5
Confidence sets

This chapter is a continuation of Chapter 4, and the same conditions hold; re-read the introduction to Chapter 4 if necessary, and the start of Sec. 4.2. In brief, interest focuses on the parameter θ in the model

$$Y \sim f(\cdot\,;\theta) \quad \text{for some } \theta \in \Omega, \tag{5.1}$$

where Y are observables and $f(y;\theta)$ is assumed to be easily computed. The parameter space is denoted

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

for simplicity, even though the parameter may be vector-valued, and the parameter space may be uncountable; typically the parameter space is a convex subset of a finite-dimensional Euclidean space. An element of Ω is denoted θ_j, and the observed value of Y is denoted y^{obs}.

Throughout this chapter we accept that it is useful to make inferences about the parameter of the statistical model. I regard this notion as unscientific, as explained in **??**. Nevertheless, confidence-set-based hypothesis testing (Sec. 4.6) is very widely practiced and, accepting that this is not going to change, it is important that we use the best confidence sets that we can. For this purpose we must sometimes suppose the 'truth' of the statistical model, and refer to the 'true' parameter, which I will denote as θ.

New notation. 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 $g(Y, \theta_j)$ depends on the value of θ. Often the value of θ will be the same value as the second argument to g, but this is not implied by simply writing $g(Y, \theta_j)$. So it is best to make the value of θ explicit, when writing about the distribution of $g(Y, \theta_j)$. Hence I write $g(Y, \theta_j)\big|_{\theta=\theta_j}$ to indicate the random quantity $g(Y, \theta_j)$ when $Y \sim f(\cdot\,;\theta_j)$.

5.1 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] Decision rules for set estimators were discussed in Sec. 4.5.

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

Definition 10 (Confidence procedure). *$C : \mathcal{Y} \to 2^\Omega$ is a level-$(1-\alpha)$ confidence procedure exactly when*

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

If the probability equals $(1-\alpha)$ for all θ_j, 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 $\Pr\{\theta_j \in C(Y); \theta_j\}$ is termed the *coverage* of C at θ_j. Thus a 95% confidence procedure has coverage of at least 95% for all θ_j, and an exact 95% confidence procedure has coverage of exactly 95% for all θ_j. The diameter of $C(y)$ can grow rapidly with its coverage.[3] In fact, the relation must be extremely 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 (Sec. 5.3.3). 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 fails a necessary condition for admissibility that was derived in Sec. 4.5. This will be discussed in Sec. 5.2.

It is helpful to distinguish between the confidence procedure C, which is a function of y, and the result when C is evaluated at $y \leftarrow y^{\text{obs}}$, which is a set in Ω. I like the terms used in Morey *et al.* (2015), which I will also adapt to P-values in Sec. 5.5.

Definition 11 (Confidence set). *$C(y^{obs})$ 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.

5.2 *Families of confidence procedures*

The trick with confidence procedures is to construct one with a specified level, or, failing that, a specified lower bound on the level. One could propose an arbitrary $C : \mathcal{Y} \to 2^{\Omega}$, and then laboriously compute the coverage for every $\theta_j \in \Omega$. 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.

It turns out that families of confidence procedures all have the same form. The key concept is *stochastic dominance*. Let X and Y be two scalar random quantities. Then X stochastically dominates Y exactly when

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

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 $F(x) := \Pr(X \dot{\leq} x)$ for $x \in \mathbb{R}$.

Definition 12 (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 $g(Y; \theta_j)\big|_{\theta=\theta_j}$.

Theorem 5.1 (Families of Confidence Procedures, FCP)**.** *Let $g : \mathcal{Y} \times \Omega \to \mathbb{R}$. Then*

$$C(y; \alpha) := \{\theta_j \in \Omega : g(y, \theta_j) > \alpha\} \tag{5.2}$$

is a family of level-$(1-\alpha)$ confidence procedures if and only if $g(Y, \theta_j)\big|_{\theta=\theta_j}$ is super-uniform for all $\theta_j \in \Omega$. $C(\cdot; \alpha)$ is exact if and only if $g(Y, \theta_j)\big|_{\theta=\theta_j}$ is uniform for all θ_j.

Proof.
($\Leftarrow$). Let $g(Y,\theta_j)\big|_{\theta=\theta_j}$ be super-uniform for all θ_j. Then, for arbitrary θ_j,

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

as required. For the case where $g(Y,\theta_j)\big|_{\theta=\theta_j}$ is uniform, the inequality is replaced by an equality.

($\Rightarrow$). This is basically the same argument in reverse. Let $C(\cdot;\alpha)$ defined in (5.2) be a level-$(1-\alpha)$ confidence procedure. Then, for arbtrary θ_j,

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

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

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

$$\alpha < \alpha' \implies C(y;\alpha) \supset C(y;\alpha'). \tag{5.3}$$

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 an unavoidable consequence of the fact that all families must be defined using (5.2) for some g.

* * *

Sec. 4.5 made a recommendation about set estimators for θ, which was that confidence procedures should be based on level sets of $\{f(y;\theta_j) : \theta_j \in \Omega\}$. This was to satisfy a necessary condition to be admissible under the loss function (4.7). Here I restate that recommendation as a property.

