APTS Statistical Inference, Lecture notes

Jonathan Rougier, University of Bristol

December 2012

The first half of this document is a crash-bang-wallop summary of the statistical material that I would present in a comprehensive second undergraduate course on statistical inference at a leading UK university. It covers both Frequentist and Bayesian approaches to estimation and hypothesis testing in some generality, with a linking section on different statistical interpretations of probability.

More advanced material is presented in the second half, covering asymptotic convergence, Decision Theory, Bayes Factors, inferential fallacies, and principles.

Throughout the documents the key results are proved to an acceptable level of rigour; longer proofs can be found in the textbooks listed at the end.

This is the first draft of these notes, and I expect there are mistakes, or passages that are unclear. So comments on this document are particularly welcome; please address them by email to me at `j.c.rougier@bristol.ac.uk`. This version of the document was created on December 1, 2012, and formatted using the `tufte-handout` class, available at `https://code.google.com/p/tufte-latex/`.

Contents

1 Probability and statistics background

At the heart of statistical inference is the notion of an *random quantity*, typically denoted X. I strongly advocate the positivist view that X is a set of operations which, if followed, will result in a value. A statement such as $\Pr(X = x)$ is to be read as "the probability that the operations described by X yields the value x", where I am adopting the standard convention that small letters denote specific possible values for X. It is a very useful discipline to insist that *random quantities have operational definitions*; it ensures that uncertainty about X is not confounded with ambiguity about the meaning of X.

random quantity

The set of possible outcomes for X is denoted $\mathcal{X}$, the *sample space* for X. The operational definition for X implies that $\mathcal{X}$ is finite, reflecting the finite precision of our instruments; that is to say, X *is always a discrete random quantity*. It may seem idiosyncratic to rule out continuous random quantities at the start of a course on statistical inference, but in fact many statisticians have advocated something similar. For example, Debabrata Basu:

sample space

> "The author holds firmly to the view that this contingent and cognitive universe of ours is in reality only finite and, therefore, discrete. In this essay we steer clear of the logical quick sands of 'infinity' and the 'infinitesimal'. Infinite and continuous models will be used in the sequel, but they are to be looked upon as mere approximations to the finite realities." (Basu, 1975, footnote, p. 4)

In fact, continuous random quantities will reappear for reasons explained in section 1.4.

A collection of random quantities is denoted $\boldsymbol{X} := (X_1, \dots, X_n) \in \boldsymbol{\mathcal{X}}$, where very often $\boldsymbol{\mathcal{X}} = \mathcal{X}^n$; likewise, $\boldsymbol{x} := (x_1, \dots, x_n)$.[1] In lectures and handwritten notes I will use an underscore, e.g. $\boldsymbol{X} \equiv \underline{X}$ and $\boldsymbol{x} \equiv \underline{x}$. Where it is necessary to specify the observed values of $\boldsymbol{X}$, these will be denoted $\boldsymbol{x}^{\text{obs}}$.

1.1 Probability basics

At this stage we assume the existence of a *probability mass function (PMF)* $f_{\boldsymbol{X}}$ with domain $\boldsymbol{\mathcal{X}}$, such that

probability mass function (PMF)

$$\Pr(\boldsymbol{X} \in A) := \sum_{\boldsymbol{x} \in A} f_{\boldsymbol{X}}(\boldsymbol{x}) \qquad \text{for any } A \subset \boldsymbol{\mathcal{X}}.$$

The three axioms of the probability calculus are automatically satisfied provided that $f_{\boldsymbol{X}}(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$ and $\Pr(\boldsymbol{X} \in \boldsymbol{\mathcal{X}}) = 1$. For now I skirt over the precise interpretation of $f_{\boldsymbol{X}}$, which is discussed in more detail in section 5. If $\boldsymbol{\mathcal{X}} = \boldsymbol{Y} \times \boldsymbol{Z}$, then the *marginal PMF* for $\boldsymbol{Z}$ is defined as

marginal PMF

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) := \sum_{\boldsymbol{y}} f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y}, \boldsymbol{z}).$$

The *conditional PMF* of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ is defined as the function $f_{\boldsymbol{Y}|\boldsymbol{Z}}$ satisfying

conditional PMF

$$f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y}, \boldsymbol{z}) = f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y} \mid \boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z})$$

[1] Examples where $\boldsymbol{\mathcal{X}} \neq \mathcal{X}^n$ are given in section 11.4.

for all y and z; clearly such a function always exists and is unique excepting the case where $f_Z(z) = 0$. It is conventional to state that $f_{Y|Z}$ is undefined in this case, but also overly restrictive, since $f_Z(z) = 0$ implies that $f_{Y,Z}(y, z) = 0$, and so any value for the conditional PMF would do.

The *Law of Total Probability (LTP)* is a simple extension of the definitions,

Law of Total Probability (LTP)

$$f_Y(y) = \sum_z f_{Y,Z}(y, z) = \sum_z f_{Y|Z}(y \mid z) f_Z(z). \tag{1}$$

The LTP shows that one has the option of specifying f_Y indirectly, in terms of $f_{Y|Z}$ and f_Z. Breaking down complex probability assessments in this way is often helpful in practice. It also shows that $\min_z f_{Y|Z}(y \mid z)$ and $\max_z f_{Y|Z}(y \mid z)$ are lower and upper bounds on $f_Y(y)$, because the righthand side of (1) is a convex combination.[2]

Expectation and variance. Let X be a component of $\boldsymbol{X}$. The *expectation* of X is defined as

expectation

$$E(X) := \sum_x x f_X(x)$$

and the *variance* of X as

variance

$$\text{Var}(X) := E\left(\{X - E(X)\}^2\right).$$

These always exist if X has a bounded sample space, but do not always exist, in more general treatments of random quantities. Note that if $E(X)$ exists, it does not necessarily lie in $\mathcal{X}$. However, it *is* a convex combination of $\mathcal{X}$. Hence if all the elements of $\mathcal{X}$ are non-negative, then $E(X) \geq 0$. This gives rise to the very useful *monotonicity property of expectations*. If X and Y are two components of $\boldsymbol{X}$ and $f_{X,Y}(x, y) = 0$ unless $x \leq y$, then $E(X) \leq E(Y)$.

monotonicity property of expectations

The expectation and variance are related by various inequalities, including *Chebyshev's inequality*: if μ is the expectation of X and σ^2 the variance, then

Chebyshev's inequality

$$\Pr\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}.$$

If g is a convex scalar function of $\boldsymbol{x}$, then *Jensen's inequality* states that

Jensen's inequality

$$E\{g(\boldsymbol{X})\} \geq g(E\{\boldsymbol{X}\}).$$

(The proofs of these easy results are left as exercises.)

If X and Y are two components of $\boldsymbol{X}$, then the *covariance* between X and Y is defined as

covariance

$$\text{Cov}(X, Y) := E\left(\{X - E(X)\}\{Y - E(Y)\}\right),$$

from which we might also have defined the variance of X as $\text{Cov}(X, X)$. It is a general property that

$$\text{Cov}(X, Y)^2 \leq \text{Var}(X)\,\text{Var}(Y),$$

which might be more familiar as the statement that the correlation between X and Y always lies in the interval $[-1, 1]$; see Grimmett and Stirzaker (2001, section 3.6) for a simple proof.

[2] A convex combination of the k elements of set $\mathcal{A}$ has the form $\sum_{j=1}^{k} p_j a_j$ where each $p_j \geq 0$ and $\sum_j p_j = 1$.

Propositions. A proposition is a statement which is either true or false. Thus

propositions

$$a = \text{'I am an astronaut'}$$

is a proposition which is false, and $a(x) = \text{'}x \text{ is an astronaut'}$ is a proposition whose value is contingent on x. The *indicator function* of a proposition a is

indicator function

$$\mathbb{1}(a) := \begin{cases} 0 & \text{if } a \text{ is false} \\ 1 & \text{if } a \text{ is true.} \end{cases}$$

A proposition need not be certain; if $A := a(\boldsymbol{X})$, then A is a random quantity, and the probability that A is true is equal to the expectation of its indicator function,

$$\Pr(A) = \sum_{\boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x}) \mathbb{1}(a(\boldsymbol{x})) = E\{\mathbb{1}(A)\}.$$

The following two results are simple but useful inequalities for propositions, which follow from the link between probabilities and expectations of indicator functions. First, if a and b are two propositions and a implies b, then $\Pr(A) \leq \Pr(B)$.

Proof. So a implies b if and only if $\mathbb{1}(a(\boldsymbol{x})) \leq \mathbb{1}(b(\boldsymbol{x}))$ for all $\boldsymbol{x}$. Then, using the monotonicity property of expectations,

$$\Pr(A) = E\{\mathbb{1}(A)\} \leq E\{\mathbb{1}(B)\} = \Pr(B).$$

□

Second, if $a_1, \dots, a_k$ is a set of propositions, then

$$\Pr(A_1 \vee \cdots \vee A_k) \leq \sum_{i=1}^{k} \Pr(A_i), \tag{2}$$

where '$\vee$' denotes 'or' (inclusive disjunction). This is the *Bonferroni inequality*.

Bonferroni inequality

Proof. Here is the case with $k = 2$, and $a_1 = a$ and $a_2 = b$.

$$\mathbb{1}(a(\boldsymbol{x}) \vee b(\boldsymbol{x})) = \max\{\mathbb{1}(a(\boldsymbol{x})), \mathbb{1}(b(\boldsymbol{x}))\} \leq \mathbb{1}(a(\boldsymbol{x})) + \mathbb{1}(b(\boldsymbol{x})).$$

Then, using the monotonicity property of expectations,

$$\Pr(A \vee B) = E\{\mathbb{1}(A \vee B)\} \leq E\{\mathbb{1}(A) + \mathbb{1}(B)\} = \Pr(A) + \Pr(B).$$

The general case follows by induction. □

1.2 Statistical models

The PMF for $\boldsymbol{X}$ is often indexed by a parameter vector $\theta \in \Omega$, written

$$\Pr(\boldsymbol{X} \in A; \theta) = \sum_{\boldsymbol{x} \in A} f_{\boldsymbol{X}}(\boldsymbol{x}; \theta);$$

other operators dependent on θ, such as the expectation of functions of $\boldsymbol{X}$, are also indexed by θ. The introduction of a fixed family

f_X and a parameter θ is a common strategy to limit the set of possible probability assignments over $\mathcal{X}$ in order to be consistent with one's judgements about $\boldsymbol{X}$. Often the nature of f_X is fairly apparent, or conforms to a conventional situation, and only the value of θ is unknown. Interest then focuses on making inferences about θ based on $\boldsymbol{x}^{\text{obs}}$. The combination of f_X and Ω is termed the *statistical model*. To avoid unnecessary complications, the probability assignment in the statistical model is assumed to be an injective function of θ, i.e. two different θ's cannot give the same probability assignment; in this case the statistical model is said to be *identifiable*.[3]

statistical model

identifiable

I will take Ω to be a convex subset of $\mathbb{R}^d$. Many useful functions in statistical inference have Ω in their domain, such as the score function, (4), and the likelihood function, (5). For these functions I use the argument $t \in \Omega$. Differentiation with respect to t is denoted using '∇' (nabla):

$$\nabla := \begin{pmatrix} \frac{\partial}{\partial t_1} \\ \vdots \\ \frac{\partial}{\partial t_d} \end{pmatrix} \tag{3}$$

This is a notationally confusing area. Some authors do not distinguish between θ, the 'true but unknown parameter value' and t, a point in Ω; this can be confusing for non-experts. The convention on capital letters would suggest that the true parameter be Θ, and the point in Ω could then be θ; Schervish (1995) follows this consistently, but the size of 'Θ' makes it quite intrusive. Another possibility is to use θ_0 for the true value and θ for the point, but this gets confusing when the null hypothesis H_0 arrives.

Those $\boldsymbol{x}$ values for which $f_X(\boldsymbol{x};t) > 0$ are termed the *support of* $\boldsymbol{X}$. Many interesting statistical results concern the special case where the support of $\boldsymbol{X}$ does not depend on the value of t. These are *regular models*.[4] One very useful function for regular statistical models is the *score function*

support of $\boldsymbol{X}$

regular models

score function

$$u(\boldsymbol{x}, t) := \nabla \log f_X(\boldsymbol{x}; t), \tag{4}$$

for which $E\{u(\boldsymbol{X};\theta);\theta\} = \mathbf{0}$.

Proof. Start from the identity $\sum_x f_X(\boldsymbol{x};\theta) = 1$. Then differentiate both sides with respect to θ_j. The righthand side is zero. For a regular model, the lefthand side is

$$\begin{aligned} \frac{\partial}{\partial \theta_j} \sum_{\boldsymbol{x}} f_X(\boldsymbol{x};\theta) &= \sum_{\boldsymbol{x}} \nabla_j f_X(\boldsymbol{x};\theta) \\ &= \sum_{\boldsymbol{x}} \nabla_j \log f_X(\boldsymbol{x};\theta) \times f_X(\boldsymbol{x};\theta) \\ &= E\{u_j(\boldsymbol{X};\theta);\theta\}, \end{aligned}$$

which must therefore equal zero. □

The variance of $u(\boldsymbol{X},\theta)$ is termed the *Fisher information matrix*:

Fisher information matrix

$$i(\theta) := \operatorname{Var}\{u(\boldsymbol{X},\theta);\theta\}$$

[3] A stronger condition is that $f_X(\boldsymbol{x};\theta) \neq f_X(\boldsymbol{x};\theta')$ unless $\theta = \theta'$, for every $\boldsymbol{x} \in \mathcal{X}$. This type of strong condition is only required in precise proofs of consistency.

[4] Technical note. In fact, regular models are defined to be those for which the order of differentiation with respect to θ and summation/integration with respect to $\boldsymbol{x}$ can be reversed. See Casella and Berger (2002, section 2.4) for details. The invariance of the support of $\boldsymbol{X}$ with respect to t is the main necessary condition.

(it is a $d \times d$ matrix when θ is a vector).

A common situation is where the X's are *independent and identically distributed (IID)*, in which case

independent and identically distributed (IID)

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta), \quad \text{written } \boldsymbol{X} \overset{\text{iid}}{\sim} f_X(x;\theta),$$

for some marginal PMF f_X. If the X's are IID and f_X is regular then the score function is a simple sum:

$$u(\boldsymbol{x},t) = \nabla \log \prod_{i=1}^{n} f_X(x_i;t) = \sum_{i=1}^{n} \nabla \log f_X(x_i;t) = \sum_{i=1}^{n} u(x_i,t).$$

In this case $i(\theta) = n i_1(\theta)$, where $i_1(\theta) := \text{Var}\{u(X_1,\theta);\theta\}$.

Nuisance parameters. The statistical model $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ for $\theta \in \Omega$ is meant to be very general, accommodating simple 'student' models such as $(X_1,\ldots,X_n) \overset{\text{iid}}{\sim} \text{N}(x;\mu,\sigma^2)$, but also much more complex models where the IID structure may be buried deep inside the model, and where both $\boldsymbol{X}$ and θ represent collections of possibly quite heterogeneous quantities.[5]

In all but the simplest statistical models some of the parameters are interesting to the scientist, and others are only there for the purposes of structuring the statistical model. The ones that are *not* interesting are referred to as *nuisance parameters*, and the name is well-earned, because they are indeed a nuisance. In order to discuss nuisance parameters at the appropriate places below, I will partition the parameter space as $\Omega = \Omega' \times \Omega''$, where $\theta = (\theta',\theta'')$ and θ'' are the nuisance parameters.

nuisance parameters

This division of θ into two parts is made without loss of generality, as long as the techniques being used are *transformation invariant*. Which is to say, if $g : \theta \mapsto \psi$ is a bijective function, then all inferences made about θ from the model $f_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ are equivalent to inferences made about ψ from the model $f_{\boldsymbol{X}}^{\psi}(\boldsymbol{x};\psi) = f_{\boldsymbol{X}}(\boldsymbol{x};g^{-1}(\psi))$, and *vice versa*. In this case consistent inferences can be made about any function of θ, say $\psi' = g_1(\theta)$, as long as this function can be embedded within a bijective function $\psi = (\psi',\psi'') = (g_1(\theta),g_2(\theta)) = g(\theta)$.

transformation invariant

Transformation invariance is also an important general concern, because the precise representation of the parameters in the statistical model is purely a matter of convention. And in fact in many cases the convention has not settled down, and the same distribution can be parameterised in several different ways. In applied work it is always a good idea to state $f_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ explicitly as a formula, to remove any ambiguity; a statement such as '$\boldsymbol{X}$ is IID Gamma' does not perfectly identify the two components of θ.

1.3 The exponential family of distributions

One very common class of statistical models have the general form

$$g_{\boldsymbol{X}}(\boldsymbol{x};\psi) = f(\boldsymbol{x}) \exp\left\{\psi s(\boldsymbol{x}) - \kappa(\psi)\right\} \quad \text{for } \psi \in \Psi \subset \mathbb{R},$$

[5] A more general treatment would distinguish between the random quantities $\boldsymbol{X}$ and the observations $\boldsymbol{Y}$, where the latter are some function of $\boldsymbol{X}$.

where ψ is a strictly monotone function of θ and s is a scalar function. These are the one parameter *exponential family of distributions*. Common IID distributions based on the Binomial, Poisson, Geometric, Exponential, and Normal (known variance) all belong to this family. One derives a member of the one-parameter exponential family by 'exponentially tilting' a base distribution $f(\boldsymbol{x})$:

exponential family of distributions

$$g_{\boldsymbol{X}}(\boldsymbol{x};\psi) \propto f(\boldsymbol{x}) \exp\{\psi s(\boldsymbol{x})\},$$

where, in order to sum to one,

$$\kappa(\psi) := \log \sum_{\boldsymbol{x}} f(\boldsymbol{x}) \exp\{\psi s(\boldsymbol{x})\}$$

and the parameter space is defined as $\Psi := \{\psi : \kappa(\psi) < \infty\}$. It is straightforward to show that Ψ is a convex subset of $\mathbb{R}$, and that κ is a convex function of ψ (Davison, 2003, sec. 5.2). And also that $f_{\boldsymbol{X}}$ and $g_{\boldsymbol{X}}$ have the same support, which does not depend on ψ.

The exponential family generalises to a vector ψ, with the form

$$g_{\boldsymbol{X}}(\boldsymbol{x};\psi) \propto f(\boldsymbol{x}) \exp\left\{ \textstyle\sum_j \psi_j s_j(\boldsymbol{x}) \right\}.$$

This contains many more familiar distributions, including the Normal, Beta, Gamma, Multinomial, and Dirichlet.

The exponential family has special properties, some of which will be mentioned below. These properties are so special, though, that the exponential family does not provide a general basis for an exploration of statistical inference. When computation was a challenge, the exponential family was favoured because of its simplicity; now, however, there is less need for this type of restriction.

In case you are wondering, in these notes there will be no mention at all of *sufficiency*. I regard sufficiency as a very useful computational device in the case where the model admits a fixed length sufficient statistic. But, as the *Pitman-Koopmans-Darmois theorem* states, a support that does not depend on the parameter plus a fixed length sufficient statistic are more-or-less the characterisation of the exponential family; see Schervish (1995, section 2.2.3). So if I judge the exponential family too restrictive for general consideration, then I can ignore sufficiency.[6]

sufficiency

Pitman-Koopmans-Darmois theorem

1.4 Continuous random quantities

As already stated, $\mathcal{X}$ is taken to be finite, i.e. $\boldsymbol{X}$ is a discrete random quantity. However, *continuous random quantities* with uncountable state spaces are still useful, as an approximation; and because they arise naturally in the Bayesian approach, where the parameter itself (which is generally continuous) acquires a distribution. In the case that $\boldsymbol{X}$ is continuous $f_{\boldsymbol{X}}$ is taken to be a *probability density function (PDF)* with the property that

continuous random quantities

probability density function (PDF)

$$\Pr(\boldsymbol{X} \in A;\theta) = \int_A f_{\boldsymbol{X}}(\boldsymbol{x};\theta)\, \mathrm{d}x$$

[6] A slight regret: there is some beautiful mathematics in sufficiency, for example the Fisher-Neyman factorisation criterion, the Dynkin-Lehmann-Scheffé theorem on minimal sufficient statistics, and the Rao-Blackwell theorem on estimators under convex loss. Never mind.

where $A \subset \Omega$. Discrete and continuous random quantities can be unified within *measure theory*, but we are not going to worry about measure theory in these notes; see Grimmett and Stirzaker (2001) and then Rosenthal (2006) or Kingman and Taylor (1966) for an accessible introduction.

measure theory

One useful result for continuous $\boldsymbol{X}$ is the *transformation of the PDF*. If $\boldsymbol{y} = g(\boldsymbol{x})$ where g is a differentiable bijective function, then

transformation of the PDF

$$f_Y(\boldsymbol{y}) = f_X(\boldsymbol{x}) |J_g(\boldsymbol{x})|^{-1} \quad \text{where } \boldsymbol{x} := g^{-1}(\boldsymbol{y})$$

for $\boldsymbol{y} \in g(\mathcal{X})$, and zero otherwise; here

$$J_g(\boldsymbol{x}) := \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial g_1}{\partial x_n}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial g_n}{\partial x_n}(\boldsymbol{x}) \end{pmatrix}$$

is the *Jacobian matrix* of g and $|J_g(\boldsymbol{x})|$ is its determinant.

Jacobian matrix

Proof. This is just the change of variables formula for multiple integration. $\boldsymbol{Y} \in B$ if and only if $\boldsymbol{X} \in g^{-1}(B)$, and so

$$\Pr(\boldsymbol{Y} \in B) = \int_{g^{-1}(B)} f_X(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_B f_X(g^{-1}(\boldsymbol{y})) |J_{g^{-1}}(\boldsymbol{y})| \, \mathrm{d}\boldsymbol{y} = \int_B f_X(\boldsymbol{x}) |J_g(\boldsymbol{x})|^{-1} \mathrm{d}\boldsymbol{x}$$

where $\boldsymbol{x} := g^{-1}(\boldsymbol{y})$, showing that the integrand is the PDF of $\boldsymbol{Y}$. □

It is common to use f to denote either a probability mass function or a PDF, because many results are the same in either case. Below, though, I will use f for the probability mass function and π for the PDF, notably in section 6 and beyond, where I write π_θ for the PDF of θ.

1.5 The Probability Integral Transform (PIT)

Let $X \in \mathcal{X} \subset \mathbb{R}$ be a scalar random quantity with distribution function $F_X : \mathcal{X} \to [0,1]$. Define a new random quantity $Y := F_X(X)$; i.e. Y is the random quantity one gets by putting X into its own distribution function. It is a very useful fact that Y has a *sub-uniform distribution*, i.e.

sub-uniform distribution

$$F_Y(u) \leq u \quad \text{for all } u \in [0,1],$$

and that $F_Y(u) = u$ if there exists an $x \in \mathcal{X}$ such that $u = F_X(x)$.

Proof. First, consider the case where $u = F_X(x)$ for some $x \in \mathcal{X}$:

$$F_Y(u) = \Pr\{F_X(X) \leq F(x)\} = \Pr\{X \leq x\} = F_X(x) = u.$$

The 'cancellation' of F at the second equality occurs because of the bijective relationship between x and $F(x)$ for $x \in \mathcal{X}$.[7] This proves the second part of the claim: in the case where X is a continuous random quantity, the points u in $(0,1)$ are in a bijective relationship with the points x in $\mathcal{X}$, and Y is uniformly distributed.

