4
Confidence sets

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

In this chapter we have the tricky situation in which a specified function $g : \mathcal{Y} \times \Omega \to \mathbb{R}$ becomes a random quantity when Y is a random quantity. Then the distribution of $g(Y, \theta)$ depends on the value in Ω controlling the distribution of Y, which need not be the same value as θ in the argument. However, in this chapter the value in Ω controlling the distribution of Y will always be the same value as θ. Hence the random quantity $g(Y, \theta)$ has the distribution induced by $Y \sim f(\bullet; \theta)$.

4.1 Confidence procedures and confidence sets

A confidence procedure is a special type of decision rule for the problem of set estimation. Hence it is a function of the form $C : \mathcal{Y} \to 2^\Omega$, where 2^Ω is the set of all sets of Ω.[1]

[1] In this chapter I am using 'C' for a confidence procedure, rather than 'δ' for a decision rule.

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

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

If the probability equals $(1-\alpha)$ for all θ, then C is an *exact* level-$(1-\alpha)$ confidence procedure.[2]

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

The value $\Pr\{\theta \in C(Y); \theta\}$ is termed the *coverage* of C at θ. Thus a 95% confidence procedure has coverage of at least 95% for all θ, and an exact 95% confidence procedure has coverage of exactly 95% for all θ. The coverage of a confidence procedure is determined by its sampling distribution. Thus the decision to use, say, a 95% confidence procedure for inference about Θ violates the Likelihood Principle, according to Theorem 2.3 (see also Section 2.6.3).

Decision rules for set estimators were discussed in Section 3.6. A 95% confidence procedure is *not* a Bayes Rule for the loss function in (3.7). Nevertheless, confidence procedures *can* satisfy the condition that they are level sets of the likelihood function; i.e.

$$C(y) = \{\theta \in \Omega : f(y; \theta) \geq g(y)\} \tag{4.1}$$

for some g. I term this the *level set property (LSP)*. Recollect that the LSP is akin to a necessary condition for C to be an admissible set estimator under the loss function in (3.7), by Theorem 3.5.

The diameter of $C(y)$ can grow rapidly with its coverage.[3] In fact, the relation must be extrememly convex when coverage is nearly one, because, in the case where $\Omega = \mathbb{R}$, the diameter at coverage $= 1$ is unbounded. So an increase in the coverage from, say 95% to 99%, could correspond to a doubling or more of the diameter of the confidence procedure. For this reason, exact confidence procedures are highly valued, because a conservative 95% confidence procedure can deliver sets that are much larger than an exact one.

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

But, immediately a note of caution. It seems obvious that exact confidence procedures should be preferred to conservative ones, but this is easily exposed as a mistake. Suppose that $\Omega = \mathbb{R}$. Then the following procedure is an exact level-$(1-\alpha)$ confidence procedure for θ. First, draw a random variable U with a standard uniform distribution.[4] Then set

[4] See footnote 6.

$$C(y) := \begin{cases} \mathbb{R} & U \leq 1-\alpha \\ \{0\} & \text{otherwise.} \end{cases} \qquad (\dagger)$$

This is an exact level-$(1-\alpha)$ confidence procedure for θ, but also a meaningless one because it does not depend on y. If it is objected that this procedure is invalid because it includes an auxiliary random variable, then this rules out the method of generating approximately exact confidence procedures using bootstrap calibration (**??**). And if it is objected that confidence procedures must depend on y, then (†) could easily be adapted so that y is the seed of a numerical random number generator for U. So something else is wrong with (†). In fact, it lacks the LSP (see above, (4.1)).

It is helpful to distinguish between the confidence procedure C, which is a function of y, and the result when C is evaluated at the observations y^{obs}, which is a set in Ω. I like the terms used in Morey et al. (2016), which I will also adapt to p-values in Section 4.4.

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

So a confidence procedure is a function, and a confidence set is a set. If $\Omega \subset \mathbb{R}$ and $C(y^{\text{obs}})$ is convex, i.e. an interval, then a confidence set (interval) is represented by a lower and upper value. We should write, for example, "using procedure C, the 95% confidence interval for θ is $[0.55, 0.74]$", inserting "exact" if the confidence procedure C is exact.

