6
Model diagnostics and model choice

It is a complicated business to estabish whether a statistical model is appropriate, and we should beware of facile answers to complicated questions. This chapter covers model choice (Sec. 6.1), hypothesis testing (Sec. 6.2), significance levels (Sec. 6.3), and confidence sets (Sec. 6.4). Some strong opinions will be expressed.

6.1 Model choice

Consider the case where there are competing proposals for the PMF of a set of random quantities $\boldsymbol{X}$. This would usually arise where the client has several groups of experts, with incompatible beliefs. For example, the client might be the Intergovernmental Panel on Climate Change (IPCC) and the experts might be the different climate modelling groups, where each group has one or more simulators of future weather, which can be used to induce a PMF (see Rougier and Goldstein, 2014). Or else the client might be the State of California, and the experts might be different earthquake modelling groups.[1]

These are large-scale examples, but there are many smaller-scale ones as well, such as a catastrophe modelling company which requires a storm model, or an engineering consultancy which requires a fatigue model, or a pharmaceutical company which requires a model for metabolic processes, and so on. In each case a search of the literature will reveal a number of alternatives. For reasons of cost the client would like to choose a single one, but it is important to appreciate that she does not have to.

[1] See `http://www.cseptesting.org/` for an example of a large experiment to choose between different models for earthquakes.

6.1.1 Belief pooling and model averaging

Let the competing models for $\boldsymbol{X}$ be represented as the PMFs

$$\{\mathrm{p}_m\}_{m\in\mathcal{M}}.$$

The client has an application, and for this application she would like to select a subset of these models—say just one for simplicity—because it is cheaper than maintaining and using all of the models. We will analyse this decision as though she could proceed with all

of the models, in order to decide on the appropriate criterion for selecting just one.

The client always has the option to combine her models into a single 'super-model'. There are two main approaches:

$$\mathrm{p}(\boldsymbol{x}) = \sum_m \mathrm{p}_m(\boldsymbol{x}) \cdot w_m \qquad \text{linear pooling}$$
$$\mathrm{p}(\boldsymbol{x}) \propto \prod_m \mathrm{p}_m(\boldsymbol{x})^{w_m} \qquad \text{logarithmic pooling,}$$

For linear pooling it is necessary and sufficient that $\boldsymbol{w} \in \mathbb{S}^{|\mathcal{M}|-1}$, in order that $\mathrm{p}(\cdot)$ is always a PMF.[2] For logarithmic pooling it is necessary and sufficient that $w_m \geq 0$. One advantage of linear pooling over logarithmic pooling is immediately apparent: there is no need to compute a normalising constant.

[2] Recall that $\mathbb{S}^k$ denotes the k-dimensional unit simplex, see (1.2).

The reason for two approaches is that both approaches have attractive and unattractive theoretical properties.[3] Linear pooling has the attractive property that it preserves a common ordering across the probabilities of random propositions. In other words, if every model in $\mathcal{M}$ implies that $\Pr_m(P) \leq \Pr_m(Q)$, then $\Pr(P) \leq \Pr(Q)$ in the pooled model as well. Another attractive property is that $\mathrm{p}(\boldsymbol{x})$ depends only on the probabilities assigned to $\boldsymbol{x}$. A third is that the support of $\mathrm{p}(\cdot)$ is the union of the supports of the $\mathrm{p}_m(\cdot)$'s. Logarithmic pooling does not have any of these properties.

[3] More details are available in Cooke (1991, ch. 11); note the typo on p. 172, item 6, where the inequalities go the wrong way. See also the recent survey by French (2011).

On the other hand, logarithmic pooling is invariant to the order of pooling and conditioning. With linear pooling the result will typically be different if we pool first and then condition, or if we condition first and then pool—as shown below. This implies that every model in $\mathcal{M}$ might treat X_1 and X_2 as probabilistically independent, and yet the pooled model might not.[4] Logarithmic pooling has the very unattractive property that the support of $\mathrm{p}(\cdot)$ is the intersection of the supports of the $\mathrm{p}_m(\cdot)$'s; in other words it takes only one model to assert $\mathrm{p}_m(\boldsymbol{x}) = 0$ to ensure that $\mathrm{p}(\boldsymbol{x}) = 0$.

[4] Cooke (1991, p. 174) notes that this is not necessarily an attractive property if the models themselves disagree one with another about the marginal probabilities of X_1 and X_2.

Overall, linear pooling seems to have won out due to its practical simplicity, and its intuitive form when pooling first and then conditioning. The default position would be to take the weights equal, but it is useful to have the flexibility to go further. For example, two similar models could share the weight of one model, or a model that was apparently deficient (e.g. missing a process) could be down-weighted. For the climate simulators used by the IPCC, the default position of the IPCC is to give all of the models equal weight. But most climate scientists would definitely have a view about non-equal weights, reflecting all sorts of things like simulator genealogies, and the accumulated experience of the research groups.

As in Chapter 4, represent the dataset as the truth of the proposition $Q := q(\boldsymbol{X})$, where q is a first-order sentence. Now consider the

effect of conditioning the linearly pooled model:

$$\begin{aligned}
\mathrm{p}^*(\boldsymbol{x}) &:= \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q) \\
&\propto \mathbb{1}_{q(\boldsymbol{x})}\ \mathrm{p}(\boldsymbol{x}) && \text{Muddy table theorem, Thm 3.3} \\
&= \mathbb{1}_{q(\boldsymbol{x})} \sum_m \mathrm{p}_m(\boldsymbol{x}) \cdot w_m && \text{linear pooling} \\
&= \sum_m \mathbb{1}_{q(\boldsymbol{x})}\, \mathrm{p}_m(\boldsymbol{x}) \cdot w_m \\
&= \sum_m \mathrm{p}^*_m(\boldsymbol{x}) \cdot \Pr_m(Q)\, w_m
\end{aligned}$$

where

$$\mathrm{p}^*_m(\boldsymbol{x}) := \Pr_m(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q) = \frac{\mathbb{1}_{q(\boldsymbol{x})}\ \mathrm{p}_m(\boldsymbol{x})}{\Pr_m(Q)}$$

by the Muddy table theorem again. Reincorporating the normalising constant $\Pr(Q)^{-1}$ then gives

$$\mathrm{p}^*(\boldsymbol{x}) = \sum_m \mathrm{p}^*_m(\boldsymbol{x}) \cdot w^*_m$$

where

$$w^*_m := \frac{\Pr_m(Q)\, w_m}{\Pr(Q)} = \frac{\Pr_m(Q)\, w_m}{\sum_{m'} \Pr_{m'}(Q)\, w_{m'}}.$$

So conditioning the linear pooled model on Q has two parts: conditioning each model on Q, and updating the weights from $\boldsymbol{w}$ to $\boldsymbol{w}^*$. Combining multiple models into an inference in this way is termed *Bayesian Model Averaging (BMA)*; see Hoeting *et al.* (1999) for a review.