Definition 13 (Level Set Property, LSP)**.** *A confidence procedure C has the Level Set Property exactly when*

$$C(y) = \{\theta_j \in \Omega \text{ such that } f(y;\theta_j) > c\}$$

for some c which may depend on y. A family of confidence procedures has the LSP exactly when $C(\cdot\,;\alpha)$ has the LSP for all α, for which c may depend on y and α.

A family of confidence procedures does not necessarily have the LSP. So it is not obvious, but highly gratifying, that it is possible to construct families of confidence procedures with the LSP. Three different approaches are given in the next section.

5.3 Methods for constructing confidence procedures

All three of these methods produce families of confidence procedures with the LSP. This is a long section, and there is a summary in Sec. 5.3.4.

5.3.1 Markov's inequality

Here is a result that has pedagogic value, because it can be used to generate an uncountable number of families of confidence procedures, each with the LSP.

Theorem 5.2. *Let h be any PMF for Y. Then*

$$C(y;\alpha) := \{\theta_j \in \Omega : f(y,\theta_j) > \alpha \cdot h(y)\} \tag{5.4}$$

is a family of confidence procedures, with the LSP.

Proof. Define $g(y,\theta_j) := f(y;\theta_j)/h(y)$, which may be ∞. Then the result follows immediately from Thm 5.1 because $g(Y,\theta_j)\big|_{\theta=\theta_j}$ is super-uniform for each θ_j:

$$\begin{aligned}
\Pr\{f(Y;\theta_j)/h(Y) \dot{\leq} u;\theta_j\} &= \Pr\{h(Y)/f(Y;\theta_j) \dot{\geq} 1/u;\theta_j\} \\
&\leq \frac{\mathrm{E}\{h(Y)/f(Y;\theta_j);\theta_j\}}{1/u} \qquad \text{Markov's inequality, (??)} \\
&\leq \frac{1}{1/u} = u.
\end{aligned}$$

For the final inequality,

$$\begin{aligned}
\mathrm{E}\{h(Y)/f(Y;\theta_j);\theta_j\} &= \sum_{y\in\operatorname{supp} f(\cdot;\theta_j)} \frac{h(y)}{f(y;\theta_j)} \cdot f(y;\theta_j) \quad \text{FTP, Thm 1.2} \\
&= \sum_{y\in\operatorname{supp} f(\cdot;\theta_j)} h(y) \\
&\leq 1.
\end{aligned}$$

If $\operatorname{supp} h \subset \operatorname{supp} f(\cdot;\theta_j)$, then this inequality is an equality. □

Among the interesting choices for g, one possibility is $g \leftarrow f(\cdot;\theta_i)$, for $\theta_i \in \Omega$. Note that with this choice, the confidence set of (5.4) always contains θ_i. So we know that we can construct a level-$(1-\alpha)$ confidence procedure whose confidence sets will always contain θ_i, for any $\theta_i \in \Omega$.

This is another illustration of the fact that the definition of a confidence procedure given in Def. 10 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 $H_0 : \theta = \theta_i$ (see Sec. 4.6), and the second statistician may fail to reject it, for any $\theta_i \in \Omega$.

Actually, the situation is not as grim as it seems. Markov's inequality is very slack (refer to its proof at eq. **??**), and so the coverage of the family of confidence procedures defined in Thm 5.2

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 Sec. 5.1, 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, if one existed. And, for perhaps the most popular model in the whole of Statistics, this is the case.

5.3.2 *The Linear Model*

The Linear Model (LM) is commonly expressed as

$$Y \stackrel{\mathrm{D}}{=} X\beta + \epsilon \quad \text{where } \epsilon \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 I_n) \tag{5.5}$$

where Y is an n-vector of observables, X is a specified $n \times p$ matrix of *regressors*, β is a p-vector of *regression coefficients*, and ϵ is an n-vector of *residuals*.[8] The parameter is $(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_{++}$.

[8] Usually I would make Y and ϵ bold, being vectors, and I would prefer not to use X for a specified matrix, but this is the standard notation.

'$\mathrm{N}_n(\cdot)$' denotes the n-dimensional *Multinormal distribution* with specified expectation vector and variance matrix (see, e.g., Mardia *et al.*, 1979, ch. 3). The symbol '$\stackrel{\mathrm{D}}{=}$' denotes 'equal in distribution'; this notation is useful here because the Multinormal distribution is closed under affine transformations. Hence Y has a Multinormal distribution, because it is an affine transformation of ϵ. So the LM must be restricted to applications for which Y can be thought of, at least approximately, as a collection of n random quantities each with realm $\mathbb{R}$, and for each of which our uncertainty is approximately symmetric. Many observables fail to meet these necessary conditions (e.g. applications in which Y is a collection of counts); for these applications, we have *Generalized Linear Models (GLMs)*. GLMs retain many of the attractive properties of LMs.

Wood (2015, ch. 7) provides an insightful summary of the LM, while Draper and Smith (1998) give many practical details.