4.2 *Families of confidence procedures*

The challenge with confidence procedures is to construct one with a specified level (look back to Section 1.4). One could propose an

arbitrary $C : \mathcal{Y} \to 2^\Omega$, and then laboriously compute the coverage for every $\theta \in \Omega$. At that point one would know the level of C as a confidence procedure, but it is unlikely to be 95%; adjusting C and iterating this procedure many times until the minimum coverage was equal to 95% would be exceedingly tedious. So we need to go backwards: start with the level, e.g. 95%, then construct a C guaranteed to have this level.

Define a *family of confidence procedures* as $C : \mathcal{Y} \times [0,1] \to 2^\Omega$, where $C(\cdot;\alpha)$ is a level-$(1-\alpha)$ confidence procedure for each α. If we start with a family of confidence procedures for a specified model, then we can compute a confidence set for any level we choose.

One class of families of confidence procedures has a natural and convenient form. The key concept is *stochastic dominance*. Let X and Y be two scalar random quantities. Then X stochastically dominates Y exactly when

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

Visually, the distribution function for X is never to the left of the distribution function for Y.[5] Although it is not in general use, I define the following term.

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

Definition 4.3 (Super-uniform)**.** The random quantity X is *super-uniform* exactly when it stochastically dominates a standard uniform random quantity.[6]

[6] A standard uniform random quantity being one with distribution function $F(u) = \max\{0, \min\{u, 1\}\}$.

In other words, X is super-uniform exactly when $\Pr(X \leq u) \leq u$ for all $0 \leq u \leq 1$. Note that if X is super-uniform then its support is bounded below by 0, but not necessarily bounded above by 1. Now here is a representation theorem for families of confidence procedures.[7]

[7] Look back to 'New notation' at the start of the Chapter for the definition of $g(Y;\theta)$.

Theorem 4.1 (Families of Confidence Procedures, FCP)**.** *Let $g : \mathcal{Y} \times \Omega \to \mathbb{R}$. Then*

$$C(y;\alpha) := \{\theta \in \Omega : g(y,\theta) > \alpha\} \tag{4.2}$$

is a family of level-$(1-\alpha)$ confidence procedures if and only if $g(Y,\theta)$ is super-uniform for all $\theta \in \Omega$. C is exact if and only if $g(Y,\theta)$ is uniform for all θ.

Proof.
($\Leftarrow$). Let $g(Y,\theta)$ be super-uniform for all θ. Then, for arbitrary θ,

$$\begin{aligned}\Pr\{\theta \in C(Y;\alpha);\theta\} &= \Pr\{g(Y,\theta) > \alpha;\theta\} \\ &= 1 - \Pr\{g(Y,\theta) \leq \alpha;\theta\} \\ &= 1 - (\leq \alpha) \geq 1 - \alpha\end{aligned}$$

as required. For the case where $g(Y,\theta)$ is uniform, the inequality is replaced by an equality.

($\Rightarrow$). This is basically the same argument in reverse. Let $C(\cdot;\alpha)$ defined in (4.2) be a level-$(1-\alpha)$ confidence procedure. Then, for arbtrary θ,

$$\Pr\{g(Y,\theta) > \alpha;\theta\} \geq 1 - \alpha.$$

Hence $\Pr\{g(Y,\theta) \le \alpha;\theta\} \le \alpha$, showing that $g(Y,\theta)$ is super-uniform as required. Again, if $C(\cdot;\alpha)$ is exact, then the inequality is replaced by a equality, and $g(Y,\theta)$ is uniform. □

Families of confidence procedures have the very intuitive *nesting property*, that

$$\alpha < \alpha' \implies C(y;\alpha) \supset C(y;\alpha'). \tag{4.3}$$

In other words, higher-level confidence sets are always supersets of lower-level confidence sets from the same family. This has sometimes been used as part of the definition of a family of confidence procedures (see, e.g., Cox and Hinkley, 1974, ch. 7), but I prefer to see it as a consequence of a construction such as (4.2).

It is interesting, and highly gratifying, that it is possible to construct families of confidence procedures with the LSP (eq. 4.1). Here is a result that has pedagogic value,[8] because it can be used to generate an uncountable number of families of confidence procedures, each with the LSP.

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

Theorem 4.2. *Let h be any PMF for Y. Then*

$$C(y;\alpha) := \{\theta \in \Omega : f(y;\theta) > \alpha \cdot h(y)\} \tag{4.4}$$

is a family of confidence procedures, with the LSP.

