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 November 27, 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 $\boldsymbol{X} := (X_1, \ldots, X_n) \in \boldsymbol{\mathcal{X}}$, where very often $\boldsymbol{\mathcal{X}} = \mathcal{X}^n$; likewise, $\boldsymbol{x} := (x_1, \ldots, x_n)$.[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] Examples where $\boldsymbol{\mathcal{X}} \neq \mathcal{X}^n$ are given in **??**.

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 **??**. If $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Y}} \times \boldsymbol{\mathcal{Z}}$, 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

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

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

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{\boldsymbol{z}} f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z}) = \sum_{\boldsymbol{z}} f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y} \mid \boldsymbol{z}) f_{\boldsymbol{Z}}(\boldsymbol{z}). \tag{1}$$

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

[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 $\sum_j p_j = 1$.

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

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

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

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

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

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

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.

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) = $ 'x 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 $a_1, \ldots, a_k$ 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

[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.

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):

$$\nabla := \begin{pmatrix} \partial/\partial t_1 \\ \vdots \\ \partial/\partial t_d \end{pmatrix}, \text{ and } \nabla^2 := \begin{pmatrix} \partial^2/(\partial t_1)^2 & \dots & \partial^2/\partial t_1 \partial t_d \\ \vdots & \ddots & \vdots \\ \partial^2/\partial t_d \partial t_1 & \dots & \partial^2/(\partial t_d)^2 \end{pmatrix}. \quad (3)$$

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 intruisive. 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 $f_X(\boldsymbol{x};\theta) > 0$ 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 θ. These are *regular models*.[4] One very useful function for parametric statistical models is the *score function*

support of $\boldsymbol{X}$

regular models

[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 θ is the main necessary condition.

score function

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

It is straightforward to prove that

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

for regular models.

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

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

which must therefore equal zero. □

The *Fisher information matrix* is defined as

Fisher information matrix

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

(it is a $d \times d$ matrix when θ is a vector); it appears at interesting times in the theory of estimators. It is straightforward to show that for regular models

$$i(\theta) = -E\{\nabla u(\boldsymbol{X}, \theta)^T; \theta\}.$$

When evaluating $i(\theta)$ it is often easier to compute the expectation of the derivatives of $u(\boldsymbol{X}, \theta)$ than to compute the variance of $u(\boldsymbol{X}, \theta)$.

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 $i_1(\theta) := \mathrm{Var}\{u(X_1, \theta); \theta\}$.

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 $(X_1, \ldots, X_n) \overset{\text{iid}}{\sim} \mathrm{N}(x; \mu, \sigma^2)$, 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]

[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 $\boldsymbol{X}$.

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 $f^{\psi}_{\boldsymbol{X}}(\boldsymbol{x}; \psi) = f_{\boldsymbol{X}}(\boldsymbol{x}; g^{-1}(\psi))$, 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 $f_X(x;\theta)$ explicitly as a formula, to remove any ambiguity; a statement such as '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

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

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 $f(x)$:

exponential family of distributions

$$g_X(x;\psi) \propto f(x)\exp\{\psi s(x)\},$$

where, in order to sum to one,

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

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_X and g_X have the same support, which does not depend on ψ.

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

$$g_X(x;\psi) \propto f(x)\exp\Big\{\textstyle\sum_j \psi_j s_j(x)\Big\}.$$

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

[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.

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)

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

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 funtion, then

transformation of the PDF

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x}) |\nabla g(\boldsymbol{x})^T|^{-1} \quad \text{where } \boldsymbol{x} := g^{-1}(\boldsymbol{y})$$

for $\boldsymbol{y} \in g(\mathcal{X})$, and zero otherwise; here $|\nabla g^T|$ is known as the *Jacobian* of the transformation, and is the determinant of the square matrix of first derivatives, remembering that g is a vector-valued function in general, with one component for each x_i.

Proof. Consider, just for simplicity, the case of scalar X and increasing g. Denote the probability distribution function of X as F_X, and similarly for Y. Then

$$F_Y(y) := \Pr\{Y \le y\} = \Pr\{g(X) \le y\} = \Pr\{X \le g^{-1}(y)\} = F_X(x)$$

with $x := g^{-1}(y)$. Differentiate with respect to y to find the PDF, giving

$$f_Y(y) = f_X(x) \frac{\mathrm{d}x}{\mathrm{d}y} = f_X(x) \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^{-1} = f_X(x) \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|^{-1},$$

the last equality following because g is increasing. This is easily generalised to the case where g is decreasing, for which $F_Y(y) = 1 - F_X(x)$. The full result, for vector $\boldsymbol{X}$ and $\boldsymbol{Y}$, is effectively the change of variables formula for multivariate integration. □

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 **??** 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 $F_Y(u + \delta) < u + \delta$. □

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.