Now I show that the Maximum Likelihood Estimator (MLE) of (5.5) is

$$\begin{aligned} \hat{\beta}(y) &= (X^T X)^{-1} X^T y \\ \widehat{\sigma^2}(y) &= n^{-1} (y - \hat{y})^T (y - \hat{y}) \end{aligned}$$

where $\hat{y} := X\hat{\beta}(y)$.

Proof. For a LM, it is more convenient to minimise $-2 \log f(y; \beta_j, \sigma_j^2)$ over (β_j, σ_j^2) than to maximise $f(y; \beta_j, \sigma_j^2)$.[9] Then

[9] Note my insistence that (β_j, σ_j^2) be considered as an element of the parameter space, *not* as the 'true' value.

$$-2 \log f(y; \beta_j, \sigma_j^2) = n \log(2\pi\sigma_j^2) + \frac{1}{\sigma_j^2} (y - X\beta_j)^T (y - X\beta_j)$$

from the PDF of the Multinormal distribution. Now use a simple device to show that this is minimised at $\beta_j = \hat{\beta}(y)$ for all values of

σ_j^2. I will write $\hat{\beta}$ rather than $\hat{\beta}(y)$:

$$\begin{aligned}(y - X\beta_j)^T(y - X\beta_j) \\ &= (y - X\hat{\beta} + X\hat{\beta} - X\beta_j)^T(y - X\hat{\beta} + X\hat{\beta} - X\beta_j) \\ &= (y - \hat{y})^T(y - \hat{y}) + 0 + (X\hat{\beta} - X\beta_j)^T(X\hat{\beta} - X\beta_j) \qquad (\dagger)\end{aligned}$$

where multiplying out shows that the cross-product term in the middle is zero. Only the final term contains β_j. Writing this term as

$$(\hat{\beta} - \beta_j)^T(X^TX)(\hat{\beta} - \beta_j)$$

shows that if X has full column rank, so that X^TX is positive definite, then $(\dagger)$ is minimised if and only if $\beta_j = \hat{\beta}$. Then

$$-2\log f(y; \hat{\beta}, \sigma_j^2) = n\log(2\pi\sigma_j^2) + \frac{1}{\sigma_j^2}(y - \hat{y})^T(y - \hat{y}).$$

Solving the first-order condition gives the MLE for $\widehat{\sigma^2}(y)$, and it is easily checked that this is a global minimum. □

Now suppose we want a confidence procedure for β. For simplicity, I will assume that σ^2 is specified, and for practical purposes I would replace it by $\widehat{\sigma^2}(y^{\text{obs}})$ in calculations. This is known as *plugging in* for σ^2. The LM extends to the case where σ^2 is not specified, but, as long as $n/(n-p) \approx 1$, it makes little difference in practice to plug in.[10]

[10] As an eminent applied statistician remarked to me: it if matters to your conclusions whether you use a standard Normal distribution or a Student-t distribution, then you probability have bigger things to worry about.

With β_j representing an element of the β-parameter space $\mathbb{R}^p$, and σ^2 specified, we have, from the results above,

$$-2\log\left(\frac{f(y;\beta_j,\sigma^2)}{f(y;\hat{\beta}(y),\sigma^2)}\right) = \frac{1}{\sigma^2}\{\hat{\beta}(y) - \beta_j\}^T(X^TX)\{\hat{\beta}(y) - \beta_j\}. \quad (5.6)$$

Now suppose we could prove the following.

Theorem 5.3. *With σ^2 specified,*

$$\frac{1}{\sigma^2}\{\hat{\beta}(Y) - \beta_j\}^T(X^TX)\{\hat{\beta}(Y) - \beta_j\}\big|_{\beta=\beta_j}$$

has a χ_p^2 distribution.

We could define the decision rule:

$$C(y;\alpha) := \left\{\beta_j \in \mathbb{R}^p : -2\log\left(\frac{f(y;\beta_j,\sigma^2)}{f(y;\hat{\beta}(y),\sigma^2)}\right) < \chi_p^{-2}(1-\alpha)\right\}. \quad (5.7)$$

where $\chi_p^{-2}(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the χ_p^2 distribution. Under Thm 5.3, (5.6) shows that C in (5.7) would be an exact level-$(1-\alpha)$ confidence procedure for β; i.e. it provides a family of exact confidence procedures. Also note that it satisfies the LSP from Def. 13.

After that build-up, it will come as no surprise to find out that Thm 5.3 is true. Substituting Y for y in the MLE of β gives

$$\hat{\beta}(Y) \overset{\text{D}}{=} (X^TX)^{-1}X^T(X\beta + \epsilon) \overset{\text{D}}{=} \beta + (X^TX)^{-1}X^T\epsilon,$$

writing σ for $\sqrt{\sigma^2}$. So the distribution of $\hat{\beta}(Y)$ is another Multinormal distribution

$$\hat{\beta}(Y) \sim \mathrm{N}_p(\beta, \Sigma) \quad \text{where } \Sigma := \sigma^2 (X^T X)^{-1}.$$

Now apply a standard result for the Multinormal distribution to deduce

$$\{\hat{\beta}(Y) - \beta_j\}^T \Sigma^{-1} \{\hat{\beta}(Y) - \beta_j\}|_{\beta=\beta_j} \sim \chi^2_p \qquad (\dagger)$$

(see Mardia *et al.*, 1979, Thm 2.5.2). This proves Thm 5.3 above. Let's celebrate this result!