[7] Technical note: here we can ignore points in $\mathcal{X}$ that have zero probability.

Otherwise, let x and x' be two consecutive values in $\mathcal{X}$, with $u = F_X(x)$ and $u' = F_X(x')$, and let $u + \delta$ be some value in the open interval (u, u'). Then

$$Y \leq u + \delta \implies X \leq x$$

and so $F_Y(u + \delta) \leq F_X(x) = u$. But we must also have $F_Y(u + \delta) \geq F_Y(u) = u$. Therefore we conclude that $F_Y(u + \delta) = u$, and hence $F_Y(u + \delta) < u + \delta$. □

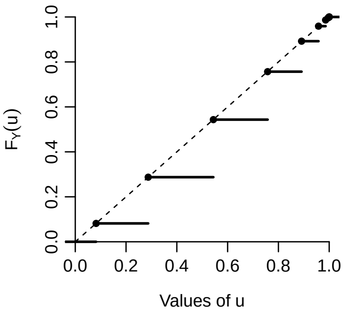

Figure 1: Distribution function of $Y = F_X(X)$, where $X \sim \text{Poisson}(x; \lambda = 2.5)$.

So the distribution function of Y looks like a staircase where each step starts from the 45° line drawn from $(0, 0)$ to $(1, 1)$; see Figure 1. For a continuous random quantity the steps are infinitesimally small, and the distribution function and the 45° line coincide.

1.6 Convergence

There are several different types of convergence for sequences of random quantities. X_n *converges in probability* to Y, written $X_n \xrightarrow{\text{P}} Y$, if

converges in probability

$$\Pr(|X_n - Y| \geq \varepsilon) \longrightarrow 0 \qquad \text{as } n \to \infty$$

for all $\varepsilon > 0$. If $X_1, X_2, \ldots$ is an IID sequence where each X_i has finite expectation μ, then

$$n^{-1}(X_1 + \cdots + X_n) \xrightarrow{\text{P}} \mu.$$

This is the *Weak Law of Large Numbers (WLLN)*. It is easily proved using Chebyshev's inequality in the case where $\text{Var}(X_i)$ is finite. In the WLLN convergence is to a constant, rather than to a random quantity.

Weak Law of Large Numbers (WLLN)

Convergence in distribution is a weaker form of convergence (i.e. it is implied by convergence in probability). X_n *converges in distribution* to Y, written $X_n \xrightarrow{\text{D}} Y$, if

converges in distribution

$$F_{X_n}(y) \longrightarrow F_Y(y) \qquad \text{as } n \to \infty$$

for all y for which F_Y is continuous. The most famous law in statistics, the *Central Limit Theorem (CLT)*, states that if $X_1, X_2, \ldots$ is an IID sequence with finite variance, then

Central Limit Theorem (CLT)

$$\frac{n^{-1}(X_1 + \cdots + X_n) - \mu}{\sigma / \sqrt{n}} \xrightarrow{\text{D}} \text{N}(0, 1),$$

where μ and σ^2 are the mean and variance of each X_i. For convergence in distribution it is common to write a distribution on the right-hand side, as shown above, rather than a Y with a specified distribution.

2 Point and set estimators

Any function of x is termed a *statistic*. A statistic designed to esti-

statistic

mate θ is termed an *estimator*; typically these can be divided into *point estimators*, which map from $\mathcal{X}$ to a point in Ω, and *set estimators*, which map from $\mathcal{X}$ to a set in Ω. If s is a statistic, then $S := s(\boldsymbol{X})$ is a random quantity, whose distribution depends on θ. So, just to clarify, s is a function with domain $\mathcal{X}$, $s(\boldsymbol{x})$ is a value in the range of s, and S is a random quantity with sample space equal to the range of s. Where it is necessary to identify a particular point in the co-domain of s I will use s'; e.g. in expressions of the form $\Pr(S \leq s'; \theta)$.[8]

point estimators

set estimators

If s is an estimator for θ, then one important principle in statistics states that s should be judged by comparing S and θ across the range of possible values of θ. This is the basis of the Frequentist approach to statistics, which is discussed in this section and the two that follow. The Bayesian approach is introduced in section 5.

2.1 *Point estimators*

There are many ways of constructing estimators for θ, but the *maximum likelihood estimator (MLE)* is often preferred for both theoretical and practical reasons, particularly when we are confident about the adequacy of the statistical model.

maximum likelihood estimator (MLE)

The MLE $\hat{\theta}$ satisfies

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\theta}(\boldsymbol{x})) \geq f_{\boldsymbol{X}}(\boldsymbol{x}; t) \qquad \text{for all } t \in \Omega.$$

If $\theta = (\theta_1, \theta_2)$, and the MLE of θ is $\hat{\theta}(\boldsymbol{x}) = \big(\hat{\theta}_1(\boldsymbol{x}), \hat{\theta}_2(\boldsymbol{x})\big)$, then the MLE of θ_1 is defined to be $\hat{\theta}_1(\boldsymbol{x})$. Now we can show that MLEs are transformation-invariant.[9]

Proof. Let $g : \theta \mapsto \psi$ be a bijective function, so that $f^{\psi}_{\boldsymbol{X}}(\boldsymbol{x}; \psi) = f_{\boldsymbol{X}}(\boldsymbol{x}; g^{-1}(\psi))$. Then

$$\begin{aligned} & f^{\psi}_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\psi}) \geq f^{\psi}_{\boldsymbol{X}}(\boldsymbol{x}; v) \quad \text{for all } v \in g(\Omega) \\ \iff & f_{\boldsymbol{X}}(\boldsymbol{x}; g^{-1}(\hat{\psi})) \geq f_{\boldsymbol{X}}(\boldsymbol{x}; t) \quad \text{for all } t \in \Omega \\ \iff & g^{-1}(\hat{\psi}) = \hat{\theta} \\ \iff & \hat{\psi} = g(\hat{\theta}). \end{aligned}$$

This argument also extends to parts of ψ, as long as they can be embedded in a bijective function. Thus, let $\psi = (\psi_1, \psi_2) = \big(g_1(\theta), g_2(\theta)\big)$ where g is bijective as before. Then $\hat{\psi} = \big(g_1(\hat{\theta}), g_2(\hat{\theta})\big)$ as already shown, and hence $\hat{\psi}_1 = g_1(\hat{\theta})$. □

In general, $\hat{\theta}(\boldsymbol{x})$ is an implicit function of $\boldsymbol{x}$, and the value of the MLE is found by maximising the *likelihood function*,

likelihood function

$$L(t) := f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}}; t). \tag{5}$$

The value $\hat{\theta}(\boldsymbol{x}^{\text{obs}}) = \operatorname{argmax}_{t \in \Omega} L(t)$ is termed the *Maximum Likelihood (ML) estimate*. For regular models, the MLE should satisfy the first order conditions

Maximum Likelihood (ML) estimate

$$u(\boldsymbol{x}, \hat{\theta}(\boldsymbol{x})) = \mathbf{0},$$

[8] That s is a function and s' is a value is not ideal, but I think it is the least-bad compromise consistent with standard notation. Another option would be to write s everywhere as $s(\cdot)$, but to me this seems redundant and ugly.

[9] There is a more general definition of the likelihood function of θ_1 which preserves the property of transformation-invariance of the MLE, but also works in the case of non-bijective functions of θ; see Casella and Berger (2002, section 7.2.2).

excepting complications that arise at the boundary of Ω—which would usually be evidence that the model was inadequate. The situation is more complicated for non-regular models. For example, if $\boldsymbol{X} \overset{\text{iid}}{\sim} \mathrm{U}(x; 0, \theta)$, the MLE is well defined but the likelihood function is not differentiable at its maximum.

Predictions. Suppose that we would like to predict the value of an unobserved X, based upon the $\boldsymbol{x}^{\text{obs}}$ that we have. All such predictions are ultimately functions of θ. For example,

$$\Pr(X_{n+1} = x \mid \boldsymbol{X} = \boldsymbol{x}^{\text{obs}}; \theta) = \frac{f_{\boldsymbol{X}, X_{n+1}}(\boldsymbol{x}^{\text{obs}}, x; \theta)}{f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}}; \theta)}. \tag{6}$$

As long as the function on the right-hand side is or can be extended to a bijective function of θ, then the MLE of the probability of $X_{n+1} = x$ must be $f_{\boldsymbol{X}, X_{n+1}}(\boldsymbol{x}, x; \hat{\theta}(\boldsymbol{x})) / f_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\theta}(\boldsymbol{x}))$. Thus we have the general rule, that ML predictions about unobserved X are made by plugging in the value of the MLE. In the special case where $(\boldsymbol{X}, X_{n+1}) \overset{\text{iid}}{\sim} f_X(x; \theta)$, eq. (6) simplifies to

$$\Pr(X_{n+1} = x \mid \boldsymbol{X} = \boldsymbol{x}^{\text{obs}}; \theta) = f_X(x; \theta),$$

and hence the MLE is just $f_X(x; \hat{\theta}(\boldsymbol{x}))$.

2.2 *Judging point estimators*

Suppose some statistic $s : \mathcal{X} \to \Omega$ is claiming to be a point estimator for θ—how do we judge whether it is a good estimator or a poor one? One property has already been mentioned: that the estimator should be transformation-invariant, a property that the MLE possesses. There have been many suggestions for other attractive properties of estimators. I just mention a couple here.

I will use 'estimator' for both the function s and also for the random quantity $S := s(\boldsymbol{X})$. The estimator S is *unbiased* if

unbiased estimator

$$\text{bias}(S; \theta) := E(S; \theta) - \theta$$

is zero for all $\theta \in \Omega$. This is a superficially attractive criterion but leads to daft results even in simple cases, where it exists, and often does not exist. Thus the unique unbiased estimator of θ in $X \sim \text{Geometric}(x; \theta)$ is daft (Cox and Hinkley, 1974, section 8.2),[10] and the unbiased estimator of θ in $X \sim \text{Exponential}(x; \theta)$ does not exist (Schervish, 1995, section 5.1).[11] Furthermore, unbiasedness is not consistent with transformation-invariance. For example, if S is an unbiased estimator of scalar positive θ, then $1/S$ is not an unbiased estimator of $1/\theta$, because $E\{1/S; \theta\} \neq 1/E\{S; \theta\} = 1/\theta$.

A better criterion is that S has a small *mean squared error (MSE)*,

mean squared error (MSE)

$$\text{MSE}(S; \theta) := E\{(S - \theta)^2; \theta\} = \text{bias}(S; \theta)^2 + \text{Var}(S; \theta), \tag{7}$$

where the second expression follows from a simple rearrangement. The squared error is is one example of a *loss function* for point estimation, and the MSE is then the *risk*; see section 8.

[10] It is $\hat{\theta}(x) = \mathbb{1}(x = 1)$.

[11] A much more technical point is that from the point of view of Decision Theory, some unbiased estimators, where they exist, are *inadmissible*. This is demonstrated by the *Stein effect*, discussed below in section 8.2.

There is an interesting relationship between the bias and the variance, the *Cramér-Rao lower bound (CRLB)*. Taking θ scalar, for simplicity,

Cramér-Rao lower bound (CRLB)

$$\mathrm{Var}\{S;\theta\} \geq \frac{\{1+\nabla\,\mathrm{bias}(S;\theta)\}^2}{i(\theta)} \tag{8}$$

where $\nabla\,\mathrm{bias} := \mathrm{d}\,\mathrm{bias}\,/\mathrm{d}\theta$; the Fisher Information $i(\theta)$ was defined in section 1.2. The CRLB holds for regular models.

Proof of (8), *with scalar θ.* Correlation coefficients lie in the interval $[-1,1]$, or, in our case,

$$\mathrm{Cov}\{S, u(\boldsymbol{X},\theta);\theta\}^2 \leq \mathrm{Var}\{S;\theta\}\,\mathrm{Var}\{u(\boldsymbol{X},\theta);\theta\};$$

see section 1.1. As $E\{u(\boldsymbol{X},\theta);\theta\} = 0$ for regular models,

$$\begin{aligned}
\mathrm{Cov}\{S, u(\boldsymbol{X},\theta);\theta\} &= E\{S u(\boldsymbol{X},\theta);\theta\} \\
&= \sum_{\boldsymbol{x}} s(\boldsymbol{x}) \frac{\mathrm{d}}{\mathrm{d}\theta}\Big\{\log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)\Big\} f_{\boldsymbol{X}}(\boldsymbol{x};\theta) \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{\boldsymbol{x}} s(\boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x};\theta) \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\, E\{S;\theta\} \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\{\theta + \mathrm{bias}(S;\theta)\} = 1 + \nabla\,\mathrm{bias}(S;\theta).
\end{aligned}$$

The result follows from the inequality at the top of the proof.

The generalisation to vector θ is not completely straightforward, excepting the obvious case where $s(\boldsymbol{x})$ is an estimator of one component of θ. See Cox and Hinkley (1974, section 8.3). □

Estimators that attain the CRLB are termed *efficient estimators*. It would seem to be a good policy to use efficient estimators where they exist, but in fact they are a rather special case: only a subset of the exponential family can possess efficient estimators.

efficient estimators

Proof, with scalar θ. From the previous proof, the CRLB is attained if and only if the correlation between $u(\boldsymbol{X},\theta)$ and $s(\boldsymbol{X})$ is ± 1 for any specified θ. Equivalently,

$$u(\boldsymbol{x},\theta) = a(\theta)\{s(\boldsymbol{x}) - b(\theta)\},$$

where necessarily $b(\theta) = E(s(\boldsymbol{X})) = \theta + \mathrm{bias}(S;\theta)$. Integrating this expression over θ shows that the statistical model has the general form

$$\log f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = A(\theta)s(\boldsymbol{x}) + B(\theta) + C(\boldsymbol{x}),$$

i.e. $f_{\boldsymbol{X}}$ is special case of the exponential family.

Again, there is a generalisation to vector θ, although it is complicated in the same way as the CRLB is complicated in this case. □

The MSE is an intuitive criterion with which to judge point estimators, but it is important to be aware that unbiased and efficient estimators do not necessarily minimise the MSE for regular models. This is because some biased estimators increase the first term in (7) but decrease the second term by more; the CRLB shows that these estimators must have $\nabla \operatorname{bias}(S;\theta) < 0$. The Stein effect provides a good illustration; see section 8.3.

2.3 Set estimators

Set estimators of θ are also useful: crudely, their volume represents the variability in the estimator that arises from having only a finite sample. Let $\mathcal{C}$ be a function from $\mathcal{X}$ to a subset of Ω, i.e. $\mathcal{C}(\boldsymbol{x}) \subset \Omega$. The *coverage of* $\mathcal{C}$ is the probability that $\mathcal{C}(\boldsymbol{X})$ will contain θ, which is a function of θ. $\mathcal{C}$ is a *level* $(1-\alpha)$ *confidence set* for θ if its coverage is at least $1-\alpha$ for all θ. That is,

coverage of $\mathcal{C}$

level $(1-\alpha)$ confidence set

$$\Pr\{\theta \in \mathcal{C}(\boldsymbol{X}); \theta\} \geq 1-\alpha \qquad \text{for all } \theta \in \Omega.$$

Here $(1-\alpha)$ is termed the *nominal level* of the confidence set. When the converage equals the nominal level for all θ, $\mathcal{C}$ is said to be an *exact confidence set*. Typically, though, this only happens in one or two special cases, or in the limit as $n \to \infty$ for IID X's from regular models; see below, around (9). In the case where θ is a scalar parameter and $\mathcal{C}(\boldsymbol{x})$ is convex for all $\boldsymbol{x}$, $\mathcal{C}$ is a *confidence interval* for θ.

nominal level

exact confidence set

confidence interval

In the case of nuisance parameters, i.e. where $\theta = (\theta', \theta'')$ but only θ' is interesting, a level $(1-\alpha)$ confidence set for θ' is any function $\mathcal{C}'$ from $\mathcal{X}$ to subsets of Ω', with the property that

$$\Pr\{\theta' \in \mathcal{C}'(\boldsymbol{X}); \theta\} \geq 1-\alpha \qquad \text{for all } \theta \in \Omega.$$

It is reassuring that such *marginal confidence sets* can always be constructed from full confidence sets. Define the projection $\mathbb{P}$ as $\mathbb{P}(t) = t'$. If $\mathcal{C}$ is a $(1-\alpha)$ confidence set for θ, then $\mathcal{C}' := \mathbb{P}\mathcal{C}$ is a $(1-\alpha)$ confidence set for θ'.

marginal confidence sets

Proof. $\theta \in \mathcal{C}(\boldsymbol{x})$ implies $\theta' \in \mathcal{C}'(\boldsymbol{x})$, and hence, for all θ,

$$1-\alpha \leq \Pr\{\theta \in \mathcal{C}(\boldsymbol{X}); \theta\} \leq \Pr\{\theta' \in \mathcal{C}'(\boldsymbol{X}); \theta\},$$

using the result on propositions given in section 1.1. □

Of course, just because such confidence sets exist, does not mean that they have attractive properties (like having small volumes). This is discussed further in section 8.1.

This definition of confidence sets is all very well in theory, but set estimators with a known confidence level are extremely hard to construct in practice, except in very simple statistical models, and for uninformative confidence sets.[12] On the other hand, for any reasonable confidence set, it is extremely hard to work out the infimum of the coverage over Ω, and it is even harder to reverse the problem, and identify a confidence set for which the infimum of the coverage is above a target level. It is harder still to do this in the presence of nuisance parameters.

In general, the best that can be done is construct confidence sets that have approximately the right coverage in special cases, notably

[12] I.e., the equivalent of the uninformative statistical tests mentioned below in section 4.1.

where the X's are IID and the statistical model is regular. The simplest approach is to use asymptotic theory (i.e. $n \to \infty$). Here it is important to insert a *caveat*. Resorting to asymptotic constructions is an *act of desperation*, because in reality datasets are stubbornly finite, and often quite small. It is a regrettable feature of Frequentist inference that, except in a few special cases, asymptotic arguments are all that are available. The Bayesian approach makes no particular use of asymptotic arguments, but makes additional demands on the scientist that some statisticians are unwilling to accept, as discussed in section 5.

Here is one important example, based on the score function. If the X's are IID and the statistical model is regular, then the CLT implies that $u(\boldsymbol{X}, \theta) \xrightarrow{\text{D}} \text{N}_d\big(\mathbf{0}, i(\theta)\big)$, and this in turn implies that[13]

$$w(\boldsymbol{X}, \theta) := u(\boldsymbol{X}, \theta)^T i(\theta)^{-1} u(\boldsymbol{X}, \theta) \xrightarrow{\text{D}} \chi^2_d, \tag{9}$$

remembering that $u(\boldsymbol{x}, \theta)$ is a d-dimensional vector of partial derivatives, and $i(\theta)$ is a $d \times d$ variance matrix. Then

$$\mathcal{C}(\boldsymbol{x}) := \big\{t : w(\boldsymbol{x}, t) \leq F^{-1}_{\chi^2_d}(1-\alpha)\big\} \tag{10}$$

is asymptotically an exact $(1-\alpha)$ confidence set for θ, where $F^{-1}_{\chi^2_d}$ is the quantile function of the χ^2_d distribution.

Proof.

$$\begin{aligned} \Pr\big\{\theta \in \mathcal{C}(\boldsymbol{X}); \theta\big\} &= \Pr\big\{w(\boldsymbol{X}, \theta) \leq F^{-1}_{\chi^2_d}(1-\alpha); \theta\big\} \\ &= \Pr\big\{F_{\chi^2_d}(w(\boldsymbol{X}, \theta)) \leq 1-\alpha; \theta\big\} \longrightarrow 1-\alpha \end{aligned}$$

for all θ, applying the PIT (section 1.4) and (9). □

This confidence set is transformation-invariant.

Proof. Take θ scalar, for simplicity. Apply the chain rule to show that $u^\psi(\boldsymbol{x}, \psi) = u(\boldsymbol{x}, \theta)(g'(\theta))^{-1}$, where $g : \theta \mapsto \psi$ is bijective and differentiable, $g' := (\mathrm{d}/\mathrm{d}\theta)g$ and $\theta = g^{-1}(\psi)$. Then it follows that $w(\boldsymbol{x}, \theta) = w^\psi(\boldsymbol{x}, \psi)$ for all $\boldsymbol{x}$.

The general proof is the same, but more fiddly. □

Thus we have established, at least in the case where the X's are IID and the statistical model is regular, that transformation-invariant confidence sets for θ (and consequently for θ') can be approximated, and that this approximation ought to be reasonable when n is large.

What about when n is not large? In that case the nominal coverage of a set such as $\mathcal{C}$ in (10) may differ substantially from the actual coverage, in a way that varies with θ. A computer-intensive resampling approach termed *the bootstrap* provides an elegant correction for this, notably the *prepivoting* approach of Beran (1987). The bootstrap is one of the two most important innovations in computational statistics in the last forty years.

the bootstrap

prepivoting

The other being Markov Chain Monte Carlo (MCMC); see section 6.6.

[13] This is a standard result in Normal distribution theory; see Mardia *et al.* (1979, chapter 3).

This is not the place to review the extensive and still-developing literature on the bootstrap; instead, see the review of Davison *et al.* (2003, section 4)—and the other papers in this number of *Statistical Science*, celebrating the silver anniversary of the bootstrap—and Young and Smith (2005, chapter 11); the broader perspective in Efron (1998) is also interesting.

One loose end needs to be tidied up. The Fisher information $i(\theta)$ in (9) can be replaced by an estimate, for convenience, and the asymptotic properties still hold as long as the estimator converges appropriately for large n. One suggestion is the *expected Fisher Information*, $i(\hat{\theta}(\boldsymbol{x}^{\text{obs}}))$. As shown in section 7.1, $\hat{\theta}(\boldsymbol{X}) \xrightarrow{\text{P}} \theta$ for IID X's from a regular model, and so $i(\hat{\theta}(\boldsymbol{X})) \xrightarrow{\text{P}} i(\theta)$, because i is a continuous function of θ. In practice, however, a different approximation is often preferred, the *observed Fisher Information*, denoted J, which is defined below in section 7.2. Efron and Hinkley (1978) provide a theoretical justification for preferring J, which is that J is closer to $\text{Var}(u(\boldsymbol{X},\theta) \mid s(\boldsymbol{X}) = s^{\text{obs}}; \theta)$, where s is an experimental ancillary statistic (see section 11.5).

expected Fisher Information

3 Goodness of fit tests

Tests of goodness of fit are also known as (pure) *significance tests*. One starts with a hypothesis about the data $\boldsymbol{X}$, and then examines whether the observations $\boldsymbol{x}^{\text{obs}}$ are consistent with that hypothesis. Cox and Hinkley (1974, chapter 3) provide more details.

significance tests

3.1 Simple hypothesis

A *simple hypothesis* completely defines the distribution of $\boldsymbol{X}$:

simple hypothesis

$$H_0 : \boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}).$$

A simple hypothesis is evaluated according to a *P-value*, which is any statistic with a sub-uniform distribution under H_0. That is, a P-value is a statistic p with the property that if $P := p(\boldsymbol{X})$ then

P-value

$$\Pr(P \le u; H_0) \le u \qquad \text{for all } u \in [0,1].$$

Thus, a small $p^{\text{obs}} := p(\boldsymbol{x}^{\text{obs}})$, say $p^{\text{obs}} = 1\%$, indicates an outcome that would be improbable under H_0.