Proof. Define $g(y,\theta) := f(y;\theta)/h(y)$, which may be ∞. Then the result follows immediately from Theorem 4.1 because $g(Y,\theta)$ is super-uniform for each θ:

$$\begin{aligned}
\Pr\{f(Y;\theta)/h(Y) \le u;\theta\} &= \Pr\{h(Y)/f(Y;\theta) \ge 1/u;\theta\} \\
&\le \frac{\mathrm{E}\{h(Y)/f(Y;\theta);\theta\}}{1/u} && \text{Markov's inequality} \\
&\le \frac{1}{1/u} = u.
\end{aligned}$$

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

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

because h is a probability mass function. □

Among the interesting choices for h, one possibility is $h = f(\bullet;\omega)$, for some $\omega \in \Omega$. Note that with this choice, the confidence set of (4.4) always contains ω. So we know that we can construct a level-$(1-\alpha)$ confidence procedure whose confidence sets will always contain ω, for any $\omega \in \Omega$.

This is another illustration of the fact that the definition of a confidence procedure given in Definition 4.1 is too broad to be useful. But now we see that insisting on the LSP is not enough to resolve the issue. Two statisticians can both construct 95% confidence sets

for θ which satisfy the LSP, using different families of confidence procedures. Yet the first statistician may reject the null hypothesis that $H_0 : \Theta = \omega$ (see Section 3.7), and the second statistician may fail to reject it, for any $\omega \in \Omega$.

Actually, the situation is not as grim as it seems. Markov's inequality is very slack, and so the coverage of the family of confidence procedures defined in Theorem 4.2 is likely to be much larger than $(1 - \alpha)$, e.g. much larger than 95%. Remembering the comment about the rapid increase in the diameter of the confidence set as the coverage increases, from Section 4.1, a more likely outcome is that $C(y; 0.05)$ is large for many different choices of h, in which case no one rejects the null hypothesis.

All in all, it would be much better to use an exact family of confidence procedures which satisfy the LSP, if one existed. And, for perhaps the most popular model in the whole of Statistics, this is the case. This is the Linear Model; you will recognise it as the Normal or Gaussian model, usually in the form of a linear regression. I do not cover it here; see, e.g., Wood (2017, ch. 1). This model is a *very* special case, and it is unfortunate that so many people who are learning statistics have their intuition shaped by it.

4.3 *Marginalisation*

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

Theorem 4.3 (Confidence Procedure Marginalisation, CPM)**.** *Suppose that $g : \theta \mapsto \phi$, and that C is a level-$(1 - \alpha)$ procedure for θ. Then gC is a level-$(1 - \alpha)$ confidence procedure for ϕ.*[9]

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

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

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

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

This result shows that we can derive level-$(1 - \alpha)$ confidence procedures for functions of θ directly from level-$(1 - \alpha)$ confidence procedures for θ. But it also shows that the coverage of such derived procedures will typically be more than $(1 - \alpha)$, even if the original confidence procedure is exact.

4.4 *P-values*

There is a general theory for p-values, also known as *significance levels*, which is outlined in Section 4.4.2, and critiqued in Section 4.4.3. But first I want to focus on p-values as used in Hypothesis Tests, which is a very common situation. In this section I will take it for granted that a family of good confidence procedures has been selected.

4.4.1 *P-values and confidence sets*

Hypothesis Tests (HTs) were discussed in Section 3.7. In a binary HT the parameter space is partitioned as

$$\Omega = \{H_0, H_1\},$$

where often H_0 is a very small set, commonly degenerate. We 'reject' H_0 at a significance level of α exactly when a level-$(1-\alpha)$ confidence set $C(y^{\text{obs}}; \alpha)$ does not intersect H_0. Otherwise we 'fail to reject' H_0 at a significance level of α, in the common case where H_0 is degenerate.

In practice, then, a hypothesis test with a significance level of 5% (or any other specified value) returns one bit of information, 'reject', or 'fail to reject'. We do not know whether the decision was borderline or nearly conclusive; i.e. whether, for rejection, H_0 and $C(y^{\text{obs}}; 0.05)$ were close, or well-separated. We can increase the amount of information if C is a family of confidence procedures, in the following way.

Definition 4.4 (P-value, confidence set)**.** Let $C(\cdot\,; \alpha)$ be a family of confidence procedures. The p-value of H_0 is the smallest value α for which $C(y^{\text{obs}}; \alpha)$ does not intersect H_0.