Theorem 5.4. *For the LM with σ^2 specified, C defined in* (5.7) *is a family of exact confidence procedures for β, which has the LSP.*

Of course, when we plug-in for σ^2 we slightly degrade this result, but not by much if $n/(n-p) \approx 1$.

This happy outcome where we can find a family of exact confidence procedures with the LSP is more-or-less unique to the regression parameters in the LM. but it is found, approximately, in the large-n behaviour of a much wider class of models, including GLMs, as explained next.

5.3.3 *Wilks confidence procedures*

There is a beautiful theory which explains how the results from Sec. 5.3.2 generalise to a much wider class of models than the LM. The theory is quite strict, but it almost-holds over relaxations of some of its conditions. Stated informally, if $Y := (Y_1, \ldots, Y_n)$ and

$$f(y; \theta_j) = \prod_{i=1}^{n} f_1(y_i; \theta_j) \quad \text{for some } \theta \in \Omega, \qquad (5.8)$$

(see Sec. 3.1) and f_1 is a *regular model*, and the parameter space Ω is a convex subset of $\mathbb{R}^p$ (and invariant to n), then

$$-2 \log \left(\frac{f(Y; \theta_j)}{f(Y; \hat{\theta}(Y))} \right) \Bigg|_{\theta=\theta_j} \xrightarrow{\mathrm{D}} \chi^2_p \qquad (5.9)$$

where $\hat{\theta}$ is the Maximum Likelihood Estimator (MLE) of θ, and '$\xrightarrow{\mathrm{D}}$' denotes 'convergence in distribution' as n increases without bound. Eq. (5.9) is sometimes termed *Wilks's Theorem*, hence the name of this subsection.

The definition of 'regular model' is quite technical, but a working guideline is that $f_1(y_i; \theta_j)$ must be smooth and differentiable in θ_j for each y_i; in particular, supp Y_i must not depend on θ_j. Cox (2006, ch. 6) provides a summary of this result and others like it, and more details can be found in Casella and Berger (2002, ch. 10), or, for the full story, in van der Vaart (1998).

This result is true for the LM, because we showed that it is exactly true for any n provided that σ^2 is specified, and the ML plug-in for σ^2 converges on the true value as $n/(n-p) \to 1$.[11]

[11] This is a general property of the MLE, that it is *consistent* when f has the product form given in (5.8).

In general, we can use it the same way as in the LM, to derive a decision rule:

$$C(y;\alpha) := \left\{\theta_j \in \Omega : -2\log\left(\frac{f(Y;\theta_j)}{f(Y;\hat{\theta}(Y))}\right) < \chi_p^{-2}(1-\alpha)\right\}. \quad (5.10)$$

As already noted, this C satisfies the LSP. Further, under the conditions for which (5.9) is true, C is also a family of approximately exact confidence procedures.

Eq. (5.10) can be written differently, perhaps more intuitively. Define

$$L(\theta_j; y) := f(y;\theta_j)$$

known as the *likelihood function* of θ_j; sometimes the y argument is suppressed, notably when $y \leftarrow y^{\text{obs}}$. Let $\ell := \log L$, the *log-likelihood function*. Then (5.10) can be written

$$C(y;\alpha) = \left\{\theta_j \in \Omega : \ell(\theta_j; y) > \ell(\hat{\theta}(y); y) - \kappa(\alpha)\right\} \quad (5.11)$$

where $\kappa(\alpha) := \chi_p^{-2}(1-\alpha)/2$. In this procedure we keep all $\theta_j \in \Omega$ whose log-likelihood values are within $\kappa(\alpha)$ of the maximum log-likelihood. In the common case where $\Omega \subset \mathbb{R}$, (5.11) gives *'Allan's Rule of Thumb'*:[12]

- For an approximate 95% confidence procedure for a scalar parameter, keep all values of $\theta_j \in \Omega$ for which the log-likelihood is within 2 of the maximum log-likelihood.

[12] After Allan Seheult, who first taught it to me.

The value 2 is from $\chi_1^{-2}(0.95)/2 = 1.9207\ldots \approx 2$.

Bootstrap calibration. The pertinent question, as always with methods based on asymptotic properties for particular types of model, is whether the approximation is a good one. The crucial concept here is *level error*. The coverage that we want is at least $(1-\alpha)$ everywhere, which is termed the 'nominal level'. But were we to evaluate a confidence procedure such as (5.11) for a general model (not a LM) we would find that, over all $\theta_j \in \Omega$, that the minimum coverage was not $(1-\alpha)$ but something else; usually something less than $(1-\alpha)$. This is the 'actual level'. The difference is

$$\text{level error} := \text{nominal level} - \text{actual level}.$$

Level error exists because the conditions under which (5.11) provides an exact confidence procedure are not met in practice, outside the LM. Although it is tempting to ignore level error, experience suggests that it can be large, and that we should attempt to correct for level error if we can.