The standard way to construct a P-value is using a statistic s with the property that large values of $s(\boldsymbol{x})$ indicate a worrying—or at least interesting—departure from the hypothesis H_0. Letting $S := s(\boldsymbol{X})$, the P-value is defined as

$$p(s') := \Pr(S \ge s'; H_0).$$

Under H_0 the random quantity $P = p(S)$ has a *sub-uniform* distribution in general, and a uniform distribution if S is continuous.

Proof. This proof uses a nifty trick from Casella and Berger (2002, section 8.3.4). Let G be the distribution function of $-S$ under H_0.

Then

$$p(s') = \Pr(S \geq s'; H_0) = \Pr(-S \leq -s'; H_0) = G(-s').$$

Then since $P = p(S) = G(-S)$, the result follows from the PIT, see section 1.5. □

Note that one can get a large P-value by choosing a poor test statistic. For example, $s(\boldsymbol{x}) = a$ where a is any scalar constant will have a P-value of 1, no matter what the value of $\boldsymbol{x}$. So there is no basis to claim, without additional information, that a large P-value supports H_0.

It is important to be able to choose the test statistic in the light of possible alternatives to H_0. But where this is challenging, one useful portmanteau test statistic for a simple hypothesis is *Box's test statistic* (Box, 1980). This is $s(\boldsymbol{x}) = -f_{\boldsymbol{X}}(\boldsymbol{x})$. Thus Box's P-value would be

Box's test statistic

$$p^{\text{obs}} = \Pr(f_{\boldsymbol{X}}(\boldsymbol{X}) \leq f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}}); H_0).$$

A small p^{obs} indicates that $\boldsymbol{x}^{\text{obs}}$ is in the tail of the distribution specified by H_0, because $f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}})$ is relatively small.

The P-value for a simple hypothesis can easily be computed by *Monte Carlo integration*. One samples $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(r)} \overset{\text{iid}}{\sim} f_{\boldsymbol{X}}(\boldsymbol{x})$, and then computes

Monte Carlo integration

$$\hat{P}\big(\boldsymbol{x}^{\text{obs}}; \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(r)}\big) := r^{-1} \sum_{j=1}^{r} \mathbb{1}(s^{(j)} \geq s^{\text{obs}}),$$

where $s^{(j)} := s(\boldsymbol{x}^{(j)})$ and $s^{\text{obs}} := s(\boldsymbol{x}^{\text{obs}})$. According to the WLLN,

$$\hat{P}\big(\boldsymbol{x}^{\text{obs}}; \boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(r)}\big) \xrightarrow{\text{P}} E\{\mathbb{1}(S \geq s^{\text{obs}}); H_0\} = p^{\text{obs}}$$

as $r \to \infty$, and confidence intervals for the estimator with finite r can be established using the CLT.

3.2 *Composite hypothesis*

The hypothesis might also be a *composite hypothesis*,

composite hypothesis

$$H_0 : \exists \theta \in \Omega_0 \text{ such that } \boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}; \theta),$$

which involves additional unknown parameters θ. With $\theta = (\theta', \theta'')$, where θ'' are nuisance parameters, often only θ' is specified, in which case $\Omega_0 = \{\theta'_0\} \times \Omega''$. For a composite hypothesis, one common definition of the P-value is

$$p_{\Omega_0}(s') := \sup_{t \in \Omega_0} p(s'; t)$$

where $p(s'; t)$ is a P-value for the simple hypothesis $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}; t)$. This has a sub-uniform distribution under all points in Ω_0, i.e.

$$\Pr\{p_{\Omega_0}(S) \leq u; \theta\} \leq u \qquad \text{for all } \theta \in \Omega_0.$$

Proof. Follows from the fact that $p_{\Omega_0}(s') \leq u$ implies $p(s';\theta) \leq u$ for all $\theta \in \Omega_0$. Therefore

$$\Pr\{p_{\Omega_0}(S) \leq u;\theta\} \leq \Pr\{p(S;\theta) \leq u;\theta\} \leq u$$

for all $\theta \in \Omega_0$. □

One special case is worth mentioning, when θ is a scalar parameter. If the statistical model has a Monotone Likelihood Ratio (MLR) in the test statistic s, then the result that the power function is non-decreasing in θ implies that

See ahead to section 4.2 for the appropriate definitions and results.

$$\sup_{t \in \Omega_0} p(s';\theta) = p(s';\sup \Omega_0).$$

This case shows that this definition of a P-value for a composite H_0 often gives large P-values from which nothing interesting can be inferred, and modifications are required unless Ω_0 is quite constrained.

Pearson's chi-squared test

As an alternative, *Pearson's chi-squared test* (or goodness of fit test) works for both simple and composite null hypotheses, but only in the case where the X's are IID from a regular model. This involves a particular test statistic, in which the sample space $\mathcal{X}$ is first partitioned into k bins, and then

$$s(\boldsymbol{x}) = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j}$$

where o_j is the observed number in the jth bin, and e_j is the expected number in the jth bin under H_0. For a composite null hypothesis with a d-dimensional unknown parameter, each e_j is computed using the value of the MLE:

$$e_j = n \sum_{x \text{ in bin } j} f_X\big(x;\hat{\theta}(\boldsymbol{x}^{\text{obs}})\big).$$

Then $S \xrightarrow{\text{D}} \chi^2_{k-d-1}$ under H_0 (not proved here), from which a P-value can be computed.[14] A commonly-used heuristic is to take the asymptotic limit as reliable provided that $e_j \geq 5$ for all j.

Pearson's chi-squared test has the disadvantage that the result will depend on the partition of $\mathcal{X}$. It is also not very powerful, in that there is no option to tune the choice of test statistic to a particular departure from the statistical model. But it is available in standard statistical software, so it crops up a lot in applied statistics.

A final caveat. A statistical hypothesis, whether simple or composite, can never really be 'true'. A P-value tries to address the question of whether a proposed hypothesis is adequate, but it is hard to do this without an alternative model to contrast H_0 with, as addressed in section 4 and section 9. If you have a large enough n then your P-value is very likely to be small, unless you allow

[14] Or $T \xrightarrow{\text{D}} \chi^2_{k-1}$ for a simple hypothesis. The rule is to subtract one from the degrees of freedom of the χ^2 distribution for each parameter that is replaced by the value of its MLE.

the model complexity (i.e. the number of parameters) to increase with n.

Just as it is important not to get excited when your P-value is large, it is also important not to get excited when your P-value is small, if your n is large. In this situation, your model may be perfectly adequate for your purposes, even though it has not captured every wrinkle in the observations. For example, it is a mistake to conclude that paranormal effects exist just because your P-value is 0.0003 in a test with $n = 104{,}490{,}000$; see Jefferys (1990).

4 Hypothesis tests

A goodness of fit test has a single hypothesis, and the question is whether or not the observations are consistent with it. Hypothesis tests contrast competing hypotheses, which may be represented in terms of the values of the parameter of the statistical model, the model itself being taken as adequate. Thus $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}; \theta)$ for some $\theta \in \Omega$, and the competing hypotheses are represented as

$$H_0 : \theta \in \Omega_0 \quad \text{and} \quad H_1 : \theta \in \Omega_1$$

where Ω_0 and Ω_1 are both subsets of Ω, and do not intersect. The issue is then to devise a sensible rule, based on $\boldsymbol{x}$, that allows one to choose between H_0 and H_1. Lehmann and Romano (2005) is the standard reference for hypothesis testing.

In this situation it would be natural for Ω_0 and Ω_1 to partition Ω, but in fact a different approach is often used. Ω_0 is taken to be the conventional or default set of values for θ, and termed the *null hypothesis*, and Ω_1 to be the values which would indicate a particular type of departure from H_0, such as an improved treatment, and termed the *alternative hypothesis*. A very common example is where θ is a scalar, and

null hypothesis

alternative hypothesis

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta > \theta_0$$

so that $\Omega_0 = \{\theta_0\}$ and $\Omega_1 = \{t \in \Omega : t > \theta_0\}$. Thus H_0 is a simple hypothesis and H_1 is a composite hypothesis.

4.1 Basic concepts

The basis of a hypothesis test is a *rejection region*, denoted $\mathcal{R} \subset \mathcal{X}$. Elements of $\mathcal{R}$ indicate *rejecting* the null hypothesis in favour of the alternative hypothesis, while elements of $\mathcal{X} \setminus \mathcal{R}$ suggest *accepting* the null hypothesis; both rejecting and accepting are provisional, rather than final. What makes a hypothesis test fundamentally different from a goodness of fit test is that the rejection region is chosen with both H_0 and H_1 in mind.

rejection region

A rejection region is described statistically in terms of its *power function*,

power function

$$\pi(\theta) := \Pr(\boldsymbol{X} \in \mathcal{R}; \theta).$$

The *size of the test* is

size of the test

$$\alpha := \sup_{\theta \in \Omega_0} \pi(\theta)$$

and the test is said to be of *level* α_0 if $\alpha \le \alpha_0$.

level of the test

When H_0 is a simple hypothesis, so that $\Omega_0 = \{\theta_0\}$, $\alpha = \pi(\theta_0)$ is termed the *Type 1 error rate*, where a Type 1 error is the mistake of incorrectly rejecting the null hypothesis. Where H_1 is a simple hypothesis, so that $\Omega_1 = \{\theta_1\}$,

Type 1 error rate

$$\beta := 1 - \pi(\theta_1)$$

is the *Type 2 error rate*, where a Type 2 error is the mistake of incorrectly accepting the null hypothesis.

Type 2 error rate

In medical science, $1 - \alpha$ is termed the *sensitivity* of the test, and $1 - \beta$ the *specificity*; see section 10.2.

The uninformative test. The following test is useful for clarifying ideas. Suppose I have a coin which when tossed comes up heads with probability α. I construct a test which rejects H_0 in favour of H_1 if and only if the coin comes up heads. This test has $\pi(\theta) = \alpha$ for all $\theta \in \Omega$. Thus for any pair of simple hypotheses, this is a size α test with $\beta = 1 - \alpha$. This test is completely uninformative, because the decision to reject or accept is not influenced by $\boldsymbol{x}$ at all. In order to beat the uninformative test, an informative test must have $\pi(\theta_1) > \alpha$. For a composite H_1, we would want $\pi(\theta_1) \ge \alpha$ for all $\theta_1 \in \Omega_1$, with strict inequality for some of Ω_1; otherwise, we might as well just toss a coin.

Tests for which $\pi(\theta_1) \ge \pi(\theta_0)$ for all $\theta_0 \in \Omega_0$ and all $\theta_1 \in \Omega_1$ are termed *unbiased tests*.

unbiased tests

4.2 *Powerful tests*

Consider the case where both hypotheses are simple. The *Neyman-Pearson (NP) Lemma* states that of all hypothesis tests with size α, none has a smaller β (i.e. larger power for θ_1) than the one with rejection region

Neyman-Pearson (NP) Lemma

$$\mathcal{R} := \left\{ \boldsymbol{x} : \frac{f_X(\boldsymbol{x}; \theta_1)}{f_X(\boldsymbol{x}; \theta_0)} \ge c \right\}. \tag{11}$$

Proof. Consider selecting $\mathcal{R}$ to minimise the linear combination $c\alpha + \beta$:

$$\begin{aligned} c\alpha + \beta &= c \sum_{\boldsymbol{x} \in \mathcal{R}} f_X(\boldsymbol{x}; \theta_0) + \Big(1 - \sum_{\boldsymbol{x} \in \mathcal{R}} f_X(\boldsymbol{x}; \theta_1) \Big) \\ &= \sum_{\boldsymbol{x} \in \mathcal{R}} \big\{ c f_X(\boldsymbol{x}; \theta_0) - f_X(\boldsymbol{x}; \theta_1) \big\} + 1 \end{aligned}$$

which is minimised by taking as $\mathcal{R}$ exactly those $\boldsymbol{x}$ for which the term in curly brackets is non-positive, which is (11). Now fix c, which determines $\mathcal{R}$, which determines α and β. For any other rejection region $\mathcal{R}'$, with α' and β', we must have

$$c\alpha + \beta \le c\alpha' + \beta'$$

and, consequently, if $\alpha' = \alpha$ then $\beta \le \beta'$. □

Uniformly most powerful (UMP) tests. Now consider a scalar θ and the very common hypotheses

$$H_0 : \theta = \theta_0 \quad \text{and} \quad H_1 : \theta > \theta_0$$

i.e. where there is a composite alternative hypothesis. A *uniformly most powerful (UMP) test* is most powerful for all points in H_1. There is useful sufficient condition for a UMP test. A statistical model has a *monotone likelihood ratio (MLR)*, if there exists a statistic s with the property that $f_X(x; \theta_1) / f_X(x; \theta_0)$ is a non-decreasing function of $s(x)$ when $\theta_1 > \theta_0$. In this case, the rejection region

uniformly most powerful (UMP) test

monotone likelihood ratio (MLR)

$$\mathcal{R} = \{x : s(x) \geq c'\}$$

is a UMP test.

Proof. Fix $\theta_0 < \theta_1$. Then, according to the MLR property, for every c it is possible to find a c' such that

$$\frac{f_X(x; \theta_1)}{f_X(x; \theta_0)} \geq c \iff s(x) \geq c';$$

so these two rejection regions are equivalent, and yet the second makes no reference to θ_0 or θ_1. Thus this test is simultaneously most powerful for any θ_0 and all elements of H_1. □

An interesting feature of UMP tests is that the power function is non-decreasing.

Proof. Fix $\theta_0 < \theta_1$. Let α be the size of a UMP test for $H_0 : \theta = \theta_0$, and let $\beta = 1 - \pi(\theta_1)$ be its Type 2 error. Now consider the uninformative test with size $\alpha' = \alpha$, and with $\beta' = 1 - \alpha' = 1 - \alpha$. For a UMP test we must have $\beta \leq \beta'$, i.e. $1 - \pi(\theta_1) \leq 1 - \pi(\theta_0)$, and it follows immediately that $\pi(\theta_1) \geq \pi(\theta_0)$. □

So, for example, for a UMP test with $H_0 : \theta \leq \theta_0$, the size of the test is $\pi(\theta_0)$. A non-decreasing power function implies that UMP tests are unbiased when every point in Ω_0 is smaller than every point in Ω_1.

UMP tests are very rare in theory, although common in practice. MLRs are found in one-parameter exponential family distributions, see section 1.3. But there is also a partial converse to this result, which states that if the statistical model is IID and regular and a UMP test exists for some α and all n, then this statistical model belongs to the one-parameter exponential family (Pfanzagl, 1968).[15] So, they are rare in theory because the exponential family is such a special statistical model, but they are common in practice because we so often choose members of the exponential family for our statistical models.

[15] You do not need to read this paper!

4.3 *P-values for hypothesis tests*

In the ideal case, the experimenter designs his experiment to meet targets for the power function.[16] This is what happens in clinical trials, for which typically $H_0 : \theta = \theta_0$ and $H_1 : \theta > \theta_0$. The power in Ω_1 is specified in terms of the *minimal clinically relevant difference*. This requires a particular value $\theta_1 \in \Omega_1$, the smallest value for which the clinician is prepared to assert that the treatment has made a meaningful difference. Then the experimenter looks for the combination of n and $\mathcal{R}$ for which n is the smallest, while satisfying $\pi(\theta_0) \leq \alpha_0$ and $1 - \pi(\theta_1) \leq \beta_0$. Typical values would be $\alpha_0 = 5\%$ and $\beta_0 = 20\%$.

Most statisticians, however, do not get to control the size of their sample, which is simply presented to them as $\boldsymbol{x}^{\text{obs}}$. Then the *Neyman-Pearson approach* is (i) to specify the hypotheses so that the more serious error is the Type 1 error, (ii) set $\mathcal{R}$ to satisfy a target α_0, and (iii) not worry too much about the power of the test in Ω_1, beyond favouring the use of powerful test statistics such as ratios of probabilities, as suggested by the NP Lemma.

Which is a shame, because experimental design is one area where statisticians can add a huge amount of value in science. Cox (1958) provides a non-technical exposition of experimental design.

There is a more modern twist on this though. Because choosing a particular value for α_0 is a bit subjective, hypothesis tests are typically reported in terms of the *smallest* size test at which H_0 is rejected. This turns out to be a P-value.

In practice, though, a threshold such as $\alpha_0 = 5\%$ is still often used, with P-values less than this threshold being declared "statistically significant" rejections of H_0.

Proof. Almost all rejection regions can be represented in the form $\mathcal{R} = \{\boldsymbol{x} : s(\boldsymbol{x}) \geq c\}$ for some test statistic s and critical value c. The α of this rejection region is decreasing in c for each θ. So the smallest α for which $\boldsymbol{x}^{\text{obs}} \in \mathcal{R}$ under $H_0 : \theta \in \Omega_0$ is

$$\begin{aligned} \alpha^* &:= \inf_{c : t^{\text{obs}} \geq c} \sup_{\theta \in \Omega_0} \Pr\{s(\boldsymbol{X}) \geq c; \theta\} \\ &= \sup_{\theta \in \Omega_0} \Pr\{s(\boldsymbol{X}) \geq t^{\text{obs}}; \theta\} \\ &= p_{\Omega_0}(t^{\text{obs}}) \end{aligned}$$

according to the definition in section 3.2. □

Thus hypothesis tests, which are distinguished by having both a null and an alternative hypothesis, are often reported just like goodness of fit tests, in terms of the P-value of the null hypothesis; the alternative hypothesis is important only insofar as it suggests a good choice of rejection region (i.e. a good choice of test statistic). But this difference seems to permit us to accept the null hypothesis for large P-values in a hypothesis test, in a way that contradicts the assertion made at the start of section 3 that large P-values in goodness of fit tests tell us nothing about the null hypothesis. This just goes to show that the choice of test statistic is crucial in interpreting a P-value.

Computing P-values. P-values can be computed using the *duality of confidence sets and rejection regions*. This states that a confidence set can be used to construct a rejection region, and *vice versa*.

duality of confidence sets and rejection regions

[16] Although see section 8 and section 9.2 for a more thorough approach incorporating costs of errors.

Proof. This proof is for the general case of a parameter $\theta = (\theta', \theta'')$ and a composite null hypothesis of the form $H_0 : \theta' = \theta'_0$. Suppose we have a $(1 - \alpha)$ confidence set for θ', i.e. a $\mathcal{C}'$ with the property $\Pr\{\theta' \in \mathcal{C}'(\boldsymbol{x}); \theta\} \geq 1 - \alpha$ for all $\theta \in \Omega$. Then define

$$\mathcal{R} = \{\boldsymbol{x} : \theta'_0 \notin \mathcal{C}'(\boldsymbol{x})\}.$$

The size of the test is

$$\begin{aligned}
\sup_{\theta''} \Pr\{\boldsymbol{X} \in \mathcal{R}; \theta'_0, \theta''\} &= \sup_{\theta''} \Pr\{\theta'_0 \notin \mathcal{C}'(\boldsymbol{X}); \theta'_0, \theta''\} \\
&= 1 - \inf_{\theta''} \Pr\{\theta'_0 \in \mathcal{C}'(\boldsymbol{X}); \theta'_0, \theta''\} \\
&\leq 1 - (1 - \alpha) = \alpha.
\end{aligned}$$

Therefore the decision to reject the null hypothesis when $\theta'_0 \notin \mathcal{C}'(\boldsymbol{x})$ is a test with level α. The same argument also works in reverse. □

As already discussed in section 2.3, the bootstrap is a tool for generating confidence sets with approximately correct coverage. Therefore the bootstrap can also be used to perform hypothesis tests.[17] The α of the bootstrap confidence set $\mathcal{C}'$ is adjusted until θ'_0 is just outside $\mathcal{C}'(\boldsymbol{x}^{\text{obs}})$, and then this α is the P-value of the test $H_0 : \theta' = \theta'_0$.

4.4 *Multiple testing / screening*

In the previous subsections there was one null hypothesis to be tested. As the name suggests, *multiple testing* concerns the situation where many null hypotheses are being tested simultaneously, and where it is helpful to know which of the individual hypotheses is rejected. One common special case is *screening*, where the objective is to identify a subset of the hypotheses that is likely to contain the rejections. So, for example, in a statistical model of the form $\boldsymbol{X} - \theta \overset{\text{iid}}{\sim} \mathrm{N}(x; 0, \sigma^2)$ where $\theta = (\theta_1 \ldots, \theta_n)$, the simultaneous null hypothesis might be

multiple testing

screening

$$H_0 : \theta_1 = \cdots = \theta_n = 0$$

but it might be helpful to test this hypothesis by computing a P-value for each $H_0^i : \theta_i = 0$, in order to identify the i's for which the null hypothesis is rejected.

Suppose, without loss of generality, that there are n hypotheses to be tested, as in the example above, each one giving rise to a P-value p_i. The *Family-Wise Error Rate (FWER)* is the probability of at least one false rejection. We seek a multiple testing procedure based on $(p_1, \ldots, p_n)$ that has a FWER of not more than α_0, in the same way that in a single hypothesis test we seek a rejection region that has a size not more than α_0. This can always be done, according to the *Bonferroni procedure*, which is to reject H_0^i when $p_i \leq \alpha_0 / n$.

Family-Wise Error Rate (FWER)

Bonferroni procedure

Proof. Let $\mathcal{I}$ be the set of i's for which H_0^i is true, and let k be the size of $\mathcal{I}$. Let A_i be the proposition that H_0^i is rejected. According to

[17] There is a technical complication in the construction of confidence sets for hypothesis tests, due to whether H_1 is one-tailed, e.g. $\theta > \theta_0$, or two-tailed, e.g. $\theta \neq \theta_0$.

the Bonferroni inequality, (eq. 2), and the sub-uniform distribution of P-values under the null hypothesis,

$$\begin{aligned}\text{FWER} &= \Pr\big(\bigvee_{i \in \mathcal{I}} A_i\big) \\ &\leq \sum_{i \in \mathcal{I}} \Pr(A_i) = \sum_{i \in \mathcal{I}} \Pr(P_i \leq \alpha_0 / n) \\ &\leq \sum_{i \in \mathcal{I}} \alpha_0 / n = k\alpha_0 / n \leq \alpha_0.\end{aligned}$$

□

In fact, the FWER in the Bonferroni procedure is often a lot lower than α_0. There is a better procedure, the *Holm procedure*, but for really large n controlling the FWER remains extremely conservative, and often no H_0^i are rejected for a standard level such as $\alpha_0 = 0.05$. So interest has switched to controlling a less stringent feature of the test, the *False Discovery Rate (FDR)*.

Holm procedure

False Discovery Rate (FDR)

The FDR focuses on the proportion of the rejected null hypotheses that are falsely rejected. This is a random quantity, and the FDR is defined as its expectation. Then

$$\text{FDR} \leq \text{FWER},$$

so that a procedure which controls the FWER, such as the Bonferroni procedure, automatically controls the FDR.

Proof. Let V be the number of false rejections, and S be the number of true rejections, and define $Q = V/(V+S)$ if $V + S > 0$, and zero otherwise. Then FDR $:= E(Q)$. But since $Q \leq \mathbb{1}(V \geq 1)$, the monotonicity property of expectations gives

$$\text{FDR} = E\{Q\} \leq E\{\mathbb{1}(V \geq 1)\} = \Pr(V \geq 1) = \text{FWER}.$$

□

This proof suggests that the FDR is *much* less stringent than the FWER, and, indeed, a procedure with FDR $\leq 5\%$ may well have an FWER of nearly 100% for large enough n.[18] Control of the FDR was proposed by Benjamini and Hochberg (1995), who also provided a procedure for achieving a specified upper bound on the FDR. Their original proof was obscure, and a better proof is given in Ferreira and Zwinderman (2006).