This update of the weights is the reason that linear pooling is sensitive to the order of pooling and conditioning. But the update of the weights is also one of the most attractive features of this procedure. Observe that the expression for $\boldsymbol{w}^*$ has the form analysed in Sec. 4.6. Hence we can consider the stability conditions in order to determine whether, for the dataset represented by Q, the simpler expression

$$\widetilde{w}_m := \frac{\Pr_m(Q)}{\sum_{m'} \Pr_{m'}(Q)}$$

provides a good approximation to $\boldsymbol{w}^*$. If so, the precise values of the linear combination $\boldsymbol{w}$ can be neglected (without needing to be all the same), and the weights and the conditional PMF are determined entirely by the set of models and the dataset Q.

The value $\Pr_m(Q)$ is termed the *evidence* of model m. The suitability of the evidence in updating the weights depends on $\Pr_m(Q)$ being a reasonable representation of modelling group m's beliefs about the dataset. As discussed at the end of Sec. 4.6, the statistician in group m may decide that the stability conditions hold, so that he can replace a carefully-considered prior distribution π_θ with something flat and tractable. In this case the $\boldsymbol{X}$ margin of the joint distribution $(\boldsymbol{X}, \theta)$ may *not* be a reasonable representation of group m's beliefs about $\boldsymbol{X}$, and hence $\Pr_m(Q)$ may not be a reasonable representation of group m's beliefs about the dataset. This is a serious impediment to the model choice methods discussed in the next two subsections.

6.1.2 *Model choice as a decision problem*

As shown in the previous subsection, the client could have pooled her models and derived a BMA representation of her beliefs, $\mathrm{p}^*(\cdot)$. Even if she chooses not to do this, but instead to proceed with one model alone, she still has the option to assess each of the models according to how well it matches $\mathrm{p}^*(\cdot)$ in her application. And this last point is crucial: she needs a model for a reason, and this reason will help to determine which model she should choose.

Now would be a good time to refer back to Sec. 5.1 for a review of decision theory and decision rules. The client has an action set $\mathcal{A}$ and a loss function $L(a, \boldsymbol{x})$. Her best choice of action is the one that minimises her expected loss $\mathrm{E}^*\{L(a, \boldsymbol{X})\}$.[5] Both $\mathcal{A}$ and L may be somewhat simplified compared to the client's actual application, but the idea is that they are representative, in the sense that conclusions she draws from the simplified analysis are informative about how she should proceed in her actual application.

[5] I assume, purely for simplicity, that the client's choice of action has no impact on her beliefs about $\boldsymbol{X}$.

Using her BMA representation, the client can identify her best choice of action,

$$a^* := \operatorname*{argmin}_{a \in \mathcal{A}} E^*\{L(a, \boldsymbol{X})\},$$

for which her Bayes risk is defined as $R^* := E^*\{L(a^*, \boldsymbol{X})\}$. She can also identify her best choice of action using each model on its own,

$$a_m := \operatorname*{argmin}_{a \in \mathcal{A}} E^*_m\{L(a, \boldsymbol{X})\} \quad m \in \mathcal{M}.$$

When she uses the action optimal for model m in place of her optimal action a^* her risk is

$$R(m) := \mathrm{E}^*\{L(a_m, \boldsymbol{X})\}$$

where, necessarily, $R^* \leq R(m)$. The different $R(m) - R^*$ is the expected additional loss she incurs from using just model m. The optimal single model is therefore

$$\widetilde{m} := \operatorname*{argmin}_{m \in \mathcal{M}} R(m).$$

This analysis makes it clear that the choice of a single model from $\mathcal{M}$ depends on the dataset, but also on the action set and loss function; i.e. on the client's application.

To illustrate, consider two different applications. In one, the loss is approximately linear in $\boldsymbol{X}$ for each action; for example, the action is the amount of advertising spending and the loss is the negative revenue. In this case, a good model has $\mathrm{E}^*_m(\boldsymbol{X}) \approx \mathrm{E}^*(\boldsymbol{X})$. In another application, the loss is highly non-linear in $\boldsymbol{X}$ for some actions. This is typically the case for hazards, such as flooding; for example, the action is the height of the flood defenses, and the loss is the amount of land inundated. In this case, a good model has $\mathrm{p}^*_m \approx \mathrm{p}^*$ for the extreme values of $\boldsymbol{X}$.

There are just two conditions in which the choice of model does not depend sensitively on the action set and loss function. If the

relative likelihood is almost entirely concentrated on one model, the maximum likelihood model $\hat{m}$, then the Stable estimation theorem (Thm 4.3) implies that

$$w_m^* \approx \begin{cases} 1 & m = \hat{m} \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $\mathrm{p}^*_{\hat{m}} \approx \mathrm{p}^*$, and consequently $a_{\hat{m}} \approx a^*$ no matter what the action set and loss function. Hence there is no additional expected loss from replacing $\mathcal{M}$ with the single model $\hat{m}$. Second, if the action set is really small—perhaps having only two elements—then there is a chance that the maximum likelihood action will be the same as the optimal action; but this is not something one could just assume. Except under these conditions, though, the action set and loss function *should* play a role in selecting a single model from a set of competing models.

6.1.3 *Where does this going wrong?*

There is a huge amount of confusion in statistical model choice, and many competing criteria. This reflects an unwillingness on the part of the statistician and/or the client to think explicitly about the underlying application, which I have represented above as an action set and a loss function. It goes without saying that the client is choosing between models for a reason; we should not be surprised that neglecting the reason leads to disarray.