The picture for determining the p-value is to dial up the value of α from 0 and shrink the set $C(y^{\text{obs}}; \alpha)$, until it is just clear of H_0. Of course we do not have to do this in practice. From the Representation Theorem (Theorem 4.1) we take $C(y^{\text{obs}}; \alpha)$ to be synonymous with a function $g : \mathcal{Y} \times \Omega \to \mathbb{R}$. Then $C(y^{\text{obs}}; \alpha)$ does not intersect with H_0 if and only if

$$\forall \theta \in H_0 : g(y^{\text{obs}}, \theta) \leq \alpha.$$

Thus the p-value is computed as

$$p(y^{\text{obs}}; H_0) := \max_{\theta \in H_0} g(y^{\text{obs}}, \theta), \tag{4.5}$$

for a specified family of confidence procedures (represented by the choice of g). Here is an interesting and suggestive result.[10] This will be the basis for the generalisation in Section 4.4.2.

[10] Recollect the definition of 'super-uniform' from Definition 4.3.

Theorem 4.4. *Under Definition 4.4 and* (4.5)*, $p(Y; H_0)$ is super-uniform for all $\theta \in H_0$.*

Proof. $p(y; H_0) \leq u$ implies that $g(y, \theta) \leq u$ for all $\theta \in H_0$. Hence

$$\Pr\{p(Y; H_0) \leq u; \theta\} \leq \Pr\{g(Y, \theta) \leq u; \theta\}$$

for all $\theta \in H_0$, where the final inequality follows because $g(Y, \theta)$ is super-uniform for all $\theta \in \Omega$, from Theorem 4.1. □

If interest concerns H_0, then $p(y^{\text{obs}}; H_0)$ definitely returns more information than a hypothesis test at any fixed significance level, because $p(y^{\text{obs}}; H_0) \leq \alpha$ implies 'reject H_0' at significance level α, and $p(y^{\text{obs}}; H_0) > \alpha$ implies 'fail to reject H_0' at signficance level α. But a p-value of, say, 0.045 would indicate a borderline 'reject H_0' at $\alpha = 0.05$, and a p-value of 0.001 would indicate nearly conclusive 'reject H_0' at $\alpha = 0.05$. So the following conclusion is rock-solid:

- When performing a HT, a p-value is more informative than a simple 'reject H_0' or 'fail to reject H_0' at a specified significance level (such as 0.05).

4.4.2 *The general theory of p-values*

Theorem 4.4 suggests a more general definition of a p-value, which does not just apply to hypothesis tests for parametric models, but which holds much more generally, for any PMF or model for Y. In the following f_0 is any *null model* for Y, including as a special case $f_0 = f(\bullet; \theta_0)$ for some specified $\theta_0 \in \Omega$.

Definition 4.5 (Significance procedure). $p : \mathcal{Y} \to \mathbb{R}$ is a *significance procedure* for f_0 exactly when $p(Y)$ is super-uniform under f_0. If $p_t(Y)$ is uniform under $Y \sim f_0$, then p is an *exact* significance procedure for f_0. The value $p_t(y^{\text{obs}})$ is a *significance level* or *p-value* for f_0 exactly when p is a significance procedure for f_0.

This definition can be extended to a set of PMFs for Y by requiring that p is a significance procedure for every element in the set; this is consistent with the definition of $p(y; H_0)$ in Section 4.4.1. The usual extension would be to take the maximum of the p-values over the set.[11]

[11] Although Berger and Boos (1994) have an interesting suggestion for parametric models.

For any specified f, there are a lot of significance procedures for $H_0 : Y \sim f$. An uncountable number, actually, because *every test statistic $t : \mathcal{Y} \to \mathbb{R}$ induces a significance procedure*. See Section 4.5 for the probability theory which underpins the following result.

Theorem 4.5. *Let $t : \mathcal{Y} \to \mathbb{R}$. Define*

$$p_t(y; f_0) := \Pr\{t(Y) \geq t(y); f_0\}.$$

Then $p_t(Y; f_0)$ is super-uniform under $Y \sim f_0$. That is, $p_t(\cdot\,; f_0)$ is a significance procedure for $H_0 : Y \sim f_0$. If the distribution function of $t(Y)$ is continuous, then $p_t(\cdot\,; f_0)$ is an exact *significance procedure for H_0.*

Proof.