Benjamini-Hochberg (BH) procedure

Controlling the FDR rather than the FWER seems appropriate if the cost of a few false rejections is not too high. This is typically the case in screening, where the intention is to narrow the search for rejected H_0's to a subset of the X_i's, which subset will then be subjected to a more expensive investigation. The BH procedure is very simple to implement, and it has had a major impact on screening experiments in genomics, where n can be 10^4; Efron (2008) provides an illustration, and a Bayesian interpretation of the FDR.

Another very useful application of screening is in error and outlier detection in large datasets subject to contamination, such as satellite measurements.

[18] I.e., it will almost certainly wrongly reject an H_0^i.

5 *What do statisticians mean by 'probability'?*

There is a well-defined *mathematical* theory of probability. This is the theory of finite, non-negative, countably additive set functions defined on a sigma field. One takes the existence of such a function and such a field as given, and then works out the implications of additional properties such as probabilistic independence and symmetry; see for example Kingman and Taylor (1966).

The statistician, however, needs to attach an operational meaning to probability, so that he can agree that whenever A and B are incompatible propositions, then $\Pr(A \vee B) = \Pr(A) + \Pr(B)$ is reasonable. In actual applications, he needs an operational meaning to decide whether statements such as $\Pr(X = 1) = 0.3$ or $\Pr(X_2 = 1 \mid X_1 = 1) = 0.1$ are reasonable. Frequentist statisticians and Bayesian statisticians have proposed alternative operationalisations of 'Pr', which will be briefly outlined in this section. Hacking (2001) provides a good survey of the issues.

Just to clear the air, let us note that anyone who asserts that personal judgements have no place in science has clearly not done any science; Ziman (2000) might be a good place to start.[19] Both Frequentist and Bayesian approaches to inference rely heavily on personal judgements, and claims that one approach is more or less subjective than another are largely specious. As scientists, and particularly as statisticians working in collaboration with applied scientists, we should aim for an analysis that is (i) transparent and (ii) defensible. This means we must be clear about where our judgements reside, and indicate when different plausible judgements might give rise to different conclusions.

5.1 *Frequentist probability*

Frequentist statisticians have attempted to operationalise 'Pr' in an objective way; i.e., a way that makes no reference to personal knowledge or capacities. Unfortunately, the best formal treatment, due to Richard von Mises, is clearly at odds with current practice.

This outline is taken from the concise summary of von Mises's view given in Schield and Burnham (2008). Von Mises proposed that probability concerned statements about *collectives*. A collective is an infinite sequence $X_1, X_2, \dots$ with two properties:

collectives

1. For each $a_j \in \mathcal{X}$, n_j/n tends to a well-defined limit as $n \to \infty$, where n_j is the number of X's that equal a_j in $X_1, \dots, X_n$; and

2. For each i, no gambling system to predict X_i based on the first $i - 1$ outcomes can be successful ('randomness').

Von Mises defined probability as limiting relative frequency in collectives, and was then able to conclude that the usual properties of probability follow. He acknowledged that these conditions are idealisations, but considered that they may be approximately satisfied in certain conditions, drawing analogies with common

[19] This rather dense book only contains one equation, but it happens to be Bayes's theorem. Ziman provides a good discussion of the 'Mertonian norms' of science, which include disinterestedness, which woolly-minded people sometimes confuse with 'non-subjectivity'.

practice in mathematical physics. He stressed that probability statements cannot be attached to individual X_i's, but only to collectives; in other words, he would have resisted a statement of the form $\Pr(X_i = 1) = 0.3$ in favour of "if one was to select an X at random from the collective, then the probability that it had the label '1' would be 0.3".

However, as one surveys the huge range of fields in which Frequentist statistics is being performed, it is very difficult to identify the collective. In particular, what if this experiment is stubbornly finite? What condition is sufficient for limiting relative frequencies to exist for all elements of $\mathcal{X}$? And how does one ensure that no gambling system exists, except by deliberate randomisation?

Does one really have to scramble the observations in order to do statistical inference? This seems horribly wasteful.

In practice, Frequentist statisticians have adopted a pragmatic approach. In many situations, the same operations, if repeated, would deliver a different outcome, due to uncontrolled sources of variation. The statistical model for $\boldsymbol{X}$ describes the scientist's judgement about the possible outcomes, and inferences are assessed on the basis of their performance across these outcomes, because one does not know what particular realisation of the uncontrolled variation will be involved in the actual $\boldsymbol{x}^{\text{obs}}$.

Hypothetically, if the scientist's judgement about $\boldsymbol{X}$ is always correct, then by using procedures such as level 5% hypothesis tests, the scientist can be sure that he would not wrongly reject more than 5% of the true null hypotheses, over the course of his career, were he to do thousands of hypothesis tests. It is important to appreciate that it is his life-time Type 1 error rate that is being controlled: there is no requirement for him to repeat his given experiment.

One important legacy of the Frequentist idealisation of probability as relative frequency is the view that probability statements are not well-defined for arbitrary propositions. The Frequentist statistician starts with a statistical model of the form

$$\exists \theta \in \Omega \text{ such that } \boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}; \theta),$$

where $f_{\boldsymbol{X}}$ and Ω are specified, and θ is unknown, and to be inferred from the values of a set of observations, $\boldsymbol{x}^{\text{obs}}$. If you look back through the previous sections, you will see that at no point was a probability distribution attached to θ. More than anything else, this is the shibboleth of Frequentist inference: *a strong aversion to attaching probabilities to θ*. Consequently, a 95% confidence interval for θ of $(17.2, 21.3)$ does not assert that there is a 95% probability that θ lies between 17.2 and 21.3, although it is often misinterpreted as such. Instead, $(17.2, 21.3)$ is one realisation of a random interval which has the property that it will contain θ at least 95% of the time, no matter what the value of θ happens to be: see section 2.3.

Likewise, a P-value of 0.05 does not indicate that there is a 5% probability of H_0 being correct—another common misinterpretation.

5.2 *Bayesian probability*

In the personalistic Bayesian interpretation, probability represents a person's degree of belief in a proposition, on a scale ranging from zero (certainly false) to one (certainly true). There are a number of different implementations: none claims to be the definition of probability, but each of them is available to help with the assessment of probabilities.

One simple implementation represents probabilities in terms of fair prices for gambles. For example, if £0.3 is the most I would pay for a gamble with payoff £$\mathbb{1}(X = 1)$, then I reveal that, for me, $\Pr(X = 1) \approx 0.3$. Likewise, if £0.7 is the most I would pay for a gamble with payoff £$\mathbb{1}(X = 1)$ if $Y = 1$, and which is called off if $Y \neq 1$ (in which case I get my stake back), then I reveal that, for me, $\Pr(X = 1 \mid Y = 1) \approx 0.7$. It then follows that my probability assessments are *coherent*[20] if and only if they respect the axioms of the probability calculus, and the general result that $\Pr(A, B) = \Pr(A \mid B) \Pr(B)$ follows as a theorem. This approach is personalistic because 'Pr' is *my* probability, which depends on my knowledge and capacities. *Your* probabilities are likely to be different.

This Bayesian interpretation allows probabilistic statements to be attached to any proposition that is capable of verification. Such propositions must be operationally defined. For example, a Frequentist statistician would baulk at the statement, "the probability that the 100th digit in the decimal expansion of π is a zero is 0.1". But a Bayesian would be happy to assert this: one can bet on things that one does not know; such things do not have to be 'random'. So the Bayesian interpretation of probability is much more extensive than the Frequentist interpretation. But does this interpretation extend all the way to probabilistic statements about θ? Not obviously, because, being simply an index in a family of statistical models, θ may *not* be operationally defined, except in very simple cases.

Now this may just be a matter of taste. Perhaps my reluctance to bet on things which are not operationally defined is mere squeamishness. Such reluctance could be overcome were it to be shown that many benefits would accrue after adopting the Bayesian interpretation. And, indeed, that is the message of the sections that follow this one. There are many benefits to a Bayesian approach, which all follow from specifying a probability distribution for θ. These benefits are so compelling that they can paint a very skewed picture of clear-sighted Bayesians versus confused Frequentists. So I want to set the record straight at the very start: specifying a defensible probability distribution for θ is hard for Bayesian statisticians, especially in the personalistic approach where probabilities are operationalised in terms of gambles. As a consequence of this difficulty, there is a whole spectrum of Bayesian techniques, such as nonparametric Bayes, empirical Bayes, and Bayes linear, not to mention various types of 'objective' Bayes, that operate with a reduced set of probability judgements about θ.

One should not overlook that a PDF contains an uncountable number of judgements about θ: far more than could be critiqued against any finite number of observations. In special cases the effect of these judgements can gradually be obliterated by increasing sample sizes (section 7), but in other cases they persist (section 9.4).

The personalistic approach has been developed by F.P. Ramsey, B. de Finetti, L.J. Savage, and D.V. Lindley, among others. Lindley (1985) and Jeffrey (2004) provide introductions, while Savage (1954) and de Finetti (1972, 1974/75) contain more advanced treat-

[20] Probability assessments are coherent if it is impossible to create a set of bets for which I am guaranteed to lose money. Such a set would be termed a *dutch book*.

ments. Much more detail about Bayesian inference can be found in O'Hagan and Forster (2004) or Robert (2007). For me, the best short summary of the contrast between Bayesian and Frequentist inference is Savage *et al.* (1962).

6 Bayesian inference

There has been a huge growth in Bayesian statistical inference in the last two decades. Smith (2010, section 1.0.3) gives four reasons, which I paraphrase:

1. An increasing acceptance of the value of expert judgement,
2. A shift away from simple IID-type models to large-scale highly-structured models,
3. A change in culture towards decision support, which takes explicit account of stakeholder preferences, and
4. Dramatic computational developments, both in power and algorithms.

One important subject not covered in these notes is elicitation of expert judgement; see Cooke (1991) and Garthwaite *et al.* (2005). The second point typically involves a substantial increase in the number of nuisance parameters, an area where the Frequentist statistical approach struggles. The third point involves Decision Theory (section 8), and the final point, the development of practical Monte Carlo methods, is very briefly summarised in section 6.6.

6.1 Bayesian basics

The Frequentist approach starts with a marginal statistical model for $\boldsymbol{X}$ indexed by an unknown θ. The Bayesian approach starts with a joint probability model for $(\boldsymbol{X}, \theta)$. The statistical model $f_{\boldsymbol{X}}(\boldsymbol{x}; t)$ is interpreted as the conditional probability that $\boldsymbol{X} = \boldsymbol{x}$ given $\theta = t$. Then θ itself is assigned a probability density function (PDF)—it is a PDF because Ω is typically an uncountable set (for us, it is a convex subset of $\mathbb{R}^d$). This is termed the *prior probability distribution* for θ, and denoted (by me) as π_θ. Under this interpretation of the statistical model, the joint distribution follows immediately from the definition of a conditional probability:

prior probability distribution

$$f_{\boldsymbol{X},\theta}(\boldsymbol{x}, t) = f_{\boldsymbol{X}}(\boldsymbol{x}; t)\,\pi_\theta(t).$$

The Bayesian statistical model represents a simple hypothesis about $\boldsymbol{X}$, because

$$\Pr(\boldsymbol{X} = \boldsymbol{x}) = \int f_{\boldsymbol{X},\theta}(\boldsymbol{x}, t)\,\mathrm{d}t =: f_{\boldsymbol{X}}(\boldsymbol{x}). \tag{12}$$

Here $f_{\boldsymbol{X}}$ is termed the *marginal distribution* of $\boldsymbol{X}$. *Bayes's theorem* asserts that

marginal distribution

Bayes's theorem

$$\pi^*_\theta(t) := \frac{L(t)\,\pi_\theta(t)}{f_{\boldsymbol{X}}(\boldsymbol{x}^{\mathrm{obs}})}$$

is the PDF for θ conditional on $\boldsymbol{X} = \boldsymbol{x}^{\mathrm{obs}}$.

In English, "Bayes's Theorem" is correct, rather than "Bayes' Theorem", on the authority of *Fowler's Modern English Usage*, (3rd ed., 1996, Oxford University Press), p. 61. I believe that Americans do things differently.

Proof. Let A be a subset of Ω. Then

$$\begin{aligned}
\Pr(\theta \in A \mid \boldsymbol{X} = \boldsymbol{x}^{\text{obs}}) &= \frac{\Pr(\theta \in A, \boldsymbol{X} = \boldsymbol{x}^{\text{obs}})}{\Pr(\boldsymbol{X} = \boldsymbol{x}^{\text{obs}})} \\
&= \frac{\int_A f_{\boldsymbol{X},\theta}(\boldsymbol{x}^{\text{obs}}, t)\,\mathrm{d}t}{f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}})} \\
&= \frac{\int_A f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}}; t)\pi_\theta(t)\,\mathrm{d}t}{f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}})} \\
&= \int_A \frac{L(t)\pi_\theta(t)}{f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}})}\,\mathrm{d}t = \int_A \pi_\theta^*(t)\,\mathrm{d}t,
\end{aligned}$$

as needed to be shown. A slight adjustment to the proof is necessary when $\boldsymbol{X}$ is approximated by a continuous random quantity; see Grimmett and Stirzaker (2001, section 4.6). □

In Bayes's theorem, π_θ^* is termed the *posterior probability distribution* of θ, and because the denominator is just a scaling constant, Bayes's theorem is often written

posterior probability distribution

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

The terms 'prior' and 'posterior' are unfortunate, as they give the impression of time passing, but they have stuck. It would have been better to use the labels 'marginal' and 'conditional', which is what they actually are.

This is the basic principle of Bayesian inference:

> *All inferences about θ are based on the posterior distribution.*

This may seem intuitive, but it is also a theorem of Bayesian decision theory; see section 8.1. It is completely at odds with the Frequentist approach, where inferences are based on the distribution of estimators and test statistics over the range possible parameter values. The Bayesian approach makes no reference to values of $\boldsymbol{X}$ other than the $\boldsymbol{x}^{\text{obs}}$ which actually occurred.

6.2 *Transformation-invariance*

Suppose that $g : \theta \mapsto \psi$ is a differentiable bijective function. Then the standard result on the PDF of transformations is that if $\theta \sim \pi_\theta(t)$ then

$$\pi_\psi(v) = \pi_\theta(t)\,|J_g(t)|^{-1} \quad \text{where } t = g^{-1}(v),$$

see section 1.4. The Bayesian posterior distribution obeys this relationship if the prior distribution does, and is thus transformation-invariant.

Proof.

$$\begin{aligned}
L^\psi(v)\pi_\psi(v) &= f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}}; g^{-1}(v))\pi_\psi(v) \\
&= f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}}; t)\,\pi_\theta(t)\,|J_g(t)|^{-1} \quad \text{where } t = g^{-1}(v) \\
&= L(t)\,\pi_\theta(t)\,|J_g(t)|^{-1}.
\end{aligned}$$

These both integrate to the same value, $f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}})$; the left-hand side over $v \in g(\Omega)$, the right-hand side over $t \in \Omega$. Thus

$$\pi_\psi^*(v) = \frac{L(t)\pi_\theta(t)\,|J_g(t)|^{-1}}{f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}})} = \pi_\theta^*(t)\,|J_g(t)|^{-1}$$

as required. □

6.3 Estimation

Point estimates for θ can be derived from the posterior distribution. The posterior mode is a computationally attractive option, because it does not depend on evaluating the denominator, $f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}})$, which typically requires an integration over Ω. So finding the posterior mode is no harder than finding the ML estimate. Indeed, it is the same operation if Ω is bounded and $\pi_\theta(t) \propto 1$.

The posterior mean is often preferred in practice, particularly as modern simulation methods have made it possible to sample from the posterior distribution without explicitly evaluating the denominator; see section 6.6. There is a detailed theory about which estimate of θ to prefer, but the posterior mean does a good job of approximating the minimum-loss estimate for convex loss functions; see section 8. The posterior mean is not transformation-invariant; alternatives are discussed in section 8.

Bayesian set estimates for θ are termed (posterior) *credible sets*. A *level* $(1-\alpha)$ *credible set* for θ is any set $\mathcal{C} \subset \Omega$ for which

level $(1-\alpha)$ credible set

$$\Pr\{\theta \in \mathcal{C} \mid \boldsymbol{X} = \boldsymbol{x}^{\text{obs}}\} = 1 - \alpha. \tag{13}$$

In general, credible sets can be derived from the level sets of the posterior distribution, known as *high probability density (HPD) sets*. A level $(1-\alpha)$ HPD credible set is represented as

high probability density (HPD) sets

$$\mathcal{C} := \big\{t \in \Omega : \pi_\theta^*(t) > c\big\},$$

where c is chosen to satisfy (13). HPD sets are optimally small in the sense that if a set minimses volume in Ω for a given probability content, then it must be an HPD set.

Proof. It suffices to consider $\theta \sim \pi(t)$, where π is a PDF. The proof is by contradiction. Consider a set $S \subset \Omega$ with probability content $1-\alpha$, and let t and t' be any two points on the boundary of the set. if $\pi(t) > \pi(t')$ then a volume element $\mathrm{d}t$ can be moved from t' to t, in such a way that the volume of the set is preserved, but the probability content of the set is increased by $\{\pi(t) - \pi(t')\}\mathrm{d}t$. Then the volume of the modified set can be shrunk slightly to get back to the same probability content as before. Thus, at minimum volume, $\pi(t)$ is the same for all points on the boundary of S. □

Posterior HPD sets are not transformation-invariant. A posterior credible set which *is* transformation-invariant can be found using

$$\mathcal{C} := \big\{t \in \Omega : L(t) > c\big\},$$

where c is chosen to satisfy (13), as before. These sets will typically be larger than HPD sets, but not by much if the likelihood function is strongly concentrated relative to the prior distribution (in which case they will hardly differ from posterior HPD sets).

Frequentist confidence sets can struggle with nuisance parameters. These present no difficulty in the Bayesian approach, because nuisance parameters can simply be integrated out of the posterior distribution. Thus, if $\theta = (\theta', \theta'')$ but only θ' is interesting, then

Nuisance parameters in the Bayesian approach

$$\pi^*_{\theta'}(t') = \int_{\Omega''} \pi^*_\theta(t', t'')\, \mathrm{d}t''.$$

Then a credible set for θ' can be constructed from this marginal posterior distribution. For a scalar θ or θ', an alternative to an HPD credible set is the credible interval which comprises the $\alpha/2$ and $1 - \alpha/2$ quantiles of the marginal posterior distribution. This is termed a $1 - \alpha$ *equi-tailed interval* (or centered interval); it *is* transformation-invariant.

equi-tailed interval

6.4 Prediction

If X_{n+1} is an unobserved random quantity, then

$$\begin{aligned}\Pr(X_{n+1} = x \mid \boldsymbol{X} = \boldsymbol{x}^{\text{obs}}) \\ &= \int f_{X_{n+1}|\boldsymbol{X}}(x \mid \boldsymbol{x}^{\text{obs}}; t)\, \pi^*_\theta(t)\, \mathrm{d}t \\ &\overset{?}{=} \int f_X(x; t)\, \pi^*_\theta(t)\, \mathrm{d}t\end{aligned}$$

where the first equality is just the Law of Total Probability, and is the general result, and the second equality follows in the special case that $(X_1, \dots, X_{n+1}) \overset{\text{iid}}{\sim} f_X(x; \theta)$. So the Bayesian prediction is different from the 'plug-in' ML prediction (section 2.1), because it includes uncertainty about θ.

6.5 Hypothesis tests

If both H_0 and H_1 are composite hypotheses, that occupy well-defined volumes in Ω, then deciding between them is simply a matter of finding out which one has the larger posterior probability content. For example, if θ is a scalar and

$$H_0 : \theta \le \theta_0 \qquad H_1 : \theta > \theta_0$$

then the posterior probability of H_0 is simply

$$\Pr(H_0 \mid \boldsymbol{X} = \boldsymbol{x}^{\text{obs}}) = \int_{-\infty}^{\theta_0} \pi^*_\theta(t)\, \mathrm{d}t.$$

The more interesting case is where one or both hypothesis are degenerate, for example in the case $H_0 : \theta = \theta_0$. In this we must have $\Pr(\theta = \theta_0 \mid \boldsymbol{X} = \boldsymbol{x}^{\text{obs}}) = 0$, because θ is a continuous quantity—point hypotheses cannot literally be true. One way around this is to embody the notion of 'adequacy' which underlies the Frequentist approach with a blurred point hypothesis of the form $H_0 : \theta \in \theta_0 \pm \varepsilon$ for some small ε. But this is an unsatisfactory fix. A better approach is to use *Bayes factors* to evaluate competing hypotheses about $\boldsymbol{X}$; see section 9.

6.6 Bayesian computation

As for the bootstrap, these notes are *not* the place for a detailed discussion of Bayesian computation; see, e.g., Besag *et al.* (1995), Robert (2007, chapter 6), Gelman *et al.* (2003), or Robert and Casella (1999). So just one or two brief remarks.

For practical inference, a sample from the posterior distribution π_θ^* is usually adequate. This is because the marginal posterior distribution of θ' can be appoximated by the histogram of the sampled values of θ' (i.e. one simply ignores the nuisance parameters θ''). And, if required, the conditional posterior distribution of $\theta \mid \{g(\theta) = a\}$ can be approximated by the histogram of those samples satisfying the condition. Sampling is also suitable for intractable problems where the statistical model can be simulated, but the likelihood function cannot be evaluated pointwise; see Beaumont *et al.* (2002).

The challenge for Bayesian inference, then, is to sample from the posterior distribution. Typically this means sampling from the distribution which is proportional to $L(t)\pi_\theta(t)$, because the denominator $f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}})$ is not available explicitly, being an intractable high-dimensional integral, see (12). There are simple methods to do this when the parameter has a small dimension, such as *rejection sampling* and, to compute posterior expectations, *importance sampling*. But when the parameter has a large dimension, as many modern inferences do, then the method of choice is *Markov chain Monte Carlo (MCMC)*.

rejection sampling

importance sampling

Markov chain Monte Carlo (MCMC)

In MCMC, one constructs a Markov chain on Ω which has the property of being stationary, and for which the stationary distribution is the target distribution π_θ^*. It turns out that there are literally an infinite number of such chains, and the skill is in choosing a good one which mixes rapidly, so that the autocorrelation in the chain is low for all points in Ω, or at least for those points in Ω with non-negligible posterior probability. Much of the recent interest in MCMC has been in automating the process of choosing the chain, and in particular, of allowing it to be 'self-tuning', without at the same time doing too much violence to its Markov property; see Andrieu and Thoms (2008).

MCMC is 'asymptotically exact', in the sense that an infinite number of iterations would be required to be sure of drawing one realisation from the posterior distribution. In practice, though, diagnostic evaluation of the chain, or of parallel chains, is used to check the much weaker condition that there is no evidence of non-convergence. A much more elegant technique, *perfect simulation*, can be used to sample directly, although this has not yet caught on

perfect simulation

due in part to computational difficulties in an uncountable Ω; see MacKay (2003, chapter 32) for an outline.

6.7 More on the prior distribution

Improper prior distributions. Suppose initially that Ω is a closed and bounded subset of $\mathbb{R}^d$. Then if π_θ is a non-negative continuous function which is positive somewhere on Ω then the integral of π_θ over Ω is strictly positive and finite. So π_θ can function as a PDF, suitably normalised so that it integrates to one. In fact, the normalisation constant is not required, because it cancels in the numerator and denominator of the posterior distribution.