If we strike out both the action set and the loss function, then the main thing left to judge the model on is $\Pr_m(Q)$, the 'evidence' of model m. As outlined at the end of the previous subsection, if the evidence of one of the models is a lot larger than the sum of the evidences of the other models, and if the action set is small, then selecting the model with the largest evidence is a defensible choice. Otherwise, further checks are required.

At the very least, the statistician would need to check that the model with the largest evidence was a member of a subset of the models which together made up most of the relative likelihood, and for which p^*_m was similar across the subset. Checking p^*_m is important, because different models will connect Q and $\boldsymbol{X}$ in different ways. 'Similar' is tough to quantify, because what is really needed is an assessment in terms of the action set and the loss function. But one could imagine a metric on PMFs for $\boldsymbol{X}$ which tried to reflect aspects of the PMF of $\boldsymbol{X}$ which are important in the client's application. In some cases this could be as simple as checking the each of the p^*_m's had similar expectations (linear loss function) or expectations and variances (quadratic or convex loss function, see Sec. 5.3).

But at this point another difficulty looms. In a parametric approach, $\mathrm{p}_m(\boldsymbol{x})$ is constructed as the $\boldsymbol{X}$ margin of the joint PMF

$$\Pr_m(\boldsymbol{X} \doteq \boldsymbol{x}, \theta_m \doteq t) = \mathrm{p}_m(\boldsymbol{x}; t)\, \pi_m(t) \qquad t \in \Omega_m.$$

There are two difficulties here. First, as already discussed, the Bayesian statistician in group m may be happy (keen!) to replace his considered prior distribution π_m with a flatter more tractable alternative, on the grounds that the stability conditions of Sec. 4.6 hold. So $Pr_m(Q)$ is not representative of his group's beliefs about Q. Second, the Frequentist statistician in group m' may not want to supply a $\pi_{m'}$ at all. So we may have to do model choice without the evidences.

This is where *Information Criteria* come in. There are lots of different ones, originating with the Akaike Information Criterion (AIC). All Information Criteria have the same form, comprising a goodness-of-fit term for the dataset and a penalty for model complexity. A penalty is required, because more complex models will tend to fit the same dataset better than simpler ones; e.g. a quadratic will fit a time-series better than a straight line. The penalty guards against *over-fitting*, in which a good fit within the dataset can lead to bad fits outside it. Information Criteria are presented in a negative orientation, so that smaller is better.

For example, the Bayes Information Criterion (BIC) under the simple observation model (SOM, Sec. 2.4) is

$$BIC_m(\boldsymbol{y}) := -2\log \mathrm{p}_m\left(\boldsymbol{y}; \hat{\theta}_m(\boldsymbol{y})\right) + k_m \log n$$

where $\hat{\theta}_m$ is the Maximum Likelihood Estimator of model m, $k_m := \dim \Omega_m$, and n is the number of observations. The first term measures goodness-of-fit by the maximum of the log-likelihood, and the second term penalises the number of model parameters by the log of the number of observations. The difference in the BICs of two models is a first-order approximation to the difference in the log-evidences; see Kass and Raftery (1995, sec. 4). This may be preferred to the actual difference in the log-evidences if the actual prior distributions are too flat, but in this case the BIC is being preferred because it is *not* a good approximation, which is a delicate argument.

Information Criteria are discussed in Gelman *et al.* (2014, ch. 7). The most popular one at the moment is the DIC (Spiegelhalter *et al.*, 2002).[6] This is for a couple of reasons:

1. It is easy to compute on the back of a Monte Carlo simulation from the posterior distribution of the model parameters; indeed DIC is built-in to software tools such as BUGS (Lunn *et al.*, 2013, sec. 8.6).

2. It has a sophisticated assessment of model complexity which goes beyond simply counting the parameters. This is important for hierarchical models in which exchangeability modelling creates large numbers of highly dependent parameters (to be discussed later).

As I hope I have made clear, I doubt that Information Criteria are appropriate for helping a client to choose a model in an important

[6] Be sure to read the discussion and rejoinder of this paper. Then read Spiegelhalter *et al.* (2014).

application. But by all means use them if you—the statistician—are your own client, and you are working on something not important enough to devote much thought to.

* * *

The other approach to mention is cross-validation. This was already discussed in Sec. 5.5. Under various simplifying conditions, models can be usefully compared in terms of their out-of-sample prediction performance. The main simplifying conditions are that the SOM hold (Sec. 2.4) holds, and that there is no overt sample bias. See Gelman *et al.* (2014, ch. 7) for a discussion of cross validation in the context of model choice.

6.2 *Hypothesis testing*

Hypothesis testing is a special case of model choice, in which all of the models for $\boldsymbol{X}$ are contained within the same parametric family. Usually the choice is between two distinct subsets of the parameter space. Thus we start with the family of distributions

$$\mathrm{p}(\cdot;\theta) \quad \text{for some } \theta \in \Omega,$$

where Ω is some convex subset of $\mathbb{R}^p$, and then specify the competing subsets as

$$H_0 : \theta \in \Omega_0$$
$$H_1 : \theta \in \Omega_1$$

where $\Omega_0 \cap \Omega_1 = \emptyset$ and, ideally, Ω_0 and Ω_1 are well-separated. H_0 and H_1 are termed *hypotheses*. If a hypothesis contains only a single element of Ω it is termed a *simple hypothesis*, otherwise it is a *composite hypothesis*.

For hypothesis testing, it is necessary to marginalise the PMF of $\boldsymbol{X}$ to provide a PMF for the observables $\boldsymbol{Y}$; usually this marginalisation operation is only feasible under the SOM (Sec. 2.4). From now on 'p' represents the marginal PMF of $\boldsymbol{Y}$; hence $\mathrm{p}(\boldsymbol{y};\theta)$ is the probability of the random proposition $\boldsymbol{Y} \doteq \boldsymbol{y}$ under the PMF with index θ. The objective of hypothesis testing is to choose in favour of H_0 or H_1 using p and $\boldsymbol{y}$, and to quantify the strength of evidence on which that choice is based.