One method for making this correction is *bootstrap calibration*, described in DiCiccio and Efron (1996). Here are the steps, based on (5.11), although with a generic κ in place of the function $\kappa(\alpha)$:

$$C(y;\kappa) = \left\{\theta_j \in \Omega : \ell(\theta_j; y) > \ell(\hat{\theta}(y); y) - \kappa\right\}. \quad (5.12)$$

1. Compute a point estimate for θ, say $\hat{\theta}^{\text{obs}} := \hat{\theta}(y^{\text{obs}})$ the ML estimate. Other estimates are also possible, see Sec. 4.4.

2. For $i = 1, \ldots, m$:

   Sample $y^{(i)} \sim f(\cdot; \hat{\theta}^{\text{obs}})$, compute and record $\hat{\theta}^{(i)} := \hat{\theta}(y^{(i)})$, and $\hat{\ell}^{(i)} := \ell(\hat{\theta}^{(i)}; y^{(i)})$.

So, at the end of this process we have $\hat{\theta}^{\text{obs}}$ and the sample of values $\{y^{(i)}, \hat{\theta}^{(i)}, \hat{\ell}^{(i)}\}$ for $i = 1, \ldots, m$. Computing the ML estimate has to be a quick procedure because m needs to be large, say 1000s.

Now if we choose a particular value for κ, an empirical estimate of the coverage at $\theta = \hat{\theta}^{\text{obs}}$ is

$$\begin{aligned}\widehat{\text{cvg}}(\kappa) &:= \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\hat{\theta}^{\text{obs}} \in C(y^{(i)}; \kappa)\} \\ &= \frac{1}{m}\sum\nolimits_i \mathbb{1}\{\ell(\hat{\theta}^{\text{obs}}; y^{(i)}) > \hat{\ell}^{(i)} - \kappa\} \\ &= \frac{1}{m}\sum\nolimits_i \mathbb{1}\{\hat{\ell}^{(i)} - \ell(\hat{\theta}^{\text{obs}}; y^{(i)}) < \kappa\}.\end{aligned}$$

Therefore to set the empirical coverage to $(1-\alpha)$, κ needs to be the $(1-\alpha)$-quantile of the values

$$\{\hat{\ell}^{(i)} - \ell(\hat{\theta}^{\text{obs}}; y^{(i)})\}_{i=1}^{m}.$$

So the final step is to find this value, call it $\kappa^*(\alpha)$, and then compute the confidence set $C(y^{\text{obs}}; \kappa^*(\alpha))$ from (5.12).

This is a very complicated procedure, and it is hard to be precise about the reduction in level error that occurs (see DiCiccio and Efron, 1996, for more details). One thing that is definitely informative is the discrepancy between $\kappa^*(\alpha)$ and $\kappa(\alpha)$, which is an indicator of how well the asymptotic conditions hold. Put simply, if the discrepancy is small then either threshold will do. But if the discrepancy is large, then $\kappa(\alpha)$ will not do, and one is forced to use $\kappa^*(\alpha)$, or nothing. A large sample is required, for $(1-\alpha) = 0.95$: accurately estimating the 95th percentile is going to require about $m = 1000$ samples.[13]

[13] See Harrell and Davis (1982) for a simple estimator for quantiles.

5.3.4 *Summary*

With the Linear Model (LM) described in Sec. 5.3.2, we can construct a family of exact confidence procedures, with the LSP, for the parameters β. Additionally—I did not show it but it follows directly—we can do the same for all affine functions of the parameters β, including individual components.

In general we are not so fortunate. It is not that we cannot construct families of confidence procedures with the LSP: Sec. 5.3.1 shows that we can, in an uncountable number of different ways. But their levels will be conservative, and hence they are not very informative. A better alternative, which ought to work well in large-n simple models like (5.8) is to use Wilks's Theorem to construct a

family of approximately exact confidence procedures, which have the LSP, see Sec. 5.3.3.

The Wilks approximation can be checked and—one hopes—improved, using bootstrap calibration. Bootstrap calibration is a necessary precaution for small n or more complicated models (e.g. time series or spatial applications). But in these cases a Bayesian approach is likely to be a better choice, which is reflected in modern practice.

5.4 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 $\theta_j \in \Omega$. 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 5.5 (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.*[14]

[14] $gC := \left\{\phi_j : \phi_j = g(\theta_j) \text{ for some } \theta_j \in C\right\}$.

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

$$\Pr\{\theta_j \in C(Y); \theta_j\} \leq \Pr\{\phi_j \in gC(Y); \theta_j\}$$

for all $\theta_j \in \Omega$. 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.

There is an interesting consequence of this result based on the confidence procedures defined in Sec. 5.3.2 and Sec. 5.3.3. Taking the latter more general case, consider the family of approximately exact confidence procedures defined in (5.12). Let $g^{-1} \subset \Omega$ be the inverse image of g. Then

$$\begin{aligned} \phi_j \in gC(y;\alpha) &\iff \exists \theta_j : \phi_j = g(\theta_j) \wedge \theta_j \in C(y;\alpha) \\ &\iff \max_{\theta_j \in g^{-1}(\phi_j)} \ell(\theta_j; y) > \ell(\hat{\theta}(y); y) - \kappa(\alpha) \end{aligned}$$

The expression on the left of the final inequality is the *profile log-likelihood*,

$$\ell_g(\phi_j; y) := \max_{\theta_j \in g^{-1}(\phi_j)} \ell(\theta_j; y). \tag{5.13}$$

It provides a simple rule for computing a log-likelihood for any function of θ_j. Because gC is conservative, we would expect to

be able to reduce the threshold below $\kappa(\alpha)$ if g is not bijective. However, this is not an area where the asymptotic theory is very reliable (i.e. it takes a long time to 'kick in'). A better option here is to use bootstrap calibration to derive a $\kappa^*(\alpha)$ for g, as described in Sec. 5.3.3.