$$p_t(y; f_0) = \Pr\{t(Y) \geq t(y); f_0\} = \Pr\{-t(Y) \leq -t(y); f_0\} =: G(-t(y))$$

where G is the distribution function of $-t(Y)$ under $Y \sim f_0$. Then

$$p_t(Y; f_0) = G(-t(Y))$$

which is super-uniform under $Y \sim f_0$ according to the Probability Integral Transform (see Section 4.5, notably Theorem 4.7). The PIT also covers the case where the distribution function of $t(Y)$ is continuous, in which case $p_t(\cdot\,; f_0)$ is uniform under $Y \sim f_0$. □

Like confidence procedures, significance procedures suffer from being too broadly defined. Every test statistic induces a significance procedure. This includes, for example, $t(y) = c$ for some specified constant c; but clearly a p-value based on this test statistic is useless.[12] So some additional criteria are required to separate out good from poor significance procedures. The most pertinent criterion is:

[12] It is a good exercise to check that $t(y) = c$ does indeed induce a super-uniform $p_t(Y; f_0)$ for every f_0.

- select a test statistic for which $t(Y)$ which will tend to be larger for decision-relevant departures from H_0.

This will ensure that $p_t(Y; f_0)$ will tend to be smaller under decision-relevant departures from H_0. Thus p-values offer a 'halfway house' in which an alterntive to H_0 is contemplated, but not stated explicitly.

Here is an example. Suppose that there are two sets of observations, characterised as $\boldsymbol{Y} \overset{\text{iid}}{\sim} f_0$ and $\boldsymbol{Z} \overset{\text{iid}}{\sim} f_1$, for unspecified PMFs f_0 and f_1. A common question is whether $\boldsymbol{Y}$ and $\boldsymbol{Z}$ have the same PMF, so we make this the null hypothesis:

$$H_0 : f_0 = f_1.$$

Under H_0, $(\boldsymbol{Y}, \boldsymbol{Z}) \overset{\text{iid}}{\sim} f_0$. Every test statistic $t(\boldsymbol{y}, \boldsymbol{z})$ induces a significance procedure. A few different options for the test statistic are:

1. The sum of the ranks of $\boldsymbol{y}$ in the ordered set of $(\boldsymbol{y}, \boldsymbol{z})$. This will tend to be larger if f_0 stochastically dominates f_1.
2. As above, but with $\boldsymbol{z}$ instead of $\boldsymbol{y}$.
3. The maximum rank of $\boldsymbol{y}$ in the ordered set of $(\boldsymbol{y}, \boldsymbol{z})$. This will tend to be larger if the righthand tail of f_0 is longer than that of f_1.
4. As above, but with $\boldsymbol{z}$ instead of $\boldsymbol{y}$.
5. The difference between the maximum and minimum ranks of $\boldsymbol{y}$ in the ordered set of $(\boldsymbol{y}, \boldsymbol{z})$. This will tend to be larger if f_0 and f_1 have the same location, but f_0 is more dispersed than f_1.
6. As above, but with $\boldsymbol{z}$ instead of $\boldsymbol{y}$.
7. And so on ...

There is no 'portmanteau' test statistic to examine H_0, and in my view H_0 should always be replaced by a much more specific null hypothesis which suggests a specific test statistic. For example,

$$H_0 : f_1 \text{ stochastically dominates } f_0.$$

In this case (2.) above is a useful test statistic. It is implemented as the *Wilcoxon rank sum test* (in its one-sided variant).

4.4.3 Being realistic about significance procedures

Section 4.4.1 made the case for reporting a HT in terms of a p-value. But what can be said about the more general use of p-values to 'score' the hypothesis $H_0 : Y \sim f_0$? Let's look at the logic. As Fisher himself stated, in reference to a very small p-value,

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

Fisher encourages us to accept that rare events seldom happen, and we should therefore conclude with him that a very small p-value strongly suggests that H_0 is not true. This is uncontroversial, although how small 'very small' should be is more mysterious; Cowles and Davis (1982) discuss the origin of the $\alpha = 0.05$ convention.

But what would he have written if the p-value had turned out to be large? The p-value is only useful if we conclude something different in this case, namely that H_0 is not rejected. But this is where Fisher would run into difficulties, because H_0 is an artefact: f_0 is a distribution chosen from among a small set of candidates for our convenience. So we know *a priori* that H_0 is false: nature is more complex than we can envisage or represent. Fisher's logical disjunction is trivial because the second proposition is always true (i.e. H_0 is always false). So either we confirm what we already know (small p-value, H_0 is false) or we fail to confirm what we already know (large p-value, but H_0 is still false). In the latter case, all that we have found out is that our choice of test statistic is not powerful enough to tell us what we already know to be true.