The most studied hypothesis test, and the one for which the strongest theoretical results are available, is the case of two simple hypotheses, written

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1.$$

In this case there is agreement among statisticians that the appropriate test is based on the *likelihood ratio*,

$$B_{01}(\boldsymbol{y}) := \frac{\mathrm{p}(\boldsymbol{y};\theta_0)}{\mathrm{p}(\boldsymbol{y};\theta_1)},$$

and the test must have the form

$$B_{01}(\boldsymbol{y}) \begin{cases} < k_1 & \text{choose } H_1 \\ \text{in the middle} & \text{undecided} \\ > k_2 & \text{choose } H_0 \end{cases} \tag{6.1}$$

for some values $0 < k_1 \leq k_2$. If the client dislikes 'undecided' then k_1 and k_2 will need to be close together, maybe even the same value. But the client should understand that 'undecided' is a perfectly acceptable category for evidential support, and that suppressing it can lead to wrong choices.[7]

[7] In Scots law, for example, the judge or jury can return a verdict of 'not proven', lying somewhere between 'proven' and 'not guilty'.

Appropriate values for k_1 and k_2 are tricky to decide—they ought to depend on the consequences of each choice, but in hypothesis testing we are discouraged from taking explicit account of the client's action set and loss function.

For the Bayesian, the following equality is a direct consequence of Bayes's theorem in odds form (after Thm 3.11):

$$\frac{\Pr^*(\theta \doteq \theta_0)}{\Pr^*(\theta \doteq \theta_1)} = B_{01}(\boldsymbol{y}) \, \frac{\Pr(\theta \doteq \theta_0)}{\Pr(\theta \doteq \theta_1)} \tag{6.2}$$

where $\Pr^*$ is the probability conditional on $\boldsymbol{Y} \doteq \boldsymbol{y}$, as usual. This equality is also expressed as the mantra

$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}$$

where 'odds' is a generic term for a ratio of probabilities. So the Bayesian concerned with selecting the hypothesis with the highest posterior probability needs only to be satisfied that the likelihood ratio outweighs the prior odds. Unless he starts with strong beliefs about θ, the Bayesian will find a likelihood ratio of 20 or greater fairly compelling evidence for favouring H_0. Jeffreys (1961, Appendix B) termed this 'strong' evidence in favour of H_0.[8] So the Bayesian might well have $k_1 \leftarrow 1/20$ and $k_2 \leftarrow 20$, or something similar.

[8] See also the scale given in Kass and Raftery (1995).

Things are more complicated for the Frequentist, because he will not admit a prior PMF for the parameter, and so (6.2) is a non-starter. The traditional approach to this problem is the *Neyman-Pearson* approach, in which the sampling distributions of the random quantity $B_{01}(\boldsymbol{Y})$ under H_0 and H_1 are used to set k_1, with $k_2 \leftarrow k_1$. I am not going to describe the Neyman-Pearson approach, because I believe it is obsolete.[9] Instead, I will give '*Barnard's rule*'. I base this on a suggestion of George Barnard, in the discussion of Lindley (2000). It is necessary to suspend our disbelief and to assume that exactly one of H_0 or H_1 is 'true'.[10]

[9] Casella and Berger (2002, ch. 8) and then Lehmann and Romano (2005, ch. 3) are good references.

[10] The same suspension of disbelief is necessary in the Neyman-Pearson approach. Nowhere else in these notes do we need to make this totally bogus assertion. A model is a human construct, and its parameters are as artificial as plastic flowers. Only in very rare situations can $\theta = \theta_0$ be a statement about operationally defined quantities in the real world. I am compelled to put 'true' in scare quotes.

Theorem 6.1 (Barnard's rule)**.** *Suppose that exactly one of H_0 or H_1 is 'true'. Define an incorrect choice as choosing H_0 when H_1 is 'true', or choosing H_1 when H_0 is 'true'. Then the probability of an incorrect choice when using* (6.1) *with $k_1 \leftarrow 1/c$ and $k_2 \leftarrow c$ is never more than $1/c$.*

Proof. This is a double application of Markov's inequality (Thm 1.5). Suppose that H_0 is 'true'. Then, in an obvious notation,

$$\Pr_0\{\text{incorrect}\} = \Pr_0\{B_{01}(\boldsymbol{Y}) \leq 1/c\} = \Pr_0\{1/B_{01}(\boldsymbol{Y}) \geq c\} \leq 1/c$$

by Markov's inequality, because $1/B_{01}(\boldsymbol{Y})$ has expectation 1 under H_0 (prove this using the FTP, Thm 1.6). On the other hand, suppose that H_1 is 'true'. Then

$$\Pr_1\{\text{incorrect}\} = \Pr_1\{B_{01}(\boldsymbol{Y}) \geq c\} \leq 1/c$$

by the same reasoning. Since the probability of an incorrect choice is $\leq 1/c$ under both H_0 and H_1, it must be never more than $1/c$, because of the condition that exactly one of H_0 or H_1 is 'true'. □

With Barnard's rule, the client states, "I don't like being wrong, so I am going to set $c \leftarrow 20$ whenever I have to make a choice between two mutually exclusive simple hypotheses." This implies that the client will not be wrong more than 5% of the time.[11] Because Markov's inequality is generous (i.e. seldom tight) the client can be fairly sure that her actual probability is a lot less than 0.05, and so she might be happier with a more relaxed value for c, say $c \leftarrow 10$. In many applications, she will find that the observations leave her undecided: that's just the way it is—sometimes the evidence in $\boldsymbol{y}$ is not very compelling.

[11] It is not obvious that the client wants to control her lifetime probability of being wrong, rather than her probability of being wrong in this particular application.

Composite hypotheses. Composite hypotheses provide no additional challenges for the Bayesian, who simply sums over subsets of the posterior distribution to compute $\Pr^*(\theta \in \Omega_0)/\Pr^*(\theta \in \Omega_1)$.

Composite hypotheses are a major challenge for the Frequentist, except for one special case which is massively overused in both teaching and research (the Normal distribution). The general theory is a veritable blizzard of 'adhockery'.[12] You can take my word for this, or you can read chapters 3–10 of Lehmann and Romano (2005, pp. 56–415) and decide for yourself.

[12] 'Adhockery' has gone mainstream, but I believe the word was coined by I.J. Good, a famous statistician and Bletchley Park code-breaker.