5.5 *P-values*

There is a general theory for P-values, also known as *significance levels*, which is outlined in Sec. 5.5.2, and critiqued in Sec. 5.5.3 and Sec. 5.5.4. But first I want to focus on P-values as used in Null Hypothesis Signficance Tests, which is a very common situation.

As discussed in Sec. 5.3, we have methods for constructing families of good confidence procedures, and the knowledge that there are also families of confidence procedures which are poor (including completely uninformative). In this section I will take it for granted that a family of good confidence procedures has been used.

5.5.1 *P-values and confidence sets*

Null Hypothesis Signficance Tests (NHST) were discussed in Sec. 4.5. In a NHST the parameter space is partitioned as

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

where typically H_0 is a very small set, maybe even a singleton. 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 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 14 (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 $C(y^{obs}; \alpha)$ 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 (Thm 5.1) we know that $C(y^{\text{obs}}; \alpha)$ is synonymous with a function $g : \mathcal{Y} \times \Omega \to \mathbb{R}$, and $C(y^{\text{obs}}; \alpha)$ does not intersect with H_0 if and only if

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

Thus the p-value is computed as

$$p(y^{\text{obs}}; H_0) := \max_{\theta_j \in H_0} g(y^{\text{obs}}, \theta_j), \tag{5.14}$$

for a specified family of confidence procedures (represented by the choice of g). Here is an interesting and suggestive result.[15] This will be the basis for the generalisation in Sec. 5.5.2.

[15] Recollect the definition of 'super-uniform' from Def. 12.

Theorem 5.6. *Under Def. 14 and* (5.14)*,* $p(Y; H_0)\big|_{\theta=\theta_j}$ *is super-uniform for every* $\theta_j \in H_0$*.*

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

$$\Pr\{p(Y; H_0) \leq u; \theta_j\} \leq \Pr\{g(Y, \theta_j) \leq u; \theta_j\} \leq u \qquad : \theta_j \in H_0$$

where the final inequality follows because $g(Y, \theta_j)\big|_{\theta=\theta_j}$ is super-uniform for all $\theta_j \in \Omega$, from Thm 5.1. □

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 NHST, 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).

5.5.2 *The general theory of P-values*

Thm 5.6 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.

Definition 15 (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* $p(Y)$ *is uniform under* $Y \sim f_0$*, then* p *is an* exact *significance procedure for* f_0*. The value* $p(y^{obs})$ *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 Sec. 5.5.1. The usual extension would be to take the maximum of the p-values over the set.[16]

[16] 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 Sec. 5.6 for the probability theory which underpins the following result.

Theorem 5.7. *Let $t : \mathcal{Y} \to R$. Define*

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

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

Proof.

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

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

$$p(Y;t) = G(-t(Y))$$

which is super-uniform under $Y \sim f_0$ according to the Probability Integral Transform (see Sec. 5.6, notably Thm 5.9). The PIT also covers the case where the distribution function of $t(Y)$ is strictly increasing on the realm of $t(Y)$, in which case $p(\cdot\,;t)$ 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.[17] So some additional criteria are required to separate out good from poor significance procedures. The most pertinent criterion is:

[17] It is a good exercise to check that $t(y) = c$ does indeed induce a super-uniform $p(\cdot\,;t)$ 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 $p(Y;t)$ 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 a useful 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).

5.5.3 Being realistic about significance procedures

Sec. 5.5.1 made the case for reporting an NHST in terms of a p-value. But what can be said about the more general use of p-values to 'score' the hypothsis $H_0 : Y \sim f_0$? 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.

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 me already know to be true, namely that H_0 is false." But unfortunately that is what it means.

When $H_0 : (Y_1, \ldots, Y_n) \overset{\text{iid}}{\sim} f_0$, the power of the test statistic is closely related to the size of the sample, n, from which we infer that the size of the P-value is decreasing in n. That is to say, we expect

$$n' > n \implies p(\boldsymbol{y}^{\text{obs}}; n') < p(\boldsymbol{y}^{\text{obs}}; n) \quad \text{and} \quad \lim_{n \uparrow \infty} p(\boldsymbol{y}^{\text{obs}}; n) = 0.$$

This suggests that the P-value is a crude measure of the size of the sample, which is how many experienced statisticians interpret it.[18] It also suggests that if there is a threshold for small P-values, it should be chosen with n in mind: the idea that a single threshold such as 0.05 would serve across a range of studies with different values for n seems very naïve. Unfortunately, it also encourages cheating; see, e.g., Masicampo and Lalande (2012).