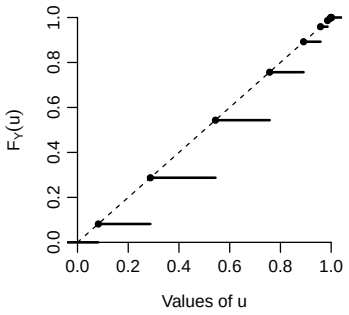

Figure 1: Distribution function of $Y = F_X(X)$, where $X \sim \text{Poisson}(x; \lambda = 2.5)$.

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)

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

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 $\boldsymbol{x}$ is termed a *statistic*. A statistic designed to estimate 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]

statistic

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 **??**.

[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.

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]

[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).

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

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

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 $\boldsymbol{X} \overset{\text{iid}}{\sim} \text{U}(x; 0, \theta)$, 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 $E\{1/S;\theta\} \neq 1/\,E\{S;\theta\} = 1/\theta$.

[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 **??**.

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 **??**.

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)

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

where $\nabla \,\text{bias} := \text{d}\,\text{bias}\,/\text{d}\theta$; 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,

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

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

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

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). □

This proof also demonstrates that an estimator attains the CRLB if and only if

$$u(\boldsymbol{x},\theta) = a(\theta)(s(\boldsymbol{x}) - \theta), \tag{9}$$

where a is any function of θ (and other values treated as known, such as n), which is necessary and sufficient to ensure that the correlation between S and $u(\boldsymbol{X},\theta)$ is ± 1 for any given θ, neglecting the trivial case where $s(\boldsymbol{x})$ is a constant. Estimators that attain the CRLB are termed *efficient estimators*. Integrating (9) and rearrang-

efficient estimators

ing shows that an estimator can attain the CRLB if and only if 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. if it is a subset of the exponential family of distributions; see section 1.3. It follows from the special properties of this family that where an estimator is efficient it is also the MLE.

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 **??**.

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 (10). 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 $\mathbb{P}(\theta) = \theta'$. 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 **??**. □

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 **??**.

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.

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

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

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 $u(\boldsymbol{X}, \theta) \xrightarrow{\mathrm{D}} \mathrm{N}_d\big(\mathbf{0}, i(\theta)\big)$, and this in turn implies that[13]

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

$$w(\boldsymbol{X}, \theta) := u(\boldsymbol{X}, \theta)^T i(\theta)^{-1} u(\boldsymbol{X}, \theta) \xrightarrow{\mathrm{D}} \chi^2_d, \tag{10}$$

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

$$\mathcal{C}(\boldsymbol{x}) := \big\{\theta : w(\boldsymbol{x}, \theta) \leq F^{-1}_{\chi^2_d}(1-\alpha)\big\} \tag{11}$$

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 (10). □

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. □

One loose end needs to be tidied up. The Fisher information $i(\theta)$ in (10) 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 **??**, $\hat{\theta}(\boldsymbol{X}) \xrightarrow{\text{P}} \theta$ 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 **??**. 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 **??**).

expected Fisher Information

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 (11) 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 **??**.

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.

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

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

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

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

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.

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

Monte Carlo integration

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

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

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

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 **??** 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.

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

Pearson's chi-squared test

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

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.

[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.

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 **??** and **??**. If you have a large enough n then your P-value is very likely to be small, unless you allow 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).

References

D. Basu, 1975. Statistical information and likelihood. *Sankhyā*, **37**(1), 1–71. With discussion. 3

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

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. 17

G. Casella and R.L. Berger, 2002. *Statistical Inference*. Pacific Grove, CA: Duxbury, 2nd edition. 6, 11, 17

D.R. Cox and D.V. Hinkley, 1974. *Theoretical Statistics*. London: Chapman and Hall. 13, 16

A.C. Davison, 2003. *Statistical Models*. Cambridge, UK: Cambridge University Press. 8

A.C. Davison, D.V. Hinkley, and G.A. Young, 2003. Recent developments in bootstrap methodology. *Statistical Science*, **18**(2), 141–157. 16

B. Efron, 1998. R.A. Fisher in the 21st century. *Statistical Science*, **13**(2), 95–114. With discussion, 114–122. 16

B. Efron and D.V. Hinkley, 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expectedfisher information. *Biometrika*, **65**(3), 457–482. 16

G.R. Grimmett and D.R. Stirzaker, 2001. *Probability and Random Processes*. Oxford University Press, 3rd edition. 4, 9

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`. 19

J.F.C. Kingman and S.J. Taylor, 1966. *Introduction to Measure and Probability*. Cambridge UK: Cambridge University Press. 9

K.V. Mardia, J.T. Kent, and J.M. Bibby, 1979. *Multivariate Analysis*. London: Harcourt Brace & Co. 15

J.S. Rosenthal, 2006. *A first look at rigorous probability theory*. Singapore: World Scientific Publishing Co. Pte. Ltd., 2nd edition. 9

M.J. Schervish, 1995. *Theory of Statistics*. New York: Springer. Corrected 2nd printing, 1997. 6, 8, 13

G.A. Young and R.L. Smith, 2005. *Essentials of Statistical Inference*. Cambridge UK: Cambridge University Press. 16