

Dawid, A. (1977). Conformity of inference patterns. In Barra, J. et al., editors, *Recent Developments in Statistcs*. North-Holland Publishing Company, Amsterdam. 16, 17

Edwards, W., Lindman, H., and Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242. 40

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, New York NY, USA. 3, 36

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236(5):119–127. Available at `http://statweb.stanford.edu/~ckirby/brad/other/Article1977.pdf`. 36

Ferguson, T. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, London, UK. 30, 32

Fisher, R. (1956). *Statistical Methods and Scientific Inference*. Edinburgh and London: Oliver and Boyd. 18

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton FL, USA, 3rd edition. Online resources at `http://www.stat.columbia.edu/~gelman/book/`. 11

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London, UK. 30

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge University Press, Cambridge, UK. 3

Hacking, I. (2014). *Why is there a Philosophy of Mathematics at all?* Cambridge University Press, Cambridge, UK. 4

Harford, T. (2014). Big data: Are we making a big mistake? *Financial Times Magazine*. Published online Mar 28, 2014. Available at `http://on.ft.com/P0PVBF`. 13

Lad, F. (1996). *Operational Subjective Statistical Methods*. John Wiley & Sons. Ltd, New York, USA. 4

Le Cam, L. (1990). Maximum likelihood: An introduction. *International Statistical Review*, 58(2):153–171. 7

Lindley, D. (1957). A statistical paradox. *Biometrika*, 44:187–192. See also Bartlett (1957). 52

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). *The BUGS Book: A Practical introduction to Bayesian Analysis*. CRC Press, Boca Raton FL, USA. 11

MacKay, D. (2009). *Sustainable Energy – Without the Hot Air*. UIT Cambridge Ltd, Cambridge, UK. available online, at `http://www.withouthotair.com/`. 4

Madigan, D., Strang, P., Berlin, J., Schuemie, M., Overhage, J., Suchard, M., Dumouchel, B., Hartzema, A., and Ryan, P. (2014). A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, 1:11–39. 10

Milner, K. and Rougier, J. (2014). How to weigh a donkey in the Kenyan countryside. *Significance*, 11(4):40–43. 34

Morey, R., Hoekstra, R., Rouder, J., Lee, M., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bullentin & Review*, 23(1):103–123. 41, 52

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. New York: Springer, 2nd edition. 6

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172. 52

Samworth, R. (2012). Stein's paradox. *Eureka*, 62:38–41. Available online at `http://www.statslab.cam.ac.uk/~rjs57/SteinParadox.pdf`. Careful readers will spot a typo in the maths. 36

Savage, L. et al. (1962). *The Foundations of Statistical Inference*. Methuen, London, UK. 22

Schervish, M. (1995). *Theory of Statistics*. Springer, New York NY, USA. Corrected 2nd printing, 1997. 3, 7, 11, 30, 32

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64(4):583–616. With discussion, pp. 616–639. 12

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B*, 76(3):485–493. 12

Stigler, S. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard University Press, Cambridge MA, USA. 3, 5, 24

Wood, S. (2017). *Generalized Linear Models: An Introduction with R*. CRC Press, Boca Raton FL, USA, 2nd edition. 47

Ziman, J. (2000). *Real Science: What it is, and what it means*. Cambridge University Press, Cambridge, UK. 24