[18] See, e.g., Andrew Gelman's blog, http://andrewgelman.com/2009/06/18/the_sample_size/.

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

5.5.4 *P-values and Generalised Likelihood Ratios*

Statisticians recognise that a P-value fails to be informative in the way that is desired because $H_0 : Y \sim f_0$ is always false. A better question to ask is: "How well does my f_0 perform compared to other candidate distributions for Y?" This is a comparative question, which does not require that any one of the candidates be true. The difficulty is that a large part of the appeal of a significance procedure is that it only requires a null model f_0 and a test statistic $t : \mathcal{Y} \to \mathbb{R}$. So this prompts the question: can we bootstrap our way to a family of candidate distributions for Y with a proper parameter space containing the null model, just from f_0 and t? And indeed we can.

This suggestion originated with David Cox in Savage *et al.* (1962, p. 84), see also Cox (1977). Let the family of distributions be

$$f(y; \theta) \propto f_0(y)\, e^{\theta \cdot t(y)} \qquad : \theta \geq 0$$

This is known as *Exponential tilting* and the result is a subset of the *Exponential family* of distributions. The constant of proportionality has to ensure that $\sum_y f(y; \theta) = 1$ for all θ in some to-be-determined parameter space Ω. It is straightforward to check that

$$f(y; \theta) = \frac{f_0(y)\, e^{\theta \cdot t(y)}}{M_T(\theta)} \qquad : M_T(\theta) < \infty \tag{5.15a}$$

where M_T is the Moment Generating Function (MGF) of $t(Y)$ under $Y \sim f_0$; see (1.10). Therefore the parameter space is defined as

$$\Omega := \{\theta \geq 0 : M_T(\theta) < \infty\}; \tag{5.15b}$$

Ω is a convex subset of $\mathbb{R}$ (see, e.g., Schervish, 1995, sec. 2.2).

How can we use this family of distributions? We could compute a 95% confidence set for θ and see whether 0 was inside it, or, better, we could compute a P-value for $H_0 : \theta = 0$. Actually, both of these are tricky because 0 is on the edge of the parameter space: we would have a problem with level error (see Sec. 5.3.3). But there is no need to follow this route. Going back to Sec. 4.6, we recollect that the ratio $f_0(y)/f_1(y)$ is the only admissible way to choose between two candidates f_0 and f_1. Furthermore, such a ratio of probabilities is directly interpretable. This suggests that we compare f_0 with the other members of the family in the same way; and since there are many of them (an uncountable number), we consider the single ratio

$$\Lambda(y; f_0) := \frac{f_0(y)}{f(y; \hat{\theta}(y))} \tag{GLR}$$

where $\hat{\theta}(y)$ is the Maximum Likelihood (ML) estimate, which gives $0 < \Lambda(y; f_0) \leq 1$. This is termed the *generalised likelihood ratio (GLR)* for f_0.[19]

[19] It is the basis of the family of Wilks approximately exact confidence procedures (see Sec. 5.3.3), but here it is being used directly, rather than being processed into a confidence set.

We can interpret the GLR as the relative support for f_0, because a small value of the ratio, say 0.05, indicates that there is another member of the family for which the observations are twenty times more probable than they are under f_0. Thus, by a smoothness argument, there are many members which are much more probable than f_0. On the other hand, a larger value, say 0.2, indicates that there is no member of the family for which the observations are more than five times more probable than they are under f_0. In some people's minds this may be enough to cast strong doubt on f_0, but I would be more cautious. If I had an *a priori* reason for selecting f_0 as a good candidate for Y, then I would want strong evidence before I gave up f_0 for another member of this family.

One unexpected but gratifying inequality relates the GLR to the original P-value:

$$\begin{aligned}
\Lambda(y; f_0) &= \frac{f_0(y)}{\max_{\theta\in\Omega} f(y;\theta)} \\
&= \min_{\theta\in\Omega} \frac{f_0(y)}{f(y;\theta)} \\
&= \min_{\theta\in\Omega} \frac{1}{e^{\theta\cdot t(y)}/M_T(\theta)} \\
&= \min_{\theta\in\Omega} e^{-\theta\cdot t(y)} M_T(\theta) \qquad (\dagger) \\
&\geq \Pr\left\{t(Y) \dot{\geq} t(y); f_0\right\} \qquad \text{Chernoff's ineq., see (1.19)} \\
&= p(y; f_0).
\end{aligned}$$

So $p(y; f_0)$ is a lower-bound for the GLR. Here then is a good reason for sticking with f_0 when the P-value is large: a large P-value rules out a small GLR. But this inequality should also make us concerned about interpreting small P-values. Markov's inequality, on which Chernoff's inequality is based, is quite relaxed, which

suggests that the GLR could be a lot higher than the P-value. So a P-value of 0.05 could correspond to a GLR of more than 0.2.