6.3 Significance levels (P-values)

Significance levels were the invention of the great statistician R.A. Fisher. Savage (1976) and Efron (1998) give fascinating commentaries on Fisher's statistical work. Fisher was also a founder of the modern theory of evolution, and nominated by Richard Dawkins as the greatest biologist since Darwin.[13]

[13] http://edge.org/conversation/who-is-the-greatest-biologist-of-all-time

6.3.1 Motivation and definition

The distinguishing feature of a significance level is the absence of an explicit alternative hypothesis. In other words, a significance level attaches a score to

$$H_0 : \theta \in \Omega_0$$

directly. Initially, consider simple hypotheses of the form $\Omega_0 \leftarrow \{\theta_0\}$; composite hypotheses will be covered in Sec. 6.3.4. Thus H_0 corresponds to

$$H_0 : \boldsymbol{Y} \sim \mathrm{p}(\cdot; \theta_0)$$

which I write as $\boldsymbol{Y} \sim \mathrm{p}_0$, where p_0 is the *null distribution*. This type of H_0 simply describes a PMF for $\boldsymbol{Y}$. There is no particular reason for it to be one member of a parametric family: it could equally well be the $\boldsymbol{Y}$-margin of a joint PMF over random quantities and parameters, although this is less common in practice (see Box, 1980, and the discussion and rejoinder).

There is some interesting theory about how to score an observation $\boldsymbol{y}$ with respect to a distribution p_0. Any score can be written as $s(\mathrm{p}_0, \boldsymbol{y})$, which I will take in the positive orientation, so that larger scores indicate a better match. A sensible constraint on scoring rules is that they are *proper*:

$$\mathrm{E}_0\{s(\mathrm{p}_0, \boldsymbol{Y})\} \geq \mathrm{E}_0\{s(\mathrm{p}', \boldsymbol{Y})\} \qquad \text{for all } \mathrm{p}',$$

where E_0 is the expectation with respect to p_0.[14] Among the proper scoring rules, perhaps the simplest is the *logarithmic scoring rule*

$$s(\mathrm{p}_0, \boldsymbol{y}) \leftarrow \log \mathrm{p}_0(\boldsymbol{y}).$$

[14] See Gneiting and Raftery (2007) for more details about proper scoring rules.

It is easy to prove that this scoring rule is proper, using *Gibbs's inequality*.[15] Bernardo and Smith (2000, sec. 2.7) make a case for favouring the logarithmic scoring rule, on the basis that it uniquely satisfies properties of smoothness and locality.

[15] Which states that $\sum_i p_i \log(p_i/q_i) \geq 0$ for probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, with equality if and only if $\boldsymbol{p} = \boldsymbol{q}$. Follows immediately from $\log(x) \leq x - 1$ with equality if and only if $x = 1$. And this latter result follows from $\log(1) = 0$ and $\log(\cdot)$ strictly concave.

Proper scoring rules are useful for comparing two PMFs for $\boldsymbol{y}$. If

$$s(\mathrm{p}_0, \boldsymbol{y}) - s(\mathrm{p}_1, \boldsymbol{y}) > 0$$

then the evidence in $\boldsymbol{y}$ favours p_0 over the alternative p_1. This conclusion would be much less compelling if the scoring rule was not proper, because we would lose the property that the score of p_0 ought to be higher (except for random variation) than that of p_1 were p_0 to be true. But the score $s(\mathrm{p}_0, \boldsymbol{y})$ on its own is much harder to interpret. Is a value such as $s(\mathrm{p}_0, \boldsymbol{y}) = -13.863$ large or small? Even with a logarithmic scoring rule, finding that

$$\log \mathrm{p}_0(\boldsymbol{y}) \approx -13.863$$

is not very helpful. We infer that $\mathrm{p}_0(\boldsymbol{y}) \approx 0.000000954$, but this small value might just reflect a huge $\mathcal{Y}$, in which any value for $\boldsymbol{y}$ has only a small probability of occurring. For example, I toss a coin 20 times and get

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

The probability of this outcome under the hypothesis

$$H_0 : \text{the coin tosses are independent and fair}$$

is $2^{-20} = 0.000000954$, but clearly the smallness of this value on its own cannot convince me that H_0 is false, since every outcome has the same probability under H_0. In fact, in this case I could easily be convinced, both before and after tossing the coin, that H_0 is true. So the smallness of $\mathrm{p}_0(\boldsymbol{y})$ can never, on its own, count as evidence against H_0.[16]

[16] This is something that Fisher got wrong. See, for example, Hacking's critique (Hacking, 1965, p. 80) of Fisher's exposition of significance levels in his final book (Fisher, 1956).

This is the basic problem that the significance level seeks to address: to construct a score $s(\mathrm{p}_0, \boldsymbol{y})$ which occupies a meaningful scale, so that we can identify small values which cause us to question the model p_0 as an appropriate representation for $\boldsymbol{Y}$. The P-value is the result.[17] In the following definition, the scalar random quantity X has a *subuniform distribution* exactly when

$$\Pr(X \dot{\leq} u) \leq u \quad \text{for all} \quad 0 \leq u \leq 1.$$

The uniform distribution is a special case of the subuniform distribution, with $\Pr(X \dot{\leq} u) = u$. The word '*statistic*' is used in its technical sense of a specified function of the observables.

[17] Here I am presenting a modern definition of a P-value, not the one advanced by Fisher.

Definition 6.1 (P-value)**.** *The statistic $p_0 : \mathcal{Y} \to [0, 1]$ is a P-value for the simple hypothesis H_0 exactly when $p_0(\boldsymbol{Y})$ has a subuniform distribution under H_0.*

In this definition, p_0 appears to be a function of $\boldsymbol{y}$ alone, but the construction of p_0 must involve p_0 as well, in order to ensure that the subuniformity property holds: hence the inclusion of the '0' subscript. A subuniform distribution is a weaker condition that uniform distribution, but without it there would not be P-values for random quantities with finite or countably-infinite realms (Sec. 6.3.3), P-values computed by sampling (also Sec. 6.3.3), or P-values for composite hypotheses (Sec. 6.3.4).