So, sticking with π_θ being non-negative and continuous, problems can only arise in the case where Ω is unbounded. In particular, it is possible to propose an *improper prior distribution*, for which the integral of π_θ over Ω is infinite. Most Bayesian statisticians seem to be relaxed about this, provided that the posterior distribution is proper, which it often is, especially in the case where $\boldsymbol{X} \overset{\text{iid}}{\sim} f_X(x;\theta)$ and the sample size n is large relative to the dimension of Ω.

improper prior distribution

I would be more cautious, because of paradoxes that can arise with improper distributions (e.g. the *marginalisation paradox*). If pressed to use an improper prior distribution on an unbounded Ω, I would want to derive the posterior distribution as the limit of the posterior distributions of an increasing sequence of bounded subsets of Ω. Improper prior distributions are discussed in more detail in section 9.4.

marginalisation paradox

Vague prior distributions. If there was a shape for the prior distribution corresponding to complete ignorance about θ, then this shape would *not* be preserved under the bijective transformation from θ to ψ. It would be bizarre indeed if I could gain knowledge simply by transforming the parameter, and so a generic prior distribution which represents complete ignorance must be a chimera. The uniform distribution, for example, is *not* representative of ignorance.

But the search goes on for principles upon which one might construct priors that 'let the observations do the talking'. Unfortunately, most suggested approaches give rise to improper priors for an unbounded parameter space. Two popular approaches are the *Jeffreys prior* for scalar θ, and the *reference prior*, which chooses a prior (not uniquely in the case of a vector θ) that maximises the Kullback-Leibler divergence between the prior and posterior distributions.

Jeffreys prior

reference prior

For IID exponential family statistical models there are also *conjugate priors*, which can be made vague in the sense of representing prior information in terms of a specified number of previous observations; e.g. letting the prior information be worth one observation.

conjugate priors

Vague prior distributions are discussed in more detail in Robert (2007, section 3.5).

7 Asymptotics

Suppose that the statistical model asserts that $\boldsymbol{X} \overset{\text{iid}}{\sim} f_X(x;\theta)$. What happens as n becomes large? One case has already been mentioned: for regular models, $u(\boldsymbol{X},\theta) \xrightarrow{\text{D}} \text{N}_d(u;\mathbf{0},i(\theta))$ according to the CLT; see section 2.3. This section gives some more general results. Much more detail about asymptotic behaviour can be found in van der Vaart (1998).

7.1 Consistency

The first important result is that the MLE $\hat{\theta}(\boldsymbol{X})$ concentrates onto (typically) a single value in Ω and that, as a consequence, the Bayesian posterior distribution for θ concentrates around the MLE, more-or-less regardless of the choice of prior distribution.

The tool for understanding this is the *Kullback-Leibler divergence*, which is defined for two statistical models f and g as

Kullback-Leibler divergence

$$\text{KL}(f,g) := \sum_{\boldsymbol{x}} f(\boldsymbol{x}) \log \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}.$$

A crucial result (easily proved using Jensen's inequality) is that $\text{KL}(f,g) \geq 0$ with $\text{KL}(f,g) = 0$ if and only if $f = g$; this is known as *Gibbs's Inequality*.

Gibbs's Inequality

Now suppose that the true distribution of $\boldsymbol{X}$ is $\boldsymbol{X} \overset{\text{iid}}{\sim} g(x)$. We treat the sample space $\mathcal{X}$ as known, and in this case we may assume that the support of f_X contains the support of g (anything else would be bizarre). For any $t_1, t_2 \in \Omega$ the log probability ratio can be rewritten as

$$\begin{aligned} \log \frac{f_X(\boldsymbol{x};t_1)}{f_X(\boldsymbol{x};t_2)} &= \log \frac{f_X(\boldsymbol{x};t_1)}{g(\boldsymbol{x})} - \log \frac{f_X(\boldsymbol{x};t_2)}{g(\boldsymbol{x})} \\ &= \log \frac{g(\boldsymbol{x})}{f_X(\boldsymbol{x};t_2)} - \log \frac{g(\boldsymbol{x})}{f_X(\boldsymbol{x};t_1)} . \end{aligned} \tag{14}$$

But

$$\log \frac{g(\boldsymbol{x})}{f_X(\boldsymbol{x};t)} = \log \frac{\prod_i g(x_i)}{\prod_i f_X(x_i;t)} = \sum_{i=1}^n \log \frac{g(x_i)}{f_X(x_i;t)},$$

and hence $\log\{g(\boldsymbol{X})/f_X(\boldsymbol{X};t)\}$ is the sum of IID bounded random quantities, and its mean must be finite. Now we can apply the WLLN to deduce that

$$\log \frac{g(\boldsymbol{X})}{f_X(\boldsymbol{X};t)} \xrightarrow{\text{P}} n\,E\left\{\log \frac{g(X)}{f_X(X;t)}; g\right\} = n\,\text{KL}(g,t),$$

which must be non-negative and finite, writing $\text{KL}(g, f_X(\cdot\,;t))$ as $\text{KL}(g,t)$, for convenience.

Applying the WLLN to both terms in (14), we deduce

$$\log \frac{f_X(\boldsymbol{X};t_1)}{f_X(\boldsymbol{X};t_2)} \xrightarrow{\text{P}} n\{\,\text{KL}(g,t_2) - \text{KL}(g,t_1)\}.$$

So if $\text{KL}(g,t_2) > \text{KL}(g,t_1)$ then $f_X(\boldsymbol{X};t_1)/f_X(\boldsymbol{X};t_2)$ will increase roughly exponentially for large enough n. Thus the MLE $\hat{\theta}(\boldsymbol{X})$ will concentrate on that value of $t \in \Omega$ for which $\mathrm{KL}(g, t)$ is smallest. In the special case that $g(x) = f_X(x; \theta_0)$, then $\mathrm{KL}(g, \theta_0) = 0$, the smallest possible value, and the ML estimator will converge on the true value θ_0, i.e. it will be a *consistent estimator*. Of course this special case seldom holds in practice, but it is still the case that the ML estimator converges to a preferred value in Ω.

consistent estimator

It follows immediately that the Bayesian posterior distribution also converges to the same preferred value in Ω as the MLE. Let t_1 and t_2 be two values in Ω for which the prior probability is non-zero. In terms of log posterior odds,

$$\log \frac{\pi_\theta^*(t_1; \boldsymbol{x})}{\pi_\theta^*(t_2; \boldsymbol{x})} = \log \frac{f_{\boldsymbol{X}}(\boldsymbol{x}; t_1)}{f_{\boldsymbol{X}}(\boldsymbol{x}; t_2)} + \log \frac{\pi_\theta(t_1)}{\pi_\theta(t_2)}.$$

And therefore, under $\boldsymbol{X} \overset{\text{iid}}{\sim} g(x)$,

$$\log \frac{\pi_\theta^*(t_1; \boldsymbol{X})}{\pi_\theta^*(t_2; \boldsymbol{X})} \xrightarrow{\text{P}} n\{\, \mathrm{KL}(g, t_2) - \mathrm{KL}(g, t_1)\} + \log \frac{\pi_\theta(t_1)}{\pi_\theta(t_2)}.$$

Hence the value $t \in \Omega$ for which $\pi_\theta(t) > 0$ and $\mathrm{KL}(g, t)$ is smallest will end up collecting all of the posterior probability, as n increases without limit. As long as the support of π_θ contains Ω, the posterior distribution for θ will converge on the MLE of θ. Because the prior distribution plays no part in the limiting posterior distribution, we can conclude that two Bayesian statisticians would agree now that, if shown the same long IID sequence of X values, they would eventually make the same inference about θ, regardless of their possibly different prior distributions.

These results imply that ML and Bayesian estimates of θ will be similar for IID models and sufficiently large n. ML and Bayesian *predictions* will be similar as well, because the convergence of the posterior distribution to the MLE implies that

$$\int_\Omega f_X(x; t)\, \pi_\theta^*(t; \boldsymbol{X})\, \mathrm{d}t \xrightarrow{\text{P}} f_X(x; \hat\theta(\boldsymbol{X})).$$

7.2 *Convergence of the posterior distribution*

Under slightly more restrictive conditions, the Bayesian posterior distribution converges to a multivariate Normal distribution. The crucial quantity is the Bayesian analogue of the Fisher Information,

$$J := -\nabla^2 \log f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}}; \hat\theta(\boldsymbol{x}^{\text{obs}})), \tag{15}$$

where ∇^2 denotes the *Hessian matrix* of second derivatives; J is termed the *observed Fisher Information*. If Ω is an open subset of $\mathbb{R}^d$, then J will be positive definite symmetric, because $\hat\theta(\boldsymbol{x}^{\text{obs}})$ maximises the log-likelihood.[21]

Hessian matrix

observed Fisher Information

Writing $\hat\theta$ for $\hat\theta(\boldsymbol{x}^{\text{obs}})$, the asymptotic behaviour of the posterior distribution is

$$\log \pi_\theta^*(t) = \log \phi_d(t; \hat\theta, J^{-1}) + \text{error that decreases in } n$$

[21] In other words, J will be a $d \times d$ variance matrix; additional technical conditions are required to ensure that the maximum is unique.

where ϕ_d is the PDF of a d-dimensional Normal distribution; i.e. the posterior distribution converges to a Normal distribution with mean $\hat{\theta}$ and with variance J^{-1}. Providing that n is reasonably large, the posterior distribution can be approximated by finding the value of the MLE and computing J, the curvature of the log-likelihood function at its maximum. This makes it easy to derive approximate HPD credible sets for θ, or for the interesting subset θ'.

Proof. Using a Taylor series, expand the log-posterior distribution around $\hat{\theta}$, to second order:

$$\log \pi^*_\theta(t) = \log \pi^*_\theta(\hat{\theta}) + (t-\hat{\theta})^T \nabla \log \pi^*_\theta(\hat{\theta}) + \frac{1}{2}(t-\hat{\theta})^T \nabla^2 \log \pi^*_\theta(\hat{\theta})(t-\hat{\theta}) + o(\|t-\hat{\theta}\|^2)$$

where the range of interest is $t = \hat{\theta} + o(1)$, according to the asymptotic consistency of the MLE.[22] Now

$$\nabla \log \pi^*_\theta(t) = \nabla \log f_X(x^{\text{obs}}; t) + \nabla \log \pi_\theta(t)$$

and so $\nabla \log \pi^*_\theta(\hat{\theta}) = \nabla \log \pi_\theta(\hat{\theta}) = O(1)$, and

$$\nabla^2 \log \pi^*_\theta(\hat{\theta}) = \nabla^2 \log f_X(x^{\text{obs}}; \hat{\theta}) + \nabla^2 \log \pi_\theta(\hat{\theta}) = -J + O(1),$$

where $-J$ is $O(n)$, because the X's are IID. Hence the dominant remainder term in the posterior distribution is $O(1)$ and

$$\log \pi^*_\theta(t) = c + O(1) - \frac{1}{2}(t-\hat{\theta})^T J (t-\hat{\theta}).$$

As n becomes large the term in J dominates the $O(1)$ term, which can be absorbed into the constant, giving, in the limit, the form of a Normal distribution with mean $\hat{\theta}$ and variance J^{-1}. □

These days, Monte Carlo methods would typically be used to sample from π^*_θ, for which no IID and large-n approximations are required; see section 6.6. But this result, and its generalisations, explains why the marginal posterior distributions of the θ's often appear to be approximately Normally distributed, especially if θ has been transformed so that $\Omega = \mathbb{R}^d$, and π_θ is vague enough to be flat in the neighbourhood of the MLE. It is also useful to have a tractable approximation to the posterior distribution; see section 9.6.

8 *Decision theory*

Statistics is not just an intellectual activity, and decision theory reflects this by considering the bigger picture, where decisions, such as whether to accept the null hypothesis, have consequences. The new ingredients are:

1. A set of possible decisions, $\mathcal{D}$,

2. A consequence for each decision, represented as a loss, $L : \Omega \times \mathcal{D} \to \mathbb{R}$, and

3. A possible rule, represented as $\delta : \mathcal{X} \to \mathcal{D}$.

[22] Here I am using the 'big O, little o' asymptotic notation introduced by Edmund Landau; see Schervish (1995, section 7.1.1).

Then the objective of decision theory is to evaluate different rules for making decisions.

It is standard to use L for both the likelihood function and the loss function; no confusion should arise because the arguments are different in the two cases.

So the main new feature of Decision Theory is the loss function L. This seems about as hard to specify as the prior distribution on θ is for a Bayesian approach, which is why Decision Theory has not become widely accepted as the unifying framework for statistical inference, except by Bayesian statisticians. But all statisticians appreciate the clarity that Decision Theory brings to theory of estimation and hypothesis testing, without necessarily wanting to use Decision Theory in their applied work.

The main tool for Decision Theory is the *risk function*, defined as the expected loss of the rule:

risk function

$$R(\theta, \delta) := E\{L(\theta, \delta(\boldsymbol{X})); \theta\}.$$

To illustrate, consider set estimation. In this case, $\mathcal{D} =$ a set of subsets of Ω, $\delta = \mathcal{C}$ where $\mathcal{C}(\boldsymbol{x})$ is a member of $\mathcal{D}$, and a simple loss would have

$$L(\theta, \mathcal{C}(\boldsymbol{x})) = \mathbb{1}(\theta \notin \mathcal{C}(\boldsymbol{x})),$$

sometimes referred to as a 0-1 loss function.[23] In this case the risk function is

$$R(\theta, \mathcal{C}) = \Pr\{\theta \notin \mathcal{C}(\boldsymbol{X}); \theta\}.$$

Thus a $(1 - \alpha)$ confidence set for θ arises from a 0-1 loss function subject to the condition that the risk never exceeds α. The same would be true for marginal confidence sets, for which $L(\theta, \mathcal{C}'(\boldsymbol{x})) = \mathbb{1}(\theta' \notin \mathcal{C}'(\boldsymbol{x}))$.

For a given rule δ, the risk is a function of θ. Two different rules will give rise to different functions, and then the pressing question is which rule do we prefer? In general this is a hard question to answer, because there is no unambiguous ordering over the rules. The exception is where one rule dominates the other, i.e.

$$R(t, \delta) \leq R(t, \delta') \qquad \text{for all } t \in \Omega,$$

with a strict inequality at at least one value of t. A rule δ' which is dominated in this way is said to be *inadmissible*. Any rule which is not inadmissible is admissible.

So one rule dominates the other if and only if, as functions of θ, they are not identical and, while they may touch, they do not cross.

inadmissible

It is important to be clear that *admissible rules are not necessarily good rules*. For example, with the quadratic loss function $L(\theta, s(\boldsymbol{x})) = |\theta - s(\boldsymbol{x})|^2$, the estimator $s(\boldsymbol{x}) = 4$ is an admissible rule for estimating θ, if $4 \in \Omega$, because it is the best rule at the point $\theta = 4$. But this is hardly a good estimator for θ. Even so, because inadmissible rules are obviously unattractive, when choosing a rule it is sensible to restrict attention to the set of admissible rules.

8.1 Bayesian decision theory

Things are a bit easier in the Bayesian approach, because of the availability of the prior distribution π_θ, taken to be proper. This

[23] Not a good choice: see section 8.1.

distribution can be used to integrate θ out of the risk, to give the *integrated risk*:

integrated risk

$$R(\delta) := \int R(t, \delta)\, \pi_\theta(t)\, \mathrm{d}t.$$

With both L and π_θ specified, the integrated risk is just a scalar function of δ. Consequently an optimal rule can be found by minimising the integrated risk:

$$\delta^* := \operatorname{argmin} R(\delta)$$

which is termed the *Bayes rule*, with $R(\delta^*)$ being the *Bayes risk*. The principle that optimal choices are made through minimising the integrated risk is a theorem from Bayesian decision theory, which is based on basic axioms for choice under uncertainty. The classic account of this is Savage (1954).

Bayes rule

Bayes risk

It looks as though finding the Bayes rule would be a really hard calculation, since we are minimising an integral over a function defined on $\mathcal{X}$. But it turns out that the Bayes rule has a very convenient characterisation: for each $\boldsymbol{x}$, $\delta^*(\boldsymbol{x})$ is the value which minimises the posterior expected loss conditional on $\boldsymbol{X} = \boldsymbol{x}$.

In other words, it looks like a calculus of variations problem. But it turns out to be a simple pointwise optimisation problem.

Proof. Writing out the integrated risk expression,

$$R(\delta) = \int \sum_{\boldsymbol{x}} L(t, \delta(\boldsymbol{x}))\, f_{\boldsymbol{X}}(\boldsymbol{x}; t)\, \pi_\theta(t)\, \mathrm{d}t.$$

Now take the integral inside the sum, and write $f_{\boldsymbol{X}}(\boldsymbol{x}; t)\, \pi_\theta(t) = f_{\theta|\boldsymbol{X}}(t \mid \boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x})$ using Bayes's Theorem. Substituting and rearranging gives

$$R(\delta) = \sum_{\boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x}) \int L(t, \delta(\boldsymbol{x}))\, f_{\theta|\boldsymbol{X}}(t \mid \boldsymbol{x})\, \mathrm{d}t.$$

Thus the integrated risk is minimised over δ by taking

$$\delta^*(\boldsymbol{x}) = \operatorname*{argmin}_d \int L(t, d)\, f_{\theta|\boldsymbol{X}}(t \mid \boldsymbol{x})\, \mathrm{d}t$$

for each $\boldsymbol{x}$, as was to be shown. □

Continuing with the illustration of confidence sets, the integrated risk of the 0-1 loss function is $\Pr\{\theta \notin \mathcal{C}(\boldsymbol{X})\}$, treating both θ and $\boldsymbol{X}$ as uncertain. This can be driven down to zero by setting $\mathcal{C}(\boldsymbol{x}) = \Omega$ for all $\boldsymbol{x}$, showing that 0-1 loss is not a very good choice for interval estimation. Instead, a more general loss function such as $L(\theta, \mathcal{C}(\boldsymbol{x})) = \mathbb{1}(\theta \notin \mathcal{C}(\boldsymbol{x})) + \beta\,\mathrm{volume}(\mathcal{C}(\boldsymbol{x}))$ for some $\beta > 0$ is required, which introduces a penalty that increases in the size of the set.

But, when faced with the difficulty of specifying β, in practice statisticians fall back on simple rules such as equitailed confidence intervals or credible intervals.

8.2 *Conditions for admissibility*

This diversion into Bayes rules is not just for an elegant result. Bayes rules and admissible rules are very closely linked. The following result is not the most general, but it is easy to prove. Suppose that Ω is finite. Let $\mathcal{B}$ be the set of Bayes rules with prior distributions which have support on the whole of Ω. Then (i) all members of $\mathcal{B}$ are admissible. Furthermore, if mixed rules are allowed, then (ii) all admissible rules are in $\mathcal{B}$.[24]

Proof. Each possible rule contributes a row of the risk matrix $R = \{r_{ij}\}$, where $r_{ij} = R(\theta_j, \delta_i)$; denote row i as r_i^T. I will take R to be bounded. Let π denote a probability assignment over the θ_j's.

Proof of (i). Suppose that rule δ_1 is in $\mathcal{B}$, corresponding to r_1^T. In this case $R\pi \geq r_1^T \pi \mathbf{1}$ for some $\pi \gg \mathbf{0}$.[25] In other words, $(R - \mathbf{1} r_1^T)\pi \geq \mathbf{0}$. Because $\pi \gg \mathbf{0}$, this implies that for no i is $r_i - r_1 < 0$, i.e. δ_1 is admissible.

Proof of (ii). If the admissible rule is not a mixed rule, then let it be the first row of R. If it *is* a mixed rule, then add an additional row to the top of R. Either way, now δ_1 is an admissible rule, corresponding to r_1^T.

Let $[R]$ denote the convex hull of the rows of R: this set contains all of the rules which can be constructed by mixing. Then r_1 is admissible only if it is an extreme point in $[R]$. This requires that there is an a such that $a^T r_1 = c$, where $c := \inf_{r \in [R]} a^T r$. This is not sufficient for admissibility, however, because r_1 needs to be on the south-west boundary of $[R]$. In other words $a^T(r_1 + \mathrm{d}r) > c$ for all $\mathrm{d}r > \mathbf{0}$, which in turn implies that $a \gg \mathbf{0}$. Setting $\pi = a / \sum_j a_j$ then shows that $\pi \gg \mathbf{0}$ and $\pi^T r_1 \leq \pi^T r$ for all $r \in [R]$; which implies that r_1 is in $\mathcal{B}$. □

This proof uses simple concepts from *convex analysis*; see Whittle (2000, section 15.2). In particular, the characterisation of extreme points is the *supporting hyperplane theorem*.

To go further, and consider uncountable and possibly unbounded Ω, requires smoothness conditions which are mainly of interest to the theorist. In a nutshell, Bayes rules with support on the whole of Ω are admissible, and admissible rules are either Bayes rules, or the limit of sequences of Bayes rules with proper prior distributions. These results are described in Robert (2007, section 2.4.4 and chapter 8).

The generalisation to mixed rules is necessary because a pure rule can be admissible and yet be dominated by a mixed rule. One can always implement mixed rules: just toss a coin to decide which actual rule to use. But you can imagine the look on your client's face when you tell him that he should compute two different estimates of θ, and then toss a coin 17 times, and select the first if there are fewer than 13 heads, and the second otherwise. He will think "I might as well consult an astrologer, who will probably be much cheaper, especially if the moon is in Sagittarius."

Luckily, it is never necessary to tell the client this, if one is a Bayesian statistician. This is because if a mixed rule is admissible, there is a Bayes rule with the same integrated risk. This result holds in full generality, but here is the proof in the case where Ω is finite.

A Frequentist statistician may be unwilling to make this argument, if for him the practice of ranking rules by their integrated risk is unacceptable, because it involves a prior distribution for θ.

Proof. Let Ω be finite, of size m. Then the supporting hyperplane of an admissible mixed rule contains between 2 and m pure rules. As the integrated risk is the same for all rules on the supporting hyperplane, each of these pure rules has the same integrated risk as

[24] A mixed rule is a rule of the form $p_1\delta_1 + \cdots + p_k\delta_k$ where each $p_i \geq 0$ and $\sum_i p_i = 1$.

[25] For vector inequalities, $\boldsymbol{x} \leq \boldsymbol{y}$ if $x_i \leq y_i$ for all i; $\boldsymbol{x} < \boldsymbol{y}$ if $x_i \leq y_i$ for all i with at least one strict inequality; and $\boldsymbol{x} \ll \boldsymbol{y}$ if $x_i < y_i$ for all i.

the mixed rule. As each of these pure rules is admissible, each one is a Bayes rule. □

8.3 A cautionary tale: the Stein Effect

The general message from section 8.2 is *think very carefully before proposing a rule which is not demonstrably a Bayes rule*. Another way to make a similar point is to note that if you restrict yourself to admissible rules then you are effectively behaving *as though* you have a prior distribution over θ, even though as a card-carrying Frequentist statistician you may assert that such a thing does not exist.