This is not how people who use p-values want to interpret them. They want a large p-value to mean "No reason to reject H_0", so that when the p-value is small, they can "Reject H_0". They do not want it to mean "My test statistic is not powerful enough to tell me what I already know to be true, namely that H_0 is false." But unfortunately that is what it means.

Statisticians have been warning about misinterpreting p-values for nearly 60 years (dating from Lindley, 1957). They continue to do so in fields which use statistical methods to examine hypotheses, indicating that the message has yet to sink in. So there is now a huge literature on this topic. A good place to start is Wasserstein

and Lazar (2016) and then Greenland and Poole (2013), and then work backwards.

4.5 The Probability Integral Transform

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

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

Define the *quantile function*

$$F^-(u) := \inf\{x \in \mathbb{R} : F(x) \geq u\}. \tag{4.6}$$

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

Theorem 4.6 (Probability Integral Transform, PIT)**.** *Let U have a standard uniform distribution. If F^- is the quantile function of X, then $F^-(U)$ and X have the same distribution.*

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

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

because then

$$\Pr\{F^-(U) \leq x\} = \Pr\{U \leq F(x)\} = F(x)$$

as required. So stare at Figure 4.1 for a while.

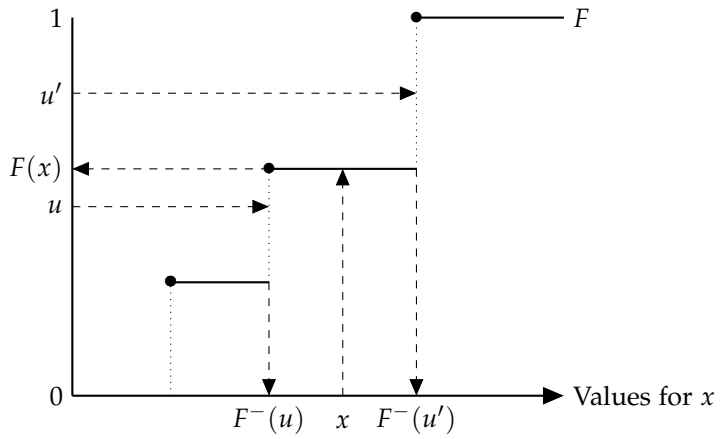

Figure 4.1: Figure for the proof of Theorem 4.6. The distribution function F is non-decreasing and continuous from the right. The quantile function F^- is defined in (4.6).

It is easy to check that

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

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

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

Taking the contrapositive of this second implication gives

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

which is the other half of (†). □

Theorem 4.6 is the basis for the following result; recollect the definition of a super-uniform random quantity from Definition 4.3. This result is used in Theorem 4.5.

Theorem 4.7. *If F is the distribution function of X, then $F(X)$ has a super-uniform distribution. If F is continuous then $F(X)$ has a uniform distribution.*

Proof. Check from Figure 4.1 that $F(F^-(u)) \geq u$. Then

$$\begin{aligned} \Pr\{F(X) \leq u\} &= \Pr\{F(F^-(U)) \leq u\} \qquad \text{from Theorem 4.6} \\ &\leq \Pr\{U \leq u\} \\ &= u. \end{aligned}$$

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

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

as required. □

5
Bibliography

Bartlett, M. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, 44:533–534. 59

Basu, D. (1975). Statistical information and likelihood. *Sankhyā*, 37(1):1–71. With discussion. 14, 15, 16, 21

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York, Inc., NY, USA, second edition. 34

Berger, J. and Boos, D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89:1012–1016. 51

Berger, J. and Wolpert, R. (1988). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward CA, USA, second edition. Available online, `http://projecteuclid.org/euclid.lnms/1215466210`. 14, 19

Bernardo, J. and Smith, A. (2000). *Bayesian Theory*. John Wiley & Sons Ltd, Chichester, UK. (paperback edition, first published 1994). 26

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57:269–306. 13, 16

Birnbaum, A. (1972). More concepts of statistical evidence. *Journal of the American Statistical Association*, 67:858–861. 14, 16

Casella, G. and Berger, R. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury, 2nd edition. 1, 3

Cormen, T., Leiserson, C., and Rivest, R. (1990). *Introduction to Algorithms*. The MIT Press, Cambridge, MA. 10