6.3.2 *Difficulties with interpretation*

The basic idea with a P-value is that a value of $p_0(\boldsymbol{y})$ close to zero indicates an event in the lefthand tail of the distribution that would be implied by the truth of H_0. For example, since

$$\Pr_0\{p_0(\boldsymbol{Y}) \dot{\leq} 0.005\} \leq 0.005$$

we conclude that the outcome $p_0(\boldsymbol{y}) = 0.005$ is no larger than the 0.5th percentile of the distribution of $p_0(\boldsymbol{Y})$ under H_0. An outcome this far into the tail of a distribution is unusual, and leads us to consider whether in fact H_0 is 'true', or even adequate. But this apparently simply story is full of subtlety.

Sub-uniformity. The subuniformity of P-values limits the conclusions that we can draw. Suppose that $p_0(\boldsymbol{Y})$ were uniform under H_0, rather than subuniform. In this case if $p_0(\boldsymbol{y}) = 0.005$ we could conclude that we were in the tail of the distribution of $p_0(\boldsymbol{Y})$ under H_0, while if $p_0(\boldsymbol{y}) = 0.35$ we could conclude that we were near the middle of the distribution. But with subuniformity we can no longer conclude the latter, because $\Pr_0\{p_0(\boldsymbol{Y}) \dot{\leq} 0.35\}$ is no longer equal to 0.35, but only no larger than 0.35. So subuniformity prevents us from interpreting middling P-values as indicating we are near the centre of the distribution of $p_0(\boldsymbol{Y})$ under H_0. In fact, with $p_0(\boldsymbol{y}) = 0.35$ we may actually be in the lefthand tail, but not know it. Similarly, $p_0(\boldsymbol{y}) = 0.09$ looks a bit improbable under H_0, being in

the lefthand tail, but with subuniformity we do not know whether we are a little into the lefthand tail, or way into the lefthand tail.

So wherever possible, we construct P-values which are uniform or nearly uniform under H_0, rather than subuniform. Sec. 6.3.4 shows that a necessary (but not sufficient) condition is that the realm of $\boldsymbol{Y}$ is large; ideally uncountably infinite.

The 'truth' of H_0. Computing P-values is, from the outset, a forlorn exercise, because we know that p_0 is not the 'true' PMF for $\boldsymbol{Y}$. It is a representation of my beliefs about $\boldsymbol{Y}$. The randomness I perceive in $\boldsymbol{Y}$ is a symptom of my lack of knowledge. So I should respond to a small P-value without any surprise: what did I expect, that nature herself would choose $\boldsymbol{y}$ according to *my* PMF?! Likewise, even in the uniform case a large P-value does not indicate that H_0 is 'true': it simply shows that $\boldsymbol{y}$ is not very informative, since it has failed to alert me to a claim which I know to be false.

It is tempting to try to finesse this difficulty by focusing on adequacy rather than 'truth'. But this does not work. If n, the number of observations is small, then the P-value will be large because $\boldsymbol{Y}$ is not very informative. But if n is large, then the P-value will be small because H_0 is not 'true'. So one might argue that in the middle there is a 'sweet spot' for n for which the P-value is informative about the adequacy of H_0 as a model for $\boldsymbol{Y}$. But there is no logic for this claim. If something is not useful for small n or for large n, there is no basis to claim that it will nevertheless be useful for middling n.[18] We should call this the *no sweet spot* argument.

[18] For a similar observation ("the P-value can be viewed as a crude measure of sample size"), see this blog entry, `http://andrewgelman.com/2009/06/18/the_sample_size/`, and the following comments.

Many P-values. For any H_0, there is an infinity of P-values (see Sec. 6.3.3). In fact, one can find a P-value to take any value in $(0, 1)$, for any $\boldsymbol{y}$. So if you have a P-value $p_0(\boldsymbol{y}) = 0.005$, which you think is quite interesting, I can counter with a $p_0'(\boldsymbol{y}) = 0.35$, which is not very interesting at all. Who is right?

Here is a recipe for a completely meaningless P-value. Use $\boldsymbol{y}$ to seed a uniform random number generator $u_1, u_2, \ldots$, and let $p_0(\boldsymbol{y})$ be the value of the trillionth term in the sequence. By any reasonable criterion this value has a uniform distribution, and hence is a valid P-value according to Def. 6.1.

DeGroot (1973) identified an additional condition which was necessary for p_0 to be a sensible P-value: $p_0(\boldsymbol{Y})$ under H_0 has to *stochastically dominate* $p_0(\boldsymbol{Y})$ under a decision-relevant alternative to H_0. A random quantity X stochastically dominates Y exactly when

$$\begin{aligned} &\Pr(X \leq v) \leq \Pr(Y \leq v) \text{ for all } v, \\ \text{and } &\Pr(X \leq v) < \Pr(Y \leq v) \text{ for some } v. \end{aligned}$$

The stochastic dominance property implies that the distribution of $p_0(\boldsymbol{Y})$ is pushed to the left under a decision-relevant alternative to H_0, and in this sense small P-values favour the alternative over H_0.

The stochastic dominance condition eliminates P-values like the uniform random number generator, because its distribution is the

same under all models. But the condition reintroduces a competitor model through the back door—the decision-relevant alternative to H_0—and thus compromises the 'purity' of the P-value as an assessment of H_0 alone. If users of P-values want to demonstrate that their particular choice of p_0 is an appropriate one, then they need to establish that it has both the subuniformity property for H_0 *and* the stochastic dominance property for some alternative to H_0. But, having produced an alternative, these users might as well be doing a hypothesis test.

Not a hypothesis test. Many people confuse P-values and hypothesis tests. A typical symptom is to compute a $p_0(\boldsymbol{y})$ less than 0.05 and then report that "the null hypothesis is rejected at a Type 1 error of 5%." The characteristic of a hypothesis test is that two hypotheses compete, and *both* of them get tested by the observations $\boldsymbol{y}$. The result is a test statistic $B_{01}(\boldsymbol{y})$ which, to the Bayesian at least, is directly meaningful, and which is used by statisticians of all tribes to construct a rule for choosing between H_0 and H_1, or remaining undecided. This does not happen with a P-value, and there is no sense in using a P-value to "reject" H_0 if it cannot be demonstrated that an alternative H_1 is better. See Goodman (1999a,b).

Not a proper scoring rule either. The function p_0 takes both p_0 and $\boldsymbol{y}$ as its arguments (as will be seen explicitly in Sec. 6.3.3), and therefore it is a scoring rule. But it is not a proper scoring rule (see Sec. 6.3.1), and therefore comparisons between P-values of different models is not a good way to choose between models. See Schervish (1996).