Here is a famous cautionary tale about inadmissible point estimators. Suppose that $(X_1, \dots, X_d) \sim \mathrm{N}_d(\boldsymbol{x}; \theta, I_d)$, where $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$. A natural estimator for θ would be $s(\boldsymbol{x}) = \boldsymbol{x}$. While this estimator is not a Bayes rule for a proper prior distribution, it is the posterior expectation in the limit that the prior variances of all of the θ_i's tends to infinity. So one might hope that it is admissible under quadratic loss.[26]

In a hugely influential paper, Stein (1955) showed that this estimator was *in*admissible for sufficiently large d under quadratic loss, and a few years later James and Stein (1961) produced an estimator that dominated it for $d > 2$. The general form of these better estimators is

$$s(\boldsymbol{x}) = \bar{x}\mathbf{1} + c(\boldsymbol{x} - \bar{x}\mathbf{1}) \tag{16}$$

where $\bar{x}$ is the sample mean and c is a function of $\boldsymbol{x}$, lying between 0 and 1. Such estimators are called *shrinkage estimators* because they shrink $\boldsymbol{x}$ towards the mean $\bar{x}$: they are clearly biased if the θ_i's vary. Efron and Morris (1977) give a non-technical presentation of the Stein effect.[27]

shrinkage estimators

The moral of this story is that reaching for the 'obvious' unbiased estimator is risky. One upshot has been the gradual abandonment of unbiasedness as a primary criterion for point estimators. Another has been for statisticians to favour Bayes rules, but to search for prior distributions for which such rules have good sampling properties, so called *matching priors*.[28] Brown *et al.* (2001) present a case study on the use of various approaches, Frequentist and Bayesian, for constructing confidence intervals for the probability parameter in a Binomial model; one recommended approach is a Bayesian credible interval, based on the Jeffreys prior. In other words, sometimes a Bayesian credible interval makes a good Frequentist confidence interval.

matching priors

8.4 Transformation invariance

Despite their origin, Bayes rules are not necessarily transformation-invariant. The 'culprit' is the loss function. Consider, for example, point estimation of a scalar parameter and a quadratic loss function $L(\theta, s(\boldsymbol{x})) = (\theta - s(\boldsymbol{x}))^2$. The Bayes rule is the posterior expecta-

[26] Under quadratic loss, it is easy to see that the Bayes rule for a point estimator is the posterior expectation.

[27] As Efron and Morris note, (16) is itself inadmissible. A Bayes rule would have the form $s(\boldsymbol{x}) = \mu\mathbf{1} + c(\boldsymbol{x} - \mu\mathbf{1})$, where μ is the common prior mean of the θ_i's, and this *would* be admissible.

[28] These tend to be Jeffreys priors or reference priors in simple cases.

tion of θ. Although the posterior distribution is transformation-invariant, the posterior expectation is not. Interestingly, in this case the posterior median *is* transformation-invariant, and this arises from the loss function $L(\theta, s(\boldsymbol{x})) = |\theta - s(\boldsymbol{x})|$.

Whether this is a practical issue is debatable. If transformation-invariance is a compelling issue then the loss function can be chosen appropriately. For example, for point estimation, a general class of transformation-invariant loss functions can be constructed by representing $L(\theta, s(\boldsymbol{x}))$ in terms of the distance between the distributions $f_{\boldsymbol{X}}(\cdot; \theta)$ and $f_{\boldsymbol{X}}(\cdot; s(\boldsymbol{x}))$. This distance would be exactly the same under the bijective transformation $g : \theta \mapsto \psi$, for which $f^{\psi}_{\boldsymbol{X}}(\cdot; \psi) = f_{\boldsymbol{X}}(\cdot; g^{-1}(\psi))$. One such distance has already been mentioned: the Kullback-Leibler divergence.[29] The availability of transformation-invariant Bayes rules is reassuring, but probably not germane in practical statistical applications: Decision Theory seems to be more valuable for its conceptual clarity than as a practical template for applied statistics.

9 Bayes factors

Bayes factors were originally introduced by Harold Jeffreys in the 1930s (see Jeffreys, 1961, 1st edition 1939). The standard reference is the review by Kass and Raftery (1995), and the introductory sections of Berger and Pericchi (2001) are also very helpful. Bayesian textbooks and reference books such as Robert (2007, chapters 5 and 7) also contain more detailed information than that presented here; some references will be given below.

A word on notation. I prefer to use $f_{\boldsymbol{X}}(\boldsymbol{x}; \theta)$ to denote the statistical model, and $\pi_\theta(t)$ to denote the prior distribution on the parameters of the model. But where there are many hypotheses, it is more convenient to use the subscript to indicate the hypothesis. So where appropriate I will write $f_j(\boldsymbol{x}; \theta_j)$ for the statistical model of the jth hypothesis, $\pi_j(t_j)$ for the prior distribution of the parameters in the jth model, which has support $\Omega_j \subset \mathbb{R}^{d_j}$, and $f_j(\boldsymbol{x})$ for the marginal probability of $\boldsymbol{X}$ in the jth model, where $f_j(\boldsymbol{x}) = \int f_j(\boldsymbol{x}; t_j) \pi_j(t_j) \, \mathrm{d}t_j$.

9.1 Definition

Consider two simple hypotheses about $\boldsymbol{X}$:

$$H_0 : \boldsymbol{X} \sim f_0(\boldsymbol{x}), \quad H_1 : \boldsymbol{X} \sim f_1(\boldsymbol{x}). \tag{17}$$

These two hypotheses do not have to be mutually exclusive: they might just be two hypotheses within a whole collection of possibilities, $H_0, H_1, \dots, H_k$. After observing $\boldsymbol{X} = \boldsymbol{x}^{\text{obs}}$, which of these two hypotheses is more probable? Applying Bayes's Theorem in its odds form,

While O_{ij} is a scalar, conventionally it is treated as a plural noun.

$$\frac{\Pr(H_0 \mid \boldsymbol{x}^{\text{obs}})}{\Pr(H_1 \mid \boldsymbol{x}^{\text{obs}})} = \frac{\Pr(\boldsymbol{X} = \boldsymbol{x}^{\text{obs}} \mid H_0)}{\Pr(\boldsymbol{X} = \boldsymbol{x}^{\text{obs}} \mid H_1)} \times \frac{\Pr(H_0)}{\Pr(H_1)}$$

[29] Technically not a distance because it is not symmetric.

or, equivalently,

$$O^*_{01} = B_{01} \times O_{01}$$

where O_{01} are the *prior odds* of H_0 versus H_1, B_{01} is the *Bayes factor* of H_0 versus H_1, and O^*_{01} are the *posterior odds*. In other words, to compute the posterior odds, take the product of the Bayes factor and the prior odds.

prior odds
Bayes factor
posterior odds

In general odds do not convert into probabilities. The exception is where we have the odds on a set of *mutually exclusive hypotheses*, say $O_{01}, \ldots, O_{0k}$. In this case it is easy to see that[30]

mutually exclusive hypotheses

$$p_0 = \frac{1}{1 + \sum_{j=1}^{k} O_{0j}^{-1}}$$

and then $p_j = p_0 \, O_{0j}^{-1}$ for $j = 1, \ldots, k$. I will write O_{ji} for O_{ij}^{-1}, and likewise B_{ji} for B_{ij}^{-1}. As Berger and Pericchi (2001) note, it is common in scientific reporting to set $O_{01} = \cdots = O_{0k} = 1$, in which case $O^*_{0j} = B_{0j}$ and

$$p^*_j = \begin{cases} 1/(1 + \sum_{j=1}^{k} B_{j0}) & j = 0 \\ B_{j0}/(1 + \sum_{j=1}^{k} B_{j0}) & j = 1, \ldots, k. \end{cases}$$

Because it is possible to infer the Bayes factors from these posterior probabilities, anyone with non-uniform prior probabilities is able to compute his personal posterior probabilities from the values of $p^*_0, p^*_1, \ldots, p^*_k$.

The Bayes factor B_{01} measures how the observations have changed the odds of H_0 versus H_1. Bayes factors are usable at two levels, as shown by the following tableau.

1. $B_{01} > 1 \iff O^*_{01} > O_{01}$ After conditioning on x^{obs}, the probability of H_0 is raised relative to H_1.
2. $B_{01} > O_{10} \iff O^*_{01} > 1$ After conditioning on x^{obs}, H_0 is more probable than H_1.

At the first level, the Bayes factor is simply about the *balance of evidence*. If $B_{01} > 1$ (resp. < 1) then the probability of H_0 is raised (resp. lowered) relative to H_1. This conclusion does not require any assessment of O_{01}. If it is possible to assess O_{01}, then more can be said. For example, $O_{10} < B_{10}$ implies that, after conditioning, H_0 is more probable than H_1. So one does not have to make an explicit judgement about O_{01}: a bound may be sufficient. If one can specify O_{01}, then B_{01} allows one to compute the posterior odds O^*_{01}. And if one can do this for a set of mutually exclusive hypotheses, one can compute the posterior probabilities of the hypotheses. This 'incremental' feature of Bayes factors is very useful in practice, when assessing uncertainties, as discussed by Spiegelhalter and Riesch (2011).

balance of evidence

[30] Multiply the top and bottom of the right-hand side by p_0.

9.2 *Bayes factors in simple hypothesis tests*

Decision theory was outlined in section 8. Here we consider simple hypothesis testing, where the new feature of decision theory is to introduce explicit costs for making an error. Suppose the parameter space and the decision space are both $\{0, 1\}$, where 0 indicates the simple hypothesis H_0, and 1 the simple hypothesis H_1. Let the loss function be

$$L(d_1, d_2) = c_0 \mathbb{1}(d_1 = 1 \wedge d_2 = 0) + c_1 \mathbb{1}(d_1 = 0 \wedge d_2 = 1)$$

i.e. a loss of c_0 for choosing H_0 when H_1 is true, and of c_1 for choosing H_1 when H_0 is true, and zero otherwise.

The Bayes rule $\delta^*(\boldsymbol{x})$ is found by minimising the posterior expected loss conditional on $\boldsymbol{X} = \boldsymbol{x}$; see section 8.1. For choosing H_0 this would be

$$\text{choose } H_0 : E(L \mid \boldsymbol{x}) = c_0 \Pr(H_1 \mid \boldsymbol{x}),$$

while for choosing H_1 this would be

$$\text{choose } H_1 : E(L \mid \boldsymbol{x}) = c_1 \Pr(H_0 \mid \boldsymbol{x}).$$

So the Bayes rule chooses H_1 if and only if

$$c_0 \Pr(H_1 \mid \boldsymbol{x}) \geq c_1 \Pr(H_0 \mid \boldsymbol{x}),$$

or $O^*_{10} \geq c_1 / c_0$; equivalently, as $O^*_{10} = B_{10}\, O_{10}$,

$$B_{10} \geq \frac{c_1}{c_0} O_{01}.$$

So the Bayes factor plays the crucial role in the Bayes rule for choosing between two simple hypotheses.

This rule is identical in structure to the Neyman-Pearson optimal hypothesis test: reject H_0 in favour of H_1 when $B_{10} \geq c$. But whereas the Neyman-Pearson approach suggests fixing the Type 1 error level at some significance level α in order to set the critical value c, Bayesian decision theory suggests that c should be determined by the relative cost of the two errors, and the prior probabilities attached to the two hypotheses. As one might expect, c is increasing in the relative cost of making an error when selecting H_1, and increasing in the relative prior probability of H_0.

Most statisticians would agree that the Bayesian approach is more sensible but, when faced with the need to be explicit about the relative costs and the prior odds, it is often less stressful to fall back on the NP approach, particularly for applications which lack serious consequences.

9.3 *Underneath the simple hypothesis*

The hypotheses H_0 and H_1 in (17) may be simple, but that does not mean that they do not involve prior probabilities. Typically, in a Bayesian analysis,

$$H_i : \boldsymbol{X} \sim f_i(\boldsymbol{x}; \theta_i) \text{ and } \theta_i \sim \pi_i(t_i).$$

So hypothesis i involves a statistical model for $\boldsymbol{X}$, and this statistical model's parameters have a prior probability distribution. Across hypotheses the statistical models may be different, and so the parameter spaces may be different. In general, then,

$$B_{ij} := \frac{f_i(\boldsymbol{x}^{\text{obs}})}{f_j(\boldsymbol{x}^{\text{obs}})} \equiv \frac{\int L_i(t_i)\pi_i(t_i)\,\mathrm{d}t_i}{\int L_j(t_j)\pi_j(t_j)\,\mathrm{d}t_j}.$$

Bayes factors are, of course, invariant to differentiable bijective transformations of the parameters in each model.

This section considers the most common form of hypothesis test. Suppose that the model is the same for both hypotheses, namely $f_{\boldsymbol{X}}$, and the hypotheses are

$$\begin{aligned} H_0 :& \boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0), \\ H_1 :& \boldsymbol{X} \sim f_1(\boldsymbol{x}) = \int f_{\boldsymbol{X}}(\boldsymbol{x};t)\pi_1(t)\,\mathrm{d}t \end{aligned} \tag{18}$$

where Ω_1 is the support of π_1, and $\theta_0 \notin \Omega_1$. A hard-line Frequentist statistician would deny that H_1 was a well-formed hypothesis, since it contains a probability distribution over the parameter space. But this statistician is really stuck unless he is in the highly restrictive situation where the statistical model has a scalar parameter with an MLR, and Ω_1 is a subset of $\mathbb{R}$ which lies entirely to the left or right of θ_0. In this case he can use a UMP test; see section 4.2.

Most statisticians would admit the validity of H_1 as a statistical model for $\boldsymbol{X}$ but point out that it is not very helpful if one cannot choose π_1 in a meaningful way. This concern was originally discussed in Edwards *et al.* (1963), who considered bounds on Bayes factors. The most general bound is

$$\begin{aligned} B_{01} &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}};\theta_0)}{\int f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}},t)\pi_1(t)\,\mathrm{d}t} \\ &\geq \frac{f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}};\theta_0)}{\sup_{\pi_1}\int f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}};t)\pi_1(t)\,\mathrm{d}t} \\ &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}};\theta_0)}{f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}};\hat{\theta}_1(\boldsymbol{x}^{\text{obs}}))} \end{aligned} \tag{19}$$

where $\hat{\theta}_1(\boldsymbol{x}^{\text{obs}})$ is the Maximum Likelihood estimator under H_1,

$$\hat{\theta}_1(\boldsymbol{x}^{\text{obs}}) := \underset{t\in\Omega_1}{\operatorname{argmax}} f_{\boldsymbol{X}}(\boldsymbol{x}^{\text{obs}};t).$$

This supposes that the prior distribution π_1 just happened to be a *Dirac delta function* concentrating its mass at the maximum likelihood value within Ω_1. This is an *extreme lower bound*, because no one in his right mind would choose a Dirac delta function for a prior distribution, and even if he did, the chance of it being exactly at the value of the MLE is effectively zero.

Dirac delta function

Having said, even this extreme lower bound is interesting. Consider perhaps the most common situation of all in hypothesis testing, where $\boldsymbol{X} \overset{\text{iid}}{\sim} \mathrm{N}(x;\mu,\sigma^2)$, which implies that, for large n,

$\bar{X} \sim \mathrm{N}(\bar{x}; \mu, s^2/n)$ to a good approximation, where $\bar{X}$ is the sample mean and s^2 is the value of the unbiased estimator of the variance of X_i. Consider the two hypotheses

$$H_0 : \mu = 0, \quad H_1 : \mu > 0.$$

In medical science these might represent H_0: the treatment is the same as the control, and H_1: the treatment is better than the control. The ML estimator for μ under H_1 is $\bar{x}$ if $\bar{x} > 0$, and zero otherwise. So if $\bar{x}^{\mathrm{obs}} \leq 0$ then $B_{01} \geq 1$. If $\bar{x}^{\mathrm{obs}} > 0$ then

$$B_{01} \geq \frac{\phi(\bar{x}^{\mathrm{obs}}; 0, s^2/n)}{\phi(\bar{x}^{\mathrm{obs}}; \bar{x}^{\mathrm{obs}}, s^2/n)} = \exp\{-(z^{\mathrm{obs}})^2/2\}, \tag{20}$$

where ϕ is the Normal PDF with specified expectation and variance, and

$$z^{\mathrm{obs}} := \frac{\bar{x}^{\mathrm{obs}}}{\sqrt{s^2/n}},$$

often referred to as the *z-score*. A value such as $z^{\mathrm{obs}} = 2$, which would imply a P-value for H_0 of 2.3%, has an extreme lower bound for the Bayes factor of $1/e^2 \approx 0.14$. Even at the lower bound value, this does not seem to be compelling evidence against H_0. So unless one already has a strong reason for thinking that H_0 is not true, $z^{\mathrm{obs}} = 2$ does not persuade one to reject H_0. This conclusion directly contradicts standard statistical practice in many applied fields, where a P-value of less than 5% is often used to reject H_0. This is discussed further in section 10.3.

z-score

The extreme lower bound in (19) can be relaxed towards something a bit more reasonable. Sellke *et al.* (2001) provide a fairly general argument for a reasonable lower bound for the Bayes factor being $B_{01} \geq -ep \log p$, where p is the P-value. This lower bound is based on a smooth prior distribution for θ, rather than the Dirac delta function used in the extreme lower bound. See also section 10.3.

Robert (2007, section 5.3.5, *Least favorable Bayesian answers*) discusses other approaches to specifying lower bounds for the Bayes factor.

9.4 *Why prior distributions matter for Bayes factors*

Let's go back to Bayesian inference about θ for a given hypothesis,

$$\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}; \theta) \text{ and } \theta \sim \pi_\theta(t)$$

where the parameters can take any value in $\mathbb{R}^d$ (without loss of generality: this can be achieved by a transformation, if necessary), and let Ω be the support of π_θ, a initially a *bounded* convex subset of $\mathbb{R}^d$. Applying Bayes theorem with observations $\boldsymbol{x}^{\mathrm{obs}}$ gives the posterior distribution

$$\pi^*_\theta(t) = \frac{L(t)\pi_\theta(t)}{f_{\boldsymbol{X}}(\boldsymbol{x}^{\mathrm{obs}})} \equiv \frac{L(t)\pi_\theta(t)}{\int L(u)\pi_\theta(u)\,\mathrm{d}u} \qquad \text{for } t \in \Omega$$

and zero otherwise. Initially I will take $\pi_\theta(t) \propto 1$ on Ω, which gives

$$\pi^*_\theta(t) = \frac{L(t) v^{-1}}{\int_\Omega L(u)\, v^{-1}\,\mathrm{d}u} = \frac{L(t)}{\int_\Omega L(u)\,\mathrm{d}u} \qquad \text{for } t \in \Omega$$

and zero otherwise, where v is the volume of Ω. Clearly the posterior distribution depends on Ω, even if it does not depend explicitly on v.

Now suppose that Ω is sufficiently large that $L(t)$ is zero on and outside its boundary. In this case, I could consider a uniform prior with a larger support, $\Omega' \supset \Omega$ with volume $v' > v$, and the new posterior distribution would be

$$\pi^{*\prime}_\theta(t) = \frac{L(t)(v')^{-1}}{\int_{\Omega'} L(u)\,(v')^{-1}\,\mathrm{d}u} = \frac{L(t)}{\int_\Omega L(u)\,\mathrm{d}u} \qquad \text{for } t \in \Omega'$$

and zero otherwise; i.e. exactly the same. So why bother to set Ω at all, if I am confident that it will contain the support of L? I can let Ω' approach $\mathbb{R}^d$, and, again,

$$\lim \pi^{*\prime}_\theta(t) = \frac{L(t)}{\int_{\mathbb{R}^d} L(u)\,\mathrm{d}u} \qquad \text{for } t \in \mathbb{R}^d$$

even though, in the limit, the prior distribution $\pi_\theta(t) \propto 1$ on $\mathbb{R}^d$ is improper.

When the support of L is not bounded, more complicated conditions are necessary to ensure that the denominator is finite. In well-behaved problems for which $\boldsymbol{X} \overset{\text{iid}}{\sim} f_X(x;\theta)$, having $n \geq d$ often suffices. Other cases may require a non-uniform prior distribution which goes to zero faster than the tails of the likelihood function. The argument then becomes more complicated because expanding to Ω' may change the shape of the prior distribution inside Ω. In practice, improper priors tend to be used mainly for location parameters and log-scale parameters, for which a uniform prior distribution is justifiably 'vague'.

Therefore, to summarise, when doing Bayesian inference about θ it is sometimes innocuous to take Ω unbounded and $\pi_\theta \propto 1$. This convenient choice allows the statistician to spend time on more difficult issues, including specifying the model f_X, and, these days, trying to get the blasted MCMC sampler to converge.

Unfortunately, though, the argument breaks down for Bayes factors. This is because Bayes factors are the ratio of marginal probabilities, and there is no early cancellation of the v's. Sticking with uniform prior distributions for the parameters in both hypotheses

$$B_{ij} := \frac{f_i(\boldsymbol{x}^{\text{obs}})}{f_j(\boldsymbol{x}^{\text{obs}})} = \frac{\int_{\Omega_i} L_i(t_i)(v_i)^{-1}\,\mathrm{d}t_i}{\int_{\Omega_j} L_j(t_j)(v_j)^{-1}\,\mathrm{d}t_j} = \frac{v_j}{v_i} \times \frac{\int_{\Omega_i} L_i(t_i)\,\mathrm{d}t_i}{\int_{\Omega_j} L_j(t_j)\,\mathrm{d}t_j}. \tag{21}$$

Expansion of Ω_i and Ω_j may well leave the second term unchanged, but it has a direct effect on the first term, which is the reciprocal of the ratio of their volumes. In particular, taking one of Ω_i or Ω_j to its limit will drive B_{ij} to 0 or ∞, regardless of $\boldsymbol{x}^{\text{obs}}$, while taking them both to their limits will definitely not work, because the limit of the ratio v_j/v_i is undefined. So (i) improper prior distributions cannot be used for Bayes factors, and (ii) the volume of the parameter space has a direct impact on the result.

This sensitivity to the prior distribution is the basis of *Lindley's paradox*. Lindley (1957) noted that it is not incompatible for the P-value of H_0 to be smaller than 5% while at the same time the Bayes factor can imply a posterior probability for H_0 of greater than 95%, even for a quite small prior probability. A discrepancy between a P-value and a Bayes factor is not a paradox, because the P-value computes an entirely different quantity that does not involve H_1. It would be strange, though, if a P-value rather than a Bayes factor was used to decide between H_0 and H_1—which of course *is* often done in practice: this is the paradox! A closely related paradox is *Bartlett's paradox*. In discussing Lindley's paper, Bartlett (1957) more-or-less made the explicit point that the Bayes factor goes to infinity as the prior distribution π_1 becomes more and more vague.

Lindley's paradox

This is discussed in more detail in section 10.3.

Bartlett's paradox

And just to add to the confusion, Lindley's paradox is sometimes called Jeffreys's paradox, or the Jeffreys-Lindley paradox.

9.5 Model selection and parsimony

Is the sensitivity of the Bayes factor to the prior a good thing or a bad thing? Well, for a statistician wanting a quiet life, it is definitely a bad thing: for Bayes factors one has to worry about the prior distributions, whereas for inference within a given hypothesis it is often possible to use a vague and sometimes improper prior distribution.

On the other hand, it gives Bayes factors a useful property: they *automatically penalise unnecessary model complexity*. If $L_i = L_j$, and Ω_i is a strict subset of Ω_j but the extra points in Ω_j are unnecessary, then the second term in (21) might be a little less than one, due to *over-fitting*, or it might be greater than one, if most of $\Omega_j \setminus \Omega_i$ is in the wrong place. But the first term will be substantially greater than 1, giving $B_{ij} > 1$ overall. Parsimony is considered to be a good property of a hypothesis, reflecting "a widespread philosophical presumption that simplicity is a theoretical virtue", often referred to as *Occam's Razor* (Baker, 2011), as well as the practical benefits of a simple model. So Bayes factors for choosing between hypotheses are naturally parsimonious.

parsimony

over-fitting

Occam's Razor

Other approaches to model selection are somewhat *ad hoc* in the way they penalise model complexity.[31] One popular approach is the *Deviance Information Criterion (DIC)* of Spiegelhalter *et al.* (2002). This popularity in part reflects the ease with which the DIC can be computed from the output of an MCMC sampler, and seems not to reflect the varied concerns of the discussants of the original paper. Some of these concerns are technical, but an obvious one, mentioned by the authors themselves (their section 9.2), is that the DIC is not invariant to bijective transformations of the parameters.