Cowles, M. and Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, 37(5):553–558. 53

Cox, D. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge, UK. 1

Cox, D. and Donnelly, C. (2011). *Principles of Applied Statistics*. Cambridge University Press, Cambridge, UK. 1

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London, UK. 16, 17, 48

Davison, A. (2003). *Statistical Models*. Cambridge University Press, Cambridge, UK. 3

Dawid, A. (1977). Conformity of inference patterns. In Barra, J. et al., editors, *Recent Developments in Statistcs*. North-Holland Publishing Company, Amsterdam. 14, 15

Edwards, A. (1992). *Likelihood*. The Johns Hopkins University Press, Baltimore, USA, expanded edition. 22

Edwards, W., Lindman, H., and Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242. 44

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236(5):119–127. Available at `http://statweb.stanford.edu/~ckirby/brad/other/Article1977.pdf`. 40

Ferguson, T. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, London, UK. 34, 36

Fisher, R. (1956). *Statistical Methods and Scientific Inference*. Edinburgh and London: Oliver and Boyd. 16, 53

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton FL, USA, 3rd edition. Online resources at `http://www.stat.columbia.edu/~gelman/book/`. 9

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London, UK. 34

Greenland, S. and Poole, C. (2013). Living with *P* values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*, 24(1):62–68. With discussion and rejoinder, pp. 69–78. 54

Hacking, I. (1965). *The Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK. 22

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge University Press, Cambridge, UK. 1

Hacking, I. (2014). *Why is there a Philosophy of Mathematics at all?* Cambridge University Press, Cambridge, UK. 2

Harford, T. (2014). Big data: Are we making a big mistake? *Financial Times Magazine*. Published online Mar 28, 2014. Available at `http://on.ft.com/P0PVBF`. 11, 19

Lad, F. (1996). *Operational Subjective Statistical Methods*. New York: John Wiley & Sons. 2

Le Cam, L. (1990). Maximum likelihood: An introduction. *International Statistical Review*, 58(2):153–171. 5, 23

Lindley, D. (1957). A statistical paradox. *Biometrika*, 44:187–192. See also Bartlett (1957). 53

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). *The BUGS Book: A Practical introduction to Bayesian Analysis*. CRC Press, Boca Raton FL, USA. 9

MacKay, D. (2009). *Sustainable Energy – Without the Hot Air*. UIT Cambridge Ltd, Cambridge, UK. available online, at `http://www.withouthotair.com/`. 2

Madigan, D., Strang, P., Berlin, J., Schuemie, M., Overhage, J., Suchard, M., Dumouchel, B., Hartzema, A., and Ryan, P. (2014). A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, 1:11–39. 8

Milner, K. and Rougier, J. (2014). How to weigh a donkey in the Kenyan countryside. *Significance*, 11(4):40–43. 38

Morey, R., Hoekstra, R., Rouder, J., Lee, M., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bullentin & Review*, 23(1):103–123. 46

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. New York: Springer, 2nd edition. 4

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Clarendon Press. 22

Pearl, J. (2016). The Sure-Thing Principle. *Journal of Causal Inference*, 4(1):81–86. 19

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall/CRC Press, Boca Raton FL, USA. 22

Samworth, R. (2012). Stein's paradox. *Eureka*, 62:38–41. Available online at `http://www.statslab.cam.ac.uk/~rjs57/SteinParadox.pdf`. Careful readers will spot a typo in the maths. 40

Savage, L. (1954). *The Foundations of Statistics*. Dover, New York, revised 1972 edition. 19

Savage, L. et al. (1962). *The Foundations of Statistical Inference*. Methuen, London, UK. 1, 22

Schervish, M. (1995). *Theory of Statistics*. Springer, New York NY, USA. Corrected 2nd printing, 1997. 1, 5, 9, 34, 36

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64(4):583–616. With discussion, pp. 616–639. 9

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B*, 76(3):485–493. 9

Stigler, S. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard University Press, Cambridge MA, USA. 1, 3, 28

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK. 25

Wasserstein, R. and Lazar, N. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133. 53

Wood, S. (2017). *Generalized Linear Models: An Introduction with R*. CRC Press, Boca Raton FL, USA, 2nd edition. 49

Ziman, J. (2000). *Real Science: What it is, and what it means*. Cambridge University Press, Cambridge, UK. 28