* * *

There is a huge literature on why P-values do not do what people would like them to do; start at Greenland and Poole (2013) and work backwards. There is also, unfortunately, quite a lot of evidence that it is easy to cheat with P-values, and that people do cheat with P-values; see, for example, Simmons *et al.* (2011) and Masicampo and Lalande (2012).[19] See the discussion on level error at the end of Sec. 6.4 for further comments.

[19] Simmons *et al.* coin the coy euphemism 'researcher degrees of freedom' to describe 'flexibility in data collection, analysis, and reporting'; i.e. ways that researchers can get their P-values lower without 'cheating'. This practice is prevalent enough to have acquired the unsavoury name of 'P-hacking'.

6.3.3 *Constructing and computing P-values*

I stated above that there is an infinity of P-values for any H_0. This subsection presents a recipe for making them. But first, a very useful general result.

Theorem 6.2 (Probability Integral Transform, PIT)**.**

Let $X \in \mathcal{X} \subset \mathbb{R}$ be a scalar random quantity with distribution function $F_X(x) := \Pr(X \leq x)$, and let $Y := F_X(X)$. Then Y has a sub-uniform distribution, *and $F_Y(u) = u$ 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 $x \in \mathcal{X}$:

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

The 'cancellation' of F at the second equality occurs because of the bijective relationship between x and $F(x)$ for $x \in \mathcal{X}$.[20] This proves the second part of the claim.

[20] 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 6.1. If X is a 'continuous' random quantity then the steps will remain infinitesimally close to the 45° line, and $F_X(X)$ will be uniform. Otherwise, and this includes random quantities with countably infinite support like the Poisson in Figure 6.1, the steps can diverge substantially from the 45° line and $F_X(X)$ can be severely subuniform.

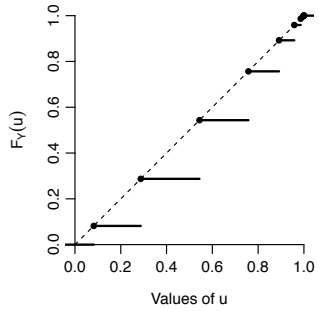

Figure 6.1: Distribution function of $Y := F_X(X)$, where $X \sim \text{Poisson}(\lambda = 2.5)$.

Now here is the recipe for making a P-value.

Theorem 6.3. *Let $t : \mathcal{Y} \to \mathbb{R}$ be a statistic. Then*

$$p_0(\boldsymbol{y}) := \Pr_0\{t(\boldsymbol{Y}) \geq t(\boldsymbol{y})\}$$

is a P-value satisfying Def. 6.1.

Proof. I use a nifty trick from Casella and Berger (2002, section 8.3.4). Define $T := t(\boldsymbol{Y})$. Let G_0 be the distribution function of $-T$ under H_0. Then

$$p_0(\boldsymbol{y}) = \Pr_0\{T \geq t(\boldsymbol{y})\} = \Pr_0\{-T \leq -t(\boldsymbol{y})\} = G_0(-t(\boldsymbol{y})).$$

Then since $p_0(\boldsymbol{Y}) = G_0(-T)$, subuniformity of $p_0(\boldsymbol{Y})$ under H_0 follows from the PIT (Thm 6.2). □

Hence there is a P-value for every test statistic, and there is an infinity of test statistics. Here is another dodgy P-value: $t(\boldsymbol{y}) = c$ (any constant will do). This does indeed have a subuniform distribution under H_0, with

$$\Pr_0\{p_0(\boldsymbol{Y}) \dot{\leq} 1\} = 1 \quad \text{and} \quad \Pr_0\{p_0(\boldsymbol{Y}) \dot{\leq} u\} = 0 \text{ for } u < 1.$$

What a useless P-value! This makes the point that of the infinity of possible P-values for H_0, many of them will be useless, or nearly so. Clearly, T needs to have a large support under H_0, in order that $p_0(\boldsymbol{Y})$ is even approximately uniform under H_0. But recollect Figure 6.1, which showed that a countably infinite support was not big enough.

* * *

Occasionally it will be possible to choose a test statistic $t(\cdot)$ with a known distribution under H_0, from which an explicit p_0 can be derived.[21] But this puts the cart before the horse—we want to choose our test statistic to reflect our application; in particular, we would like the resulting P-value to satisfy the stochastic dominance property discussed in Sec. 6.3.2. Happily, a P-value for any $t(\cdot)$ can be computed by simulation using following result, which uses exchangeability (Chapter 7).

[21] Asymptotic results are useful here; see Cox (2006, ch. 6). These give approximate P-values, in which the distribution of $p_0(\boldsymbol{Y})$ is approximately uniform under H_0. There is a level error problem with these P-values, just as in confidence sets; see Sec. 6.4.

Theorem 6.4. *For any finite sequence of scalar random quantities $X^0, X^1, \ldots, X^m$, define the rank of X^0 in the sequence as*

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

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

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

$$\Pr(R=r) = \frac{1}{m+1} \quad \text{for } r = 0, 1, \ldots, m \tag{†}$$

and zero otherwise, proving the first claim.

To prove the second claim,[22]

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

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

as required. □

Now take a statistic $t : \mathcal{Y} \to \mathbb{R}$ which has the property that larger values of $t(\boldsymbol{y})$ are suggestive of a decision-relevant alternative from H_0. Define $T := t(\boldsymbol{Y})$ and $T^j := t(\boldsymbol{Y}^j)$ where $\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^m \overset{\text{iid}}{\sim} \mathrm{p}_0$. Then $T, T^1, \ldots, T^m$ form an exchangeable sequence under H_0. Hence if

$$R(\boldsymbol{y}) := \sum_{j=1}^{m} \mathbb{1}_{-T^j \succeq -t(\boldsymbol{y})} = \sum_{j=1}^{m} \mathbb{1}_{T^j \succeq t(\boldsymbol{y})}$$

then Thm 6.4 implies that

$$P(\boldsymbol{y}) := \frac{R(\boldsymbol{y})+1}{m+1}$$

has a subuniform distribution under H_0.[23] Furthermore, the Weak Law of Large Numbers (see, e.g. Grimmett and Stirzaker, 2001, sec. 5.10) shows that

$$\lim_{m\to\infty} P(\boldsymbol{y}) = \frac{\lim_m m^{-1}\big(R(\boldsymbol{y})+1\big)}{\lim_m m^{-1}(m+1)} = \mathrm{E}_0\{\mathbb{1}_{T\dot{\geq}t(\boldsymbol{y})}\} = \Pr_0\{T \dot{\geq} t(\boldsymbol{y})\}$$

and so the asymptotic limit of $P(\boldsymbol{y})$ is the P-value defined in Thm 6.3.