For more insight, consider a classic example, where the null model is $Z \sim \mathrm{N}(0,1)$, and $t(z) = z$; this was first analysed in Edwards *et al.* (1963).[20] In this case

$$M_T(\theta) = M_Z(\theta) = \exp(\tfrac{1}{2}\theta^2)$$

and a quick calculation from (†) gives

$$\Lambda(z; f_0) = \min_{\theta\in\Omega} \exp(-\theta \cdot z + \tfrac{1}{2}\theta^2) = \exp(-\tfrac{1}{2}z^2) \qquad : z \geq 0.$$

We also have an explicit expression for the original P-value,

$$p(z; f_0) = \Pr\left\{Z \dot{\geq} z; f_0\right\} = 1 - \Phi(z) \qquad (\ddagger)$$

where Φ is the distribution function of the standard Normal distribution. Eq. (‡) can be inverted to express z as a function of the P-value, and this allows us to plot the GLR against the P-value. The result is shown in Figure 5.1. In this case, a P-value of 0.05 corresponds to a GLR of 0.26, which is hardly a strong refutation of f_0. In my view this casts serious doubt on the many thousands (literally) of f_0's which have been rejected with P-values of a little under 0.05. A GLR of 0.05 corresponds to a P-value of about 0.007. If someone forced me to provide a threshold for a P-value below which I would reject f_0, my threshold would be 0.007.

To summarize. If you feel compelled to address the question "Is f_0 a good model for Y?" and you select the test statistic t for that purpose, then you are not restricted simply to computing the P-value. You can also compute a Generalised Likelihood Ratio (GLR) for f_0 in a family of distributions which is induced by f_0 and t. If the GLR is very small, say less than 0.05, then go ahead and reject f_0—you have already identified a whole set of much better candidates. If on the other hand the GLR is not small, then you should be cautious about rejecting f_0, if you have other reasons for favouring it. The only role for the P-value in this assessment is that it provides a lower bound on the GLR, which will allow you to skip evaluating the GLR, if the P-value is not small.

[20] Here Z is the 'z-statistic',

$$(\bar{X} - \mu_0) \Big/ \sqrt{\sigma^2/n},$$

where $\bar{X}$ is the sample mean, and σ^2 may be replaced by an estimator; see Sec. 5.3.2. This classic example is also the most frequently-used statistical model of all.

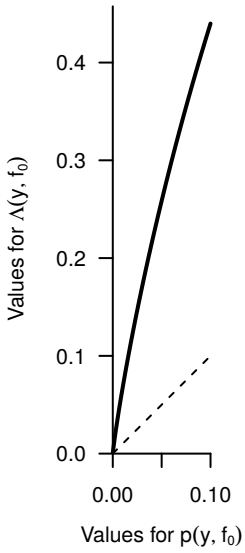

Figure 5.1: The P-value against the Generalised Likelihood Ratio (GLR) for the null model $Z \sim \mathrm{N}(0,1)$ with $t(z) = z$. The dashed line has gradient 1.

5.6 *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 $F(x) := \Pr(X \dot{\leq} x)$. 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*

$$F^-(u) := \inf\left\{x \in \mathbb{R} : F(x) \geq u\right\}. \qquad (5.16)$$

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

Theorem 5.8 (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

$$F^-(u) \leq x \iff u \leq F(x) \qquad (\dagger)$$

because then

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

as required. So stare at Figure 5.2 for a while.

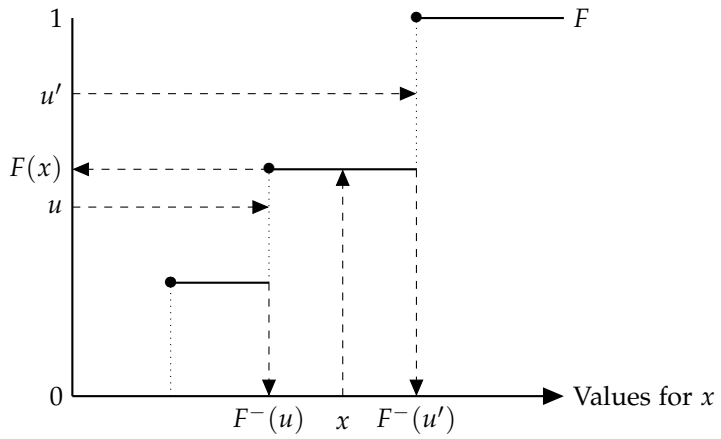

Figure 5.2: Figure for the proof of Thm 5.8. The distribution function F is non-decreasing and continuous from the right. The quantile function F^- is defined in (5.16).

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 (†). □

Thm 5.8 is the basis for the following result; recollect the definition of a super-uniform random quantity from Def. 12. This result is used in Thm 5.7.

Theorem 5.9. *If F is the distribution function of X, then $F(X)$ has a super-uniform distribution. If the support of X is convex then $F(X)$ has a uniform distribution.*

Proof. Check from Figure 5.2 that $F(F^-(u)) \geq u$. Then

$$\begin{aligned} \Pr\{F(X) \leq u\} &= \Pr\{F(F^-(U)) \leq u\} \qquad \text{from Thm 5.8} \\ &\leq \Pr\{U \leq u\} \\ &= u. \end{aligned}$$

In the case where the support of X is convex, the distribution function F is strictly increasing between lower and upper limits, in which case $F(F^-(u)) = u$ for $u \in (0,1)$. Then

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

as required. □