Deviance Information Criterion (DIC)

[31] An example is given immediately below, in section 9.6.

9.6 Computing Bayes factors

Bayes factors involve integrations over Ω_i and Ω_j, in order to compute the marginal likelihoods of hypotheses H_i and H_j. In some situations an approximate Bayes factor suffices, and the following calculation, based on the *Laplace approximation* to the marginal likelihood, is helpful.

Laplace approximation

Write $\hat{\theta}$ for the ML estimate $\hat{\theta}_i(\boldsymbol{x}^{\text{obs}})$, and expand $f_i(\boldsymbol{x}^{\text{obs}}; t) =: L_i(t)$ around $\hat{\theta}$ to give

$$\log L_i(t) \approx \log L_i(\hat{\theta}) - \frac{1}{2}(t - \hat{\theta})^T J(t - \hat{\theta}),$$

where J is the observed Fisher Information; see section 7.2. Also assume that the prior π_i is relatively flat in the region of concentration of the likelihood function, so that $\log \pi_i(t) \approx \log \pi_i(\hat{\theta})$. Then

$$\begin{aligned} f_i(\boldsymbol{x}^{\text{obs}}) &= \int_{\Omega_i} \exp\left\{ \log L_i(t) + \log \pi_i(t) \right\} \mathrm{d}t \\ &\approx L_i(\hat{\theta})\pi_i(\hat{\theta}) \int_{\Omega_i} \exp\left\{ -\frac{1}{2}(t - \hat{\theta})^T J(t - \hat{\theta}) \right\} \mathrm{d}t \\ &= L_i(\hat{\theta})\pi_i(\hat{\theta}) |2\pi J^{-1}|^{1/2} \\ &= L_i(\hat{\theta})\pi_i(\hat{\theta}) (2\pi)^{d_i/2} |J|^{-1/2}, \end{aligned}$$

where the last equality follows because J is $d_i \times d_i$. Further simplification follows if the log-likelihood is $O(n)$, e.g. if the X's are IID, for then $|J|$ is $O(n^{d_i})$, and for large n

$$\log f_i(\boldsymbol{x}^{\text{obs}}) \approx c_i + \log L_i(\hat{\theta}) - (d_i \log n)/2.$$

for some c_i which is $O(1)$. In this simple approximation, the log marginal likelihood comprises a goodness of fit term, $\log L_i(\hat{\theta})$, penalised by a model complexity term $(d_i \log n)/2$. Neglecting the $O(1)$ term, this is in fact the *Bayes Information Criterion (BIC)*,

Bayes Information Criterion (BIC)

$$\text{BIC}_i := -2 \log L_i(\hat{\theta}) + d_i \log n,$$

and motivates (although it does not justify) the use of the BIC in model selection. If H_i has a smaller BIC than H_j, then under the conditions given above, one might proceed as though the Bayes factor favours H_i over H_j, arguing that

$$B_{ij} \approx \exp\left\{-(\text{BIC}_i - \text{BIC}_j)/2\right\},$$

although cautiously, as the suppressed $O(1)$ term might be important.[32]

As in section 7.2, Bayesian statisticians can eschew these approximations, as they have the technology to sample directly from the posterior distribution π_i^*. For example, Gelfand and Dey (1994) note that, for any PDF h with support Ω_i,

$$f_i(\boldsymbol{x}^{\text{obs}})^{-1} = \int_{\Omega_i} \frac{h(t)}{L_i(t)\pi_i(t)} \pi_i^*(t)\, \mathrm{d}t,$$

[32] In fact, the $O(1)$ term *is* often important, and experience suggests that the BIC tends to over-penalise complex models when n is small.

and they suggest the estimate

$$\widehat{f}_i(x^{\text{obs}}) = \left\{ \frac{1}{r} \sum_{j=1}^{r} \frac{h(t^{(j)})}{L_i(t^{(j)})\pi_i(t^{(j)})} \right\}^{-1} \tag{22}$$

where $t^{(1)}, \dots, t^{(r)}$ are sampled from the posterior distribution π_i^*. Here the PDF h can be chosen to approximate the posterior distribution π_i^*, for example, using the Normal approximation given in section 7.2.[33] Such a choice is helpful for stabilising the terms in the sum, and reducing the variability of the estimator (which means that a smaller r is required for a given accuracy).

A more exciting approach is to arrange to sample jointly across the set of all models, and then infer the Bayes factors for each pair of models (or the posterior odds) according to the residence time of the sampler in each model. This can be done using *reversible jump MCMC (RJ-MCMC)*, see Green (1995). It is quite tricky to get the proposal to transition effectively between models with different parameter spaces, though, so this is a technique for experts only. Most of us are better off running a sampler for each model, and computing the marginal likelihood and the Bayes factors using an estimator such as (22).

reversible jump MCMC (RJ-MCMC)

10 Inferential fallacies

People are not very good when reasoning about uncertainty. This means that we make mistakes, and we fail to spot the mistakes of others (and, sometimes, we can be deliberately misled by others). Such mistakes lead to incorrect decisions in the doctor's surgery and in the courtroom, to mention just two venues, and can have catastrophic consequences. Two non-technical books about reasoning and uncertainty are Gigerenzer (2003) and Senn (2003).

10.1 Deduction and inference

Deduction is about drawing valid conclusions from given premises. Consider the premise that H implies E. For example,

Astronauts are good with heights.

This example illustrates that material implication comes in many forms. Formally, in this case, we have $H(x)$: x is an astronaut, and $E(x)$: x is good with heights. The complete statement is $\forall x \in \mathcal{S}\, \{H(x) \to E(x)\}$, where $\mathcal{S}$ is the set of all people.

From this we can validly deduce that (i) if I am an astronaut then I am good with heights. And (ii) if I am bad with heights then I am not an astronaut. We *cannot* deduce that if I am good with heights then I am an astronaut. This horrible error is known as the *fallacy of the transposed conditional*.

fallacy of the transposed conditional

The same fallacy occurs in statistical inference, where the dichotomous $\{0, 1\}$ outcomes of logic are replaced by a sequence of probabilities. For example, if astronauts are probably good with heights, then I cannot infer that because I am good with heights I am probably an astronaut. In other words, just because $\Pr(E \mid H)$ is close to one, I cannot infer from E that H is probably true.

[33] Again, it would be helpful to transform Ω_i to $\mathbb{R}^{d_i}$.

Likewise, just because $\Pr(E \mid H)$ is close to zero, I cannot infer from E that H is probably false. This is because *improbable events occur all the time*. For example, I've just tossed a coin ten times, and I got

$$T, T, T, H, T, T, H, T, H, T.$$

Let this be the evidence, E. Consider the hypothesis H that the coin is fair. Then $\Pr(E \mid H) = 2^{-10}$, which is less than one in a thousand. But it would be ludicrous for me to infer that the coin was not fair.[34] The exception here is when $\Pr(E \mid H) = 0$, in which case I can deduce that H is false.

Both of these inferential errors have the same underlying structure. There is a hypothesis H and there is evidence E. We know the values of $\Pr(E \mid H)$ and we would like to make a statement about H on the basis of E, namely assign a value to $\Pr(H \mid E)$. But in fact it is not possible to make such a statement with only the information available, and to do so commits a fallacy.

As H and not-H are mutually exclusive, we can infer their probabilities conditional on the evidence from the posterior odds ratio for H versus not-H,

$$O^* = \frac{\Pr(H \mid E)}{\Pr(\text{not-}H \mid E)} = \frac{\Pr(E \mid H)}{\Pr(E \mid \text{not-}H)} \frac{\Pr(H)}{\Pr(\text{not-}H)}.$$

In order to determine O^* we need both the Bayes factor for H versus not-H, and also the prior odds—effectively just the prior probability of H, since $\Pr(\text{not-}H) = 1 - \Pr(H)$. The *P-value fallacy* asserts that with $\Pr(E \mid H)$ to hand, one needs neither $\Pr(E \mid \text{not-}H)$ nor $\Pr(H)$; this will be discussed in section 10.3. The slightly milder *base rate fallacy* asserts that given the Bayes factor one does not need $\Pr(H)$.

P-value fallacy

base rate fallacy

10.2 *The base rate fallacy*

The natural domain of the base rate fallacy is the doctor's surgery. An asymptomatic patient undergoes a screening, which returns a positive result. The hypothesis H is that the patient has the disease. The evidence E is the positive result. The doctor knows the sensitivity and the specificity of the screening test; they are

Remember: *sensitivity* is how good the test is at identifying the disease when it is present, and *specificity* is how good the test is at not raising false alarms; see section 4.1.

$$\begin{aligned} \text{sensitivity} :& \Pr(E \mid H) = 0.50 \\ \text{specificity} :& 1 - \Pr(E \mid \text{not-}H) = 0.97. \end{aligned}$$

He reasons, "well, the specificity is basically one, so I guess the probability that my patient has the disease is 50%".[35] Wrong! He cannot make an inference about $\Pr(H \mid E)$ without also knowing the *base rate* of the disease, which is its relative frequency in the population of which his patient is a member, which provides $\Pr(H)$. In the illustration this is 0.3%, i.e. 3 in 1,000. So the posterior odds for H versus not-H are in fact

base rate

$$O^* = \frac{0.5}{1 - 0.97} \frac{0.003}{1 - 0.003} \approx 5\%$$

I hope it goes without saying that you should not enrol in a medical screening campaign without first knowing the sensitivity and specificity of the test, and the base rate of the disease in a cohort of people like you. You should also be clear about the nature of the intervention if the presence of the disease is suspected, and its success. All of these things are necessary in order for you to give informed consent; see Gigerenzer

[34] In fact the 'coin' was fair—I used a random number generator. In this example you might reply, "But actually the outcome was three heads in ten tosses, which has a probability of 0.12, i.e not that ususual at all." To which I would reply "One can always make an event seem less improbable by suppressing information, so that is hardly an enlightening argument."

[35] This example for colorectal cancer is taken from Gigerenzer (2003), chapter 6, Figures 6-1 and 6-2.

which, since the odds are close to zero, implies that the posterior probability of H is also about 5%. It is a horrible error for a doctor to tell his patient that probability of having the disease is one in two, when in fact it is one in twenty.

10.3 The P-value fallacy

The natural domain of the P-value fallacy is the courtroom, where it is termed the *prosecutor's fallacy*. The prosecutor erroneously (and perhaps maliciously) argues as follows. "The probability of crime-scene evidence E under the hypothesis H that the defendant is innocent is almost zero, and consequently he is almost certainty guilty", or, in our notation

prosecutor's fallacy

$$\text{“}\Pr(E \mid H) = \varepsilon \text{ implies that } \Pr(\text{not-}H \mid E) = 1 - \varepsilon.\text{”}$$

The prosecutor has clearly transposed the two arguments, and equated $\Pr(E \mid H)$ and $\Pr(H \mid E)$. Wrong!

We can see what is going on using Bayes's theorem. As $\Pr(E \mid \text{not-}H)$ is presumably one, the posterior odds for H versus not-H are

$$O^* = \frac{\varepsilon}{1} O$$

where O are the prior odds. As H and not-H are mutually exclusive, the posterior probability of H is

$$p^* = \frac{O^*}{1 + O^*} = \frac{\varepsilon O}{1 + \varepsilon O}.$$

Now if $O = 1$, then $p^* = \varepsilon/(1 + \varepsilon) \approx \varepsilon$, which is what the prosecutor is claiming. So in fact the prosecutor has assumed, implicitly, that the defendant is equally likely to be innocent or guilty, which is an assumption that he ought to be required to defend in the courtroom, in front of the defence lawyer and the judge (who would both be merciless, one hopes). Lindley (1991) discusses Bayesian reasoning in the courtroom.

As the name implies, though, the real P-value fallacy is being committed in statistics, where small P-values are being used to reject H_0 in favour of H_1, the alternative hypothesis. The illogic is identical to the prosecutor's fallacy, but with the additional complications that in statistics we cannot take $\Pr(E \mid H_1) = 1$, and, often, nor are H_0 and H_1 mutually exclusive. So the P-value fallacy in statistics is *worse* than the prosecutor's fallacy. The P-value fallacy is discussed in Goodman (1999a,b).

It is intuitive that we cannot use evidence to choose between hypotheses unless we know the probability of the evidence under both hypotheses. Bayes factors make this explicit. P-values cannot tell us what we need to know, because they omit any reference to H_1. Carefully designed statistical trials based on UMP tests are somewhat exempt from this criticism, but not from the criticism that the decision in such trials should depend on the costs of errors, as explained in section 9.2.

What, then, of the claim by R.A. Fisher, perhaps the most influential statistician of the C20th, about the inferential content of P-values. In an infamous passage, he writes, of a very small P-value:

> The force with which such a conclusion is supported is logically that of a simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the [hypothesis] is not true. (Fisher, 1956, p. 39)

Fisher would like to turn A = 'an event which is exceptionally rare under H has occurred', which is true, into something more. Although he says "simple disjunction" he must mean 'partition', since he presents us with the argument that in rejecting A we must therefore conclude B = 'the null hypothesis is not true'. But of course $\{A, B\}$ is not a partition, and, anyhow, we have no particular reason to reject A, given that improbable events occur all the time.

I should add that Fisher (1956) is one of the outstanding books in statistics, giving a wealth of evidence of Fisher's brilliance, and also that his "rich and manifold personality shows a few contradictions" (de Finetti, 1972, p. 171). Basu (1975, p. 38) also has an amusing example of Fisher's "euphoria".

The relationship between P-values and Bayes factors can be made explicit in one simple but very common case, discussed in section 9.3. Under the null hypothesis $\mu = 0$, $Z \sim \text{N}(z; 0, 1)$, and so the P-value for the z-score is

$$p = \int_z^\infty \phi(v)\,\mathrm{d}v = 1 - \Phi(z)$$

where ϕ and Φ are the PDF and distribution function of a standard Normal. And the extreme lower bound for the Bayes factor is $B_{01} \geq \exp(-z^2/2)$. So

$$B_{01} \geq \exp\big(-\{\Phi^{-1}(1-p)\}^2/2\big),$$

where Φ^{-1} is the quantile function of the standard Normal. This lower bound is shown in Figure 2. A P-value of 1% corresponds to an extreme lower bound on the Bayes factor of 0.07, or 1/14, hardly compelling evidence against H_0, even if the lower bound is attained. And yet standard practice would be to interpret a P-value of 1% as a strong reason to reject H_0 in favour of H_1. The Sellke *et al.* (2001) approach, mentioned in section 9.3, gives a lower bound for a P-value of 1% of 0.13, and a lower bound for a P-value of 5% of 0.41.

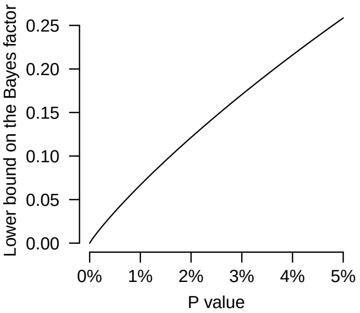

Figure 2: P values and extreme lower bounds on the Bayes factor, for the Normal case.

Does it matter? If statistics were a purely intellectual exercise, then probably not. But when statistics is used to inform decisions, then it does. In medical science, rejected null hypotheses in a laboratory lead to the assignment of animals and then humans to clinical trials. To be taken off a drug that is known to be effective and placed on an experimental drug for which the dominant source of evidence is a P-value of 1% is ethically dubious. Again, as in clinical screening, this is an issue of informed consent. But if medical scientists do not understand that a P-value of 1% is not compelling evidence against H_0, then what hope is there for the patients they recruit to their trials?

In medical science, recent research has suggested that many of the key experiments that rejected H_0 are not reproducible; see Ioannidis (2005, including the comment and rejoinder), or Begley and

Ellis (2012). Among other explanations, Ioannidis highlights the use of P-values, and the potential for experimental manipulation that might drive a summary P-value down below a critical threshold such as 5%, which is often necessary to secure a benefit for the researchers (e.g. publication in a prestigious journal).[36] Recently, Masicampo and Lalande (2012) have produced evidence of a 'bulge' of P-values just below 5% in three highly regarded journals in experimental psychology, which seems to support Ioannidis's claim. Masicampo and Lalande coyly refer to manipulation as "researcher degrees of freedom".

11 *The Likelihood Principle*

A principle is not a law; instead, it is something that one should not violate unknowingly or unthinkingly. Admissibility, for example; see section 8. This section considers some basic principles of statistical inference. This material is largely drawn from Birnbaum (1962, 1972), Savage *et al.* (1962), Basu (1975), Cox and Hinkley (1974, chapter 2), and the monograph by Berger and Wolpert (1984).

11.1 *Statistical evidence*

An experiment comprises the tuple $\mathcal{E} = \{\mathcal{X}, f_X, \Omega\}$. Two different experiments with the same parameter θ (and hence the same Ω) will be denoted $\mathcal{E}_1$ and $\mathcal{E}_2$: both the sample space and the statistical model can vary between the two experiments. An *experimental outcome* comprises the experiment *and* the value that results, i.e. $(\mathcal{E}_1, x_1)$ or $(\mathcal{E}_2, x_2)$. The crucial question is under what conditions two different experimental outcomes might give the exactly the same inferences about θ. We represent this situation by writing

experimental outcome

$$(\mathcal{E}_1, x_1) \sim (\mathcal{E}_2, x_2)$$

because this is an equivalance relation.

This section considers three principles which might be used to determine when $(\mathcal{E}_1, x_1) \sim (\mathcal{E}_2, x_2)$. The first, and simplest, is the *Invariance Principle (IP)*.

Invariance Principle (IP)

Definition 1 (Invariance Principle, IP)**.** *If $\mathcal{E}_1 = \mathcal{E}_2 = \mathcal{E}$ and $f_X(x; t) = f_X(x'; t)$ for all $t \in \Omega$, then $(\mathcal{E}, x) \sim (\mathcal{E}, x')$.*

To me, the IP asserts that we believe the statistical model f_X to be adequate, because otherwise we might think that a value x' which was different from x would tell us something different about θ, even though our model says otherwise. It is called the Indifference Principle because it is a special case of a more general principle that the inference about θ should respect the symmetries of the statistical model; see Cox and Hinkley (1974, section 2.3) or Basu (1975). For example, if the X's are IID, then the same inference should be made for any permutation of x.

The second principle is the *Weak Conditionality Principle (WCP)*.

Weak Conditionality Principle (WCP)

[36] Looking ahead to section 11.3, I suggest that this type of manipulation is easier for P-values than for Bayes factors, because P-values do not obey the Likelihood Principle, and so it is easier to compute an inappropriate P-value. There is plenty of evidence to support this abuse of P-values in Goldacre (2012, e.g. *Trials that stop early*, p. 184ff), but do not read this book if you want to preserve any illusions about the pharmaceutical industry operating in the interests of patients.

The story behind this principle was originally introduced by David Cox, and discussed in Cox (2006, section 4.3). There are two weighing scales in the laboratory downstairs: an accurate one and an inaccurate one. The accurate one is often busy, in which case I would use the inaccurate one, which is always available. I take my sample to the laboratory, and the accurate scales are free, and so I use them to make my measurement. When inferring the true weight of my sample, should I allow for the fact that I might have used the inaccurate scales, but did not? I doubt that anyone would say yes. This situation where the experiment to be performed is itself random is called a *mixed experiment*.

mixed experiment

Definition 2 (Weak Conditionality Principle, WCP). *Let $\mathcal{E}^*$ be the mixed experiment where experiment $\mathcal{E}_1$ is selected with known probability p_1 and experiment $\mathcal{E}_2$ is selected with probability $p_2 = 1 - p_1$. The experimental outcome for the mixed experiment is $(\mathcal{E}^*, (i, \boldsymbol{x}_i))$. The WCP asserts that*

$$(\mathcal{E}^*, (i, \boldsymbol{x}_i)) \sim (\mathcal{E}_i, \boldsymbol{x}_i). \qquad \text{(WCP)}$$

The WCP asserts that experiments which were not performed are immaterial; this has to be a sensible principle, otherwise where would we stop!

The third principle is the *Likelihood Principle (LP)*. It is helpful to introduce some more notation. Write $L^i_{\boldsymbol{x}_i}(t) := f^i_{\boldsymbol{X}_i}(\boldsymbol{x}_i; t)$. In this notation the IP asserts that

Likelihood Principle (LP)

$$L_{\boldsymbol{x}}(t) = L_{\boldsymbol{x}'}(t) \text{ for all } t \in \Omega \implies (\mathcal{E}, \boldsymbol{x}) \sim (\mathcal{E}, \boldsymbol{x}') \qquad \text{(IP)}$$

Now write

$$L^1_{\boldsymbol{x}_1} \propto L^2_{\boldsymbol{x}_2}$$

if there exists a $c > 0$ such that $L^1_{\boldsymbol{x}_1}(t) = cL^2_{\boldsymbol{x}_2}(t)$ for all $t \in \Omega$, where c might depend on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. (This is also an equivalence relation.) Informally, if $L^1_{\boldsymbol{x}_1} \propto L^2_{\boldsymbol{x}_2}$ then the two likelihood functions have the same shape.

Definition 3 (Likelihood Principle, LP). *Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two different experiments with the same parameter θ. Then*

$$L^1_{\boldsymbol{x}_1} \propto L^2_{\boldsymbol{x}_2} \implies (\mathcal{E}_1, \boldsymbol{x}_1) \sim (\mathcal{E}_2, \boldsymbol{x}_2). \qquad \text{(LP)}$$

In other words, if two experimental outcomes have the same shaped likelihood functions then they are equivalent for inferences about θ. This principle is highly contentious, because while Bayesian inference obeys it, Frequentist inference does not. This is discussed in section 11.3.

11.2 Birnbaum's bombshell

There are logical relationships between these three principles. One easy one is that LP $\implies$ IP, which follows immediately by letting $\mathcal{E}_2 = \mathcal{E}_1$ and $\boldsymbol{x}_2 = \boldsymbol{x}'$. A more interesting one is that LP $\implies$ WCP.

Proof. We have

$$L^*_{(i,x_i)}(t) = f^*_{(I,X_I)}(i, x_i; t) = p_i f^i_{X_i}(x_i; t)$$

because to get outcome (i, x_i) from the mixed experiment first one has to select experiment i, and then get x_i from this experiment. But $f^i_{X_i}(x_i; t) = L^i_{x_i}(t)$, by definition. Therefore $L^*_{(i,x_i)} \propto L^i_{x_i}$, and hence the LP implies that $(\mathcal{E}^*, (i, x_i)) \sim (\mathcal{E}_i, x_i)$, which is the WCP. □

Alan Birnbaum's bombshell, dropped in a slightly different form in Birnbaum (1962), was that (IP, WCP) $\implies$ LP. This proof comes from Birnbaum (1972).

Proof. Consider two experimental outcomes $(\mathcal{E}_1, x_1)$ and $(\mathcal{E}_2, x_2)$ for which $L^1_{x_1} \propto L^2_{x_2}$; i.e. for which there exists a $c > 0$ such that

$$f^1_{X_1}(x_1; t) = c f^2_{X_2}(x_2; t) \qquad \text{for all } t \in \Omega.$$

Consider the mixed experiment $\mathcal{E}^*$ with probability $p_1 = 1/(c+1)$ of experiment $\mathcal{E}_1$ and probability $p_2 = c/(c+1)$ of experiment $\mathcal{E}_2$. Then

$$\begin{aligned} L^*_{(1,x_1)}(t) = p_1 f^1_{X_1}(x_1; t) &= \frac{1}{c+1} f^1_{X_1}(x_1; t) \\ &= \frac{c}{c+1} f^2_{X_2}(x_2; t) = p_2 f^2_{X_2}(x_2; t) = L^*_{(2,x_2)}(t) \end{aligned}$$

for all $t \in \Omega$. So the IP implies that $(\mathcal{E}^*, (1, x_1)) \sim (\mathcal{E}^*, (2, x_2))$. But the WCP implies that $(\mathcal{E}^*, (i, x_i)) \sim (\mathcal{E}_i, x_i)$ for $i = 1, 2$, and, since $\sim$ is an equivalence relation, we infer that

$$L^1_{x_1} \propto L^2_{x_2} \implies (\mathcal{E}_1, x_1) \sim (\mathcal{E}_2, x_2)$$

which is the LP. □

So, in conclusion,

$$\text{(IP, WCP)} \iff \text{LP}.$$

It is illogical to accept both the IP and the WCP, both of which most people would accept, but also to deny the LP.

11.3 *Implications*

- Frequentist methods violate the LP because they use the sampling distribution of X to make inferences; e.g. for the assessment of properties of estimators, and for P-values, confidence sets, and hypothesis tests.

- The MLE as a point estimator for θ respects the LP, because the location of maximum depends only on the shape of the likelihood function.

- Bayesian methods respect the LP, because all likelihood functions with the same shape give rise to exactly the same posterior distribution for θ.

P-values violate the LP. Here is a simple illustration taken from Goodman (1999a). A trial on six IID patients had outcome $AAAAAB$. Consider two alternative models that are consistent with this outcome. In the first (Model 1, Binomial), the trial is over six patients. In the second (Model 2, truncated Geometric), the trial continues until the first B, or stops after six patients. Let θ be the probability of an A. Thus

$$\begin{array}{lll} \text{Model 1.} & x_1 = \text{Get five } A\text{'s in six trials} & L^1_{x_1}(t) = 6t^5(1-t) \\ \text{Model 2.} & x_2 = \text{Get a } B \text{ on the sixth trial} & L^2_{x_2}(t) = t^5(1-t). \end{array}$$

So the LP asserts that the same inference about θ should be drawn from either model, because $L^1_{x_1} \propto L^2_{x_2}$.

What about a P-value of $H_0 : \theta = 0.5$ against $H_1 : \theta > 0.5$? To compute a P-value we have to define the set of outcomes 'at least as extreme as' the actual outcome. For Model 1 this is

$$\{BAAAAA, ABAAAA, AABAAA, AAABAA, \\ AAAABA, AAAAAB, AAAAAA\},$$

while for Model 2 it is $\{AAAAAB, AAAAAA\}$. The P-value for Model 1 is

$$6\left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^6 = 0.109$$

And the P-value for Model 2 is

$$\left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^6 = 0.031;$$

i.e. the first P-value is $7/2$ times that of the second. In Model 1 H_0 might be accepted, while in Model 2 it might be rejected in favour of H_1. Hence the LP is violated.

The P-value has violated the LP because of its inclusion of parts of the sample space other than the point which actually occurred. The outcome $AAAAAB$ has the same probability in both models, but the 'at least as extreme as' set is different in the two models.

This is a good illustration to think about: *should* the answer depend on the probability of outcomes which did not occur? If you think it should then you have violated the LP and you must figure out which one of the IP and the WCP you reject.

11.4 Stopping rules

Sometimes an experimenter will decide to do n trials in his experiment, and then actually do n trials, and that will be the end of it. But often, the number of trials is not known at the start of the experiment, because the experiment is controlled by a *stopping rule* which depends on the sequence of outcomes. Here are some possible stopping rules in the case where $\mathcal{X} = \{0, 1\}$ and $\boldsymbol{X} \overset{\text{iid}}{\sim} \text{Bernoulli}(x; \theta)$:

stopping rule

1. Stop after 10 trials,
2. Stop after 3 successes,
3. Stop after 3 successes in a row,
4. Stop once we have recruited thirty women,
5. Stop once ten people have completed treatment,
6. Stop when the money runs out,
7. Stop when the P-value of $H_0 : \theta = 0.5$ drops below 0.05,
8. Do 10 trials, toss a coin, and if it is heads, do 10 more, then stop,

and so on; the open-ended rules will obviously have some upper limit, such as 100 trials. Each of these rules has a different sample space $\boldsymbol{\mathcal{X}}$; only in the first case is this equal to $\mathcal{X}^n$. The P-value has to be computed differently in each case, because the P-value reflects outcomes at least as extreme as $\boldsymbol{x}^{\text{obs}}$, which will depend on $\boldsymbol{\mathcal{X}}$.

Here is one way to cheat: keep extending the trial until you get a low P-value, where this P-value is (wrongly) computed as though $\boldsymbol{\mathcal{X}} = \mathcal{X}^n$.

This seems to present a real headache for the statistician. In order to compute an honest P-value (or do any other Frequentist procedure), the experimenter's stopping rule must be known precisely. More often than not, this meta-information about the experiment is not available, and yet statistics seems to be done. Why is this? One answer is the *Stopping Rule Principle (SRP)*.

Stopping Rule Principle (SRP)

Definition 4 (Stopping Rule Principle, SRP)**.** *Let $p_i(x_1, \dots, x_i)$ be the probability of continuing the experiment after seeing observations $x_1, \dots, x_i$. The SRP asserts that the stopping rule $\{p_0, p_1, \dots\}$ is immaterial for inferences about θ.*

This is obviously a very powerful principle, since it cuts through all the complications that arise from the variety of stopping rules that might be used in practice. What is fascinating and unexpected, though, it that it is a direct implication of the LP:

$$\text{LP} \implies \text{SRP}.$$

Proof. Let $f_i(x_i \mid x_1, \dots, x_{i-1}; \theta)$ be the probability of outcome x_i conditional on outcomes $x_1, \dots, x_{i-1}$. The case $n = 3$ is sufficient to see what it going on. Here is the likelihood function for the complete experiment, supposing it terminates after three trials:

$$\begin{aligned} L(t) &= \Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3; \theta = t) \\ &= p_0 f_1(x_1; t) \times p_1(x_1) f_2(x_2 \mid x_1; t) \times p_2(x_1, x_2) f_3(x_3 \mid x_1, x_2; t) \times (1 - p_3(x_1, x_2, x_3)) \\ &= p_0 p_1(x_1) p_2(x_1, x_2)(1 - p_3(x_1, x_2, x_3)) \prod_{i=1}^{3} f_i(x_i \mid x_1, \dots, x_{i-1}; t) \\ &\propto \prod_{i=1}^{3} f_i(x_i \mid x_1, \dots, x_{i-1}; t) = f_{\boldsymbol{X}}(\boldsymbol{x}; t), \end{aligned}$$

after dropping multiplicative terms that do not depend on t. Hence the SRP follows from the LP, because the actual experiment and the fixed-n experiment have the same shaped likelihood functions. □

In one of the funniest and most telling comments ever printed in a book on the foundations of statistical inference, L.J. Savage, who ranks alongside Fisher in the C20th pantheon of great statisticians, stated

> "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)

If your inference respects the LP (e.g. if you are using a Bayesian approach) then you get the SRP for free. If it does not (e.g. if you are using a Frequentist approach), then you will need to determine the precise stopping rule from the experimenter, and do the very complicated calculation over the $\mathcal{X}$ that is implied by this rule. If you cannot get hold of the stopping rule, or if you want to ignore it, you will need to provide an independent justification for the SRP, that is consistent with your violation of the LP.

11.5 Ancillary statistics

This may be the point to address a subtle feature of Frequentist inference. Estimators and hypothesis tests are evaluated according to their behaviour under randomness of $\boldsymbol{X}$, which currently comprises all of the random quantities in the assessment. But is it really relevant to consider *all* of the uncertainty in $\boldsymbol{X}$?

For example, if $\boldsymbol{X} = [\boldsymbol{Y}, \boldsymbol{Z}]$ and $\theta = (\theta', \theta'')$, where only θ' is interesting, it may be possible to represent the distribution of $\boldsymbol{X}$ as

$$f_{Y,Z}(y, z; \theta) = f_{Y|Z}(y \mid z; \theta') f_Z(z; \theta''). \tag{23}$$

In other words, conditioning on $\boldsymbol{Z} = z$ means that the nuisance parameters drop out. This is what happens in regression, for example, where the covariates $\boldsymbol{Z}$ are treated as known for the purposes of assessing the sampling distribution of the regression coefficients. A random quantity such $\boldsymbol{Z}$, for which

1. The marginal distribution f_Z does not depend on θ', and
2. the conditional distribution $f_{Y|Z}$ does not depend on θ''

is termed an *experimental ancillary statistic*, a term first proposed by Kalbfleisch (1975).

experimental ancillary statistic

Where (23) holds, the MLE and the Bayesian approach automatically condition on experimental ancillary statistics. In the Bayesian case, this requires that the prior distribution for θ is the product of prior distributions in θ' and θ''.

Proof.

$$\begin{aligned}
\pi^*_{\theta'}(t') &= \int_{\Omega''} \pi^*_{\theta}(t', t'')\,\mathrm{d}t'' \\
&\propto \int_{\Omega''} L(t') f_Z(z^{\text{obs}}; t'')\, \pi_{\theta'}(t')\, \pi_{\theta''}(t'')\,\mathrm{d}t' \\
&\propto L(t')\, \pi_{\theta'}(t'),
\end{aligned}$$

where $L(t') := f_{Y|Z}(y^{\text{obs}} \mid z^{\text{obs}}; t')$ is the *conditional likelihood*. □

Things are not so straightforward for the Frequentist approach, however. Let us term the principle that one should condition on experimental ancillary statistics the *Strong Conditionality Principle (SCP)*: it has to be its own principle because it is not implied by any other self-evident principle. The SCP is stronger than the WCP because the WCP required p_1 to be known, in order that it could be set apart from θ. The SCP implies the WCP by identifying $I \in \{1, 2\}$ with Z and p_1 with θ''.

Strong Conditionality Principle (SCP)

The SCP was justified by R.A. Fisher, using the notion of *relevance*. Of all the points in the sample space of $[Y, Z]$ only the points along the slice $[Y, Z = z^{\text{obs}}]$ are relevant to the particular inference for which this particular experiment has been conducted.[37] Some special cases indicate that the conditional likelihood gives a sensible Frequentist answer where the joint likelihood does not; for example, in those cases where the number of nuisance parameters increases with the number of observations, the *Neyman-Scott problem*.

Neyman-Scott problem

11.6 Summary

The value of statistical principles should not be overstated. Failure to comply is not critical, but one should not violate a principle through ignorance, and one should be prepared to defend a violation, or correct it. Consulting the score-card, the Bayesian approach is consistent with all of the principles mentioned here, while the Frequentist approach violates the LP. What is worse, most Frequentist statisticians would also accept the IP and the WCP, which is illogical. Further, in violating the LP, Frequentists statisticians need an independent justification of the SRP (which they must often invoke in practice). They also invoke a stronger conditionality principle, the SCP, which seems a bit *ad hoc* and opportunistic.

This seems to be an indefensible position, but only if you care about principles. The Frequentist statistician might well respond, "Principles are all very well, but I treat each problem as it arises, using my judgement and experience." This response is fine, but it is not a good platform from which to criticise the Bayesian approach as being too subjective.

Reading

Further details can be found in standard undergraduate statistics textbooks such as Rice (1994) and DeGroot and Schervish (2002); Silvey (1975) is short and readable. More advanced treatments, suitable for graduate students, can be found in Casella and Berger (2002), Young and Smith (2005) or Davison (2003); the latter has much additional material that makes it very useful for applied statistics. My main reference books are Cox and Hinkley (1974), Schervish (1995), and Robert (2007). All of the above books are

[37] Like many of Fisher's insights, this one has a slightly mystical flavour; and it also contradicted his views on randomisation, as discussed in Efron (1998).

quite heavy (literally); Cox (2006) is better for travelling and reflection.[38] Hacking (2001) has a broader and much less technical treatment of statistical induction. Grimmett and Stirzaker (2001) has more details on the probability calculus.

[38] I have lost my copy of Cox (2006), with copious marginalia: please look out for it!

Blogs. The following statisticians write blogs which include material on statistical inference (these are just the ones I know about):

- Larry Wasserman, `http://normaldeviate.wordpress.com/`
- Andrew Gelman, `http://andrewgelman.com/`
- Christian Robert, `http://xianblog.wordpress.com/`
- David Spiegelhalter, `http://understandinguncertainty.org/blog`
- Cosmo Shalizi, `http://masi.cscs.lsa.umich.edu/~crshalizi/weblog/`

In addition, there is often statistical material by

- Steve McIntyre, `http://climateaudit.org/`

on the vexed subject of climate science, and

- Ben Goldacre, `http://www.badscience.net/`

on the equally vexed subject of medical science and evidence-based policy.

References

C. Andrieu and J. Thoms, 2008. A tutorial on adaptive MCMC. *Statistics and Computing*, **18**, 343–373. 33

A. Baker, 2011. Simplicity. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/archives/sum2011/entries/simplicity/`. 48

M.S. Bartlett, 1957. A comment on D.V. Lindley's statistical paradox. *Biometrika*, **44**, 533–534. 48, 64

D. Basu, 1975. Statistical information and likelihood. *Sankhyā*, **37**(1), 1–71. With discussion. 4, 53, 54

M.A. Beaumont, W. Zhang, and D.J. Balding, 2002. Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035. 33

C.G. Begley and L.M. Ellis, 2012. Raise standards for preclinical cancer research. *Nature*, **483**, 531–533. 53

Y. Benjamini and Y. Hochberg, 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**(1), 289–300. 25

R. Beran, 1987. Prepivoting to reduce level error of confidence sets. *Biometrika*, **74**(3), 457–468. 16

J. Berger and R. Wolpert, 1984. *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics, second edition. Available online, `http://projecteuclid.org/euclid.lnms/1215466210`. 54

J.O. Berger and L.R. Pericchi, 2001. Objective Bayesian methods for model selection: Introduction and comparison. *IMS Lecture Notes – Monograph Series*, **38**, 135–207. 42, 43

J. Besag, P. Green, D. Higdon, and K. Mengerson, 1995. Bayesian computation and stochastic systems. *Statistical Science*, **10**(1), 3–41. With discussion 42–66. 33

A. Birnbaum, 1962. On the foundations of statistical inference. *Journal of the American Statistical Association*, **57**, 269–306. 54, 56

A. Birnbaum, 1972. More concepts of statistical evidence. *Journal of the American Statistical Association*, **67**, 858–861. 54, 56

G.E.P. Box, 1980. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, **143**(4), 383–430. With discussion. 18

L.D. Brown, T.T. Cai, and A. DasGupta, 2001. Interval estimation for a binomial proportion. *Statistical Science*, **16**(2), 101–117. With discussion, pp 117–133. 41

G. Casella and R.L. Berger, 2002. *Statistical Inference*. Pacific Grove, CA: Duxbury, 2nd edition. 7, 12, 17, 60

R.M. Cooke, 1991. *Experts in Uncertainty; Opinion and Subjective Probability in Science*. New York & Oxford: Oxford University Press. 29

D. R. Cox, 1958. *Planning of Experiments*. New York: John Wiley & Sons, Inc. 23

D.R. Cox, 2006. *Principles of Statistical Inference*. Oxford University Press. 55, 61

D.R. Cox and D.V. Hinkley, 1974. *Theoretical Statistics*. London: Chapman and Hall. 13, 14, 17, 54, 60

A.C. Davison, 2003. *Statistical Models*. Cambridge, UK: Cambridge University Press. 9, 60

A.C. Davison, D.V. Hinkley, and G.A. Young, 2003. Recent developments in bootstrap methodology. *Statistical Science*, **18**(2), 141–157. 17

B. de Finetti, 1972. *Probability, Induction and Statistics*. London: John Wiley & Sons. 28, 53

B. de Finetti, 1974/75. *Theory of Probability*. London: Wiley. Two volumes (2nd vol. 1975); A.F.M. Smith and A. Machi (trs.). 28

M. H. DeGroot and M.J. Schervish, 2002. *Probability and Statistics*. Reading, Mass.: Addison-Wesley Publishing Co., 3rd edition. 60

W. Edwards, H. Lindman, and L.J. Savage, 1963. Bayesian statistical inference for psychological research. *Psychological Review*, **70**(3), 193–242. 45

B. Efron, 1998. R.A. Fisher in the 21st century. *Statistical Science*, **13**(2), 95–114. With discussion, 114–122. 17, 60

B. Efron, 2008. Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, **23**(1), 1–22. 25

B. Efron and D.V. Hinkley, 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**(3), 457–482. 17

B. Efron and C. Morris, 1977. Stein's paradox in statistics. *Scientific American*, **236**(5), 119–127. 41

J.A. Ferreira and A.H. Zwinderman, 2006. On the Benjamini-Hochberg method. *The Annals of Statistics*, **34**(4), 1827–1849. 25

R.A. Fisher, 1956. *Statistical Methods and Scientific Inference*. Edinburgh and London: Oliver and Boyd. 53

P.H. Garthwaite, J.B. Kadane, and A. O'Hagan, 2005. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, **100**, 680–701. 29

A.E. Gelfand and D. Dey, 1994. Bayesian model choice: Asymptotic and exact calculations. *Journal Royal Statistical Society B*, **56**, 501–514. 49

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, 2003. *Bayesian Data Analysis*. Boca Raton, Florida: Chapman and Hall/CRC, 2nd edition. 33

G. Gigerenzer, 2003. *Reckoning with Risk: Learning to Live with Uncertainty*. Penguin. 50, 51

B. Goldacre, 2012. *Bad Pharma*. London: Fourth Estate. 54

S. Goodman, 1999a. Toward evidence-based medical statistics. 1: The p-value fallacy. *Annals of Internal Medicine*, **130**, 995–1004. 52, 57

S. Goodman, 1999b. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, **130**, 1005–1013. 52

S. Goodman and S. Greenland, 2007. Why most published research findings are false: Problems in the analysis. *PLoS Medicine*, **4**(4), e168. A longer version of the paper is available at `http://www.bepress.com/jhubiostat/paper135`. 63

P. Green, 1995. Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**(4), 711–732. 50

G.R. Grimmett and D.R. Stirzaker, 2001. *Probability and Random Processes*. Oxford, UK: Oxford University Press, 3rd edition. 5, 10, 30, 61

I. Hacking, 2001. *An Introduction to Probability and Inductive Logic*. Cambridge, UK: Cambridge University Press. 26, 61

J.P.A. Ioannidis, 2005. Why most published research findings are false. *PLoS Medicine*, **2**(8), e124. See also Goodman and Greenland (2007) and Ioannidis (2007). 53, 54

J.P.A. Ioannidis, 2007. Why most published research findings are false: Author's reply to Goodman and Greenland. *PLoS Medicine*, **4**(6), e215. 63

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–380. University of California Press, 1961. 41

W.H. Jefferys, 1990. Bayesian analysis of random event generator data. *Journal of Scientific Exploration*, **4**(2), 153–169. Available online, `http://bayesrules.net/papers/reg.pdf`. 20

R.C. Jeffrey, 2004. *Subjective Probability: The Real Thing*. Cambridge, UK: Cambridge University Press. Unfortunately this first printing contains quite a large number of typos. 28

H. Jeffreys, 1961. *Theory of Probability*. Oxford, UK: Oxford University Press, 3rd edition. 42

J.D. Kalbfleisch, 1975. Sufficiency and conditionality. *Biomtrika*, **62**(2), 251–259. 59

R.E. Kass and A.E. Raftery, 1995. Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795. 42

J.F.C. Kingman and S.J. Taylor, 1966. *Introduction to Measure and Probability*. Cambridge UK: Cambridge University Press. 10, 26

E.L. Lehmann and J.P. Romano, 2005. *Testing Statistical Hypotheses*. New York: Springer, 3rd edition. 20

D.V. Lindley, 1957. A statistical paradox. *Biometrika*, **44**, 187–192. See also Bartlett (1957). 48

D.V. Lindley, 1985. *Making Decisions*. London: John Wiley & Sons, 2nd edition. 28

D.V. Lindley, 1991. Subjective probability, decision analysis and their legal consequences. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **154**(1), 83–92. 52

D. MacKay, 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press. 34

K.V. Mardia, J.T. Kent, and J.M. Bibby, 1979. *Multivariate Analysis*. London: Harcourt Brace & Co. 16

E.J. Masicampo and D.R. Lalande, 2012. A peculiar prevalence of *p* values just below .05. *The Quarterly Journal of Experimental Psychology*. DOI:10.1080/17470218.2012.711335. 54

A. O'Hagan and J. Forster, 2004. *Bayesian Inference*, volume 2b of *Kendall's Advanced Theory of Statistics*. London: Edward Arnold, 2nd edition. 29

J. Pfanzagl, 1968. A characterization of the one parameter exponential family by existence of uniformly most powerful tests. *Sankhyā*, **30**(2), 147–156. 22

J.A. Rice, 1994. *Mathematical Statistics and Data Analysis*. Wadsworth Publishing Co. Inc., 2nd edition. 60

C.P. Robert, 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer. 29, 33, 34, 40, 42, 46, 60

C.P. Robert and G. Casella, 1999. *Monte Carlo Statistical Methods*. New York: Springer. 33

J.S. Rosenthal, 2006. *A first look at rigorous probability theory*. Singapore: World Scientific Publishing Co. Pte. Ltd., 2nd edition. 10

L.J. Savage, 1954. *The Foundations of Statistics*. New York: Dover, revised 1972 edition. 28, 39

L.J. Savage *et al.*, editors, 1962. *The Foundations of Statistical Inference*. London: Methuen. 29, 54, 59

M.J. Schervish, 1995. *Theory of Statistics*. New York: Springer. Corrected 2nd printing, 1997. 7, 9, 13, 37, 60

M. Schield and T.V.V. Burnham, 2008. Von Mises' frequentist approach to probability. Available at `http://www.statlit.org/pdf/2008SchieldBurnhamASA.pdf`. 26

T. Sellke, M.J. Bayarri, and J.O. Berger, 2001. Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, **55**(1), 62–71. 46, 53

S. Senn, 2003. *Dicing with death: Chance, Risk and Health*. Cambridge UK: Cambridge University Press. 50

S.D. Silvey, 1975. *Statistical Inference*. Boca Raton, Florida: Chapman and Hall/CRC. 60

J.Q. Smith, 2010. *Bayesian Decision Analysis: Principle and Practice*. Cambridge, UK: Cambridge University Press. 29

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde, 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. 48

D.J. Spiegelhalter and H. Riesch, 2011. Don't know, can't know: embracing deeper uncertainties when analysing risks. *Philosophical Transactions of the Royal Society, Series A*, **369**, 1–21. 43

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206. University of California Press, 1955. 41

A.W. van der Vaart, 1998. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press. 35

P. Whittle, 2000. *Probability via Expectation*. New York: Springer, 4th edition. 40

G.A. Young and R.L. Smith, 2005. *Essentials of Statistical Inference*. Cambridge UK: Cambridge University Press. 17, 60

J. Ziman, 2000. *Real Science: What it is, and what it means*. Cambridge, UK: Cambridge University Press. 26