$P(\boldsymbol{y})$ is subuniform for all m, but it is approximately uniform for large m, because in this case

$$\frac{\lfloor u(m+1)\rfloor}{m+1} \approx u.$$

So a bigger m is preferred, because a more uniform distribution under H_0 is more informative, as discussed in Sec. 6.3.2.

In cases where it is not straightforward to simulate independent realisations from p_0, the value of $P(\boldsymbol{y})$ can be computed from an MCMC sequence from p_0. In order for the $\boldsymbol{Y}^j$'s to be exchangeable it is sufficient that they are independent, and hence these m values must be extracted from well-separated locations in the sequence. Besag and Clifford (1989) described an elegant backwards-and-forwards implementation for MCMC sampling from p_0 which produces exchangeable but not IID $\boldsymbol{Y}$'s under H_0.

6.3.4 *Composite hypotheses*

The definition in Def. 6.1 is for a simple null hypothesis. It can be extended to a composite hypothesis, written

$$H_0 : \theta \in \Omega_0$$

where the previous simple hypothesis was just the special case $\Omega_0 = \{\theta_0\}$. Composite hypotheses are common—more common than simple ones in fact. Nuisance parameters were mentioned in Sec. 4.7. Where interest is in a subset of the parameters, the other nuisance parameters remain unconstrained, and the hypothesis is composite. For example, if $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++}$, then

$$H_0 : \mu = \mu_0$$

is a composite hypothesis with $\Omega_0 = \{\mu_0\} \times \mathbb{R}_{++}$.

With a composite hypothesis, $p_0(\cdot)$ is a P-value for H_0 if it has a subuniform distribution under every element of Ω_0. This is easy to achieve, given a set of P-values for simple hypotheses.

Theorem 6.5. *Let $H_0 : \theta \in \Omega_0$ be a composite hypothesis, and $p(\cdot; t)$ be a P-value for the simple hypothesis $\theta = t$. Then*

$$p_0(\boldsymbol{y}) := \sup_{t\in\Omega_0} p(\boldsymbol{y}; t)$$

has a subuniform distribution for every $t \in \Omega_0$.

[23] Here I write both R and P as capitals, because they are functions of the random quantities $\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^m$.

Proof. Follows from the fact that

$$p_0(\boldsymbol{y}) \leq u \implies p(\boldsymbol{y}; t) \leq u$$

for all $t \in \Omega_0$. Therefore

$$\Pr_t\{p_0(\boldsymbol{Y}) \dot{\leq} u\} \leq \Pr_t\{p(\boldsymbol{Y}; t) \dot{\leq} u\} \leq u$$

for all $t \in \Omega_0$, where $\Pr_t$ is the probability under $\theta = t$. □

From this proof it is clear that $p_0(\boldsymbol{Y})$ can be extremely subuniform, even in the case where $p(\boldsymbol{Y}; t)$ is uniform for every $t \in \Omega_0$. As discussed in Sec. 6.3.2, subuniformity reduces the information in a P-value, and hence P-values for composite hypotheses are generally rather uninformative. Berger and Boos (1994) have a clever suggestion to address this, but it has not been taken up in practice.

A much more common approach, in the case of a single parameter of interest plus nuisance parameters, is to compute a confidence interval for the parameter, discussed in the next section.

6.4 *Confidence sets*

Confidence sets are a way of assessing uncertainty about the parameters without treating the parameters as random variables. I will say more about *level error* at the end of the section.

Definition 6.2 (Confidence set and coverage)**.** *$\mathcal{C}_\beta$ is a level β confidence set for θ exactly when $\mathcal{C}_\beta(\boldsymbol{y}) \subset \Omega$ and*

$$\Pr_t\left\{t \in \mathcal{C}_\beta(\boldsymbol{Y})\right\} \geq \beta \quad \textit{for all } t \in \Omega.$$

The probability on the lefthand side is defined as the coverage *of $\mathcal{C}_\beta$ at t. If the coverage is exactly β for all t, then the confidence set is 'exact'.*

There is a close relationship between confidence sets and P-values; for every P-value, there is a confidence set (and *vice versa*).[24] Thus reservations about P-values hold for confidence sets as well.

[24] The *vice versa* is that if θ_0 is on the boundary of a level β confidence set, then $1 - \beta$ is a P-value for $H_0 : \theta = \theta_0$.

Theorem 6.6. *Let $p(\cdot, t)$ be a P-value for the hypothesis $H_0 : \theta = t$. Then*

$$\mathcal{C}_\beta(\boldsymbol{y}) := \left\{t \in \Omega : p(\boldsymbol{y}, t) > 1 - \beta\right\}$$

is a level β confidence set for θ. If the P-value is exact, then the confidence set is exact as well.

From this construction it is immediate that, for the same P-value, $\beta \leq \beta'$ implies that $\mathcal{C}_\beta(\boldsymbol{y}) \subset \mathcal{C}_{\beta'}(\boldsymbol{y})$, so that these confidence sets are always nested. While this property is not in the definition of a confidence set, anything else would seem bizarre.

Proof. This proof uses the subuniformity property of P-values.

$$\begin{aligned}\Pr_t\{t \in \mathcal{C}_\beta(\boldsymbol{Y})\} &= \Pr_t\{p(\boldsymbol{Y}, t) > 1 - \beta\} \\ &= 1 - \Pr_t\{p(\boldsymbol{Y}, t) \leq 1 - \beta\} \\ &\geq 1 - (1 - \beta) = \beta,\end{aligned}$$

where the inequality follows from the P-value being subuniform. In the case where the P-value is uniform, the inequality is replaced by an equality, and the confidence set is exact. □

A more general definition of confidence sets holds for any function of θ. If $g : \theta \mapsto \psi$, then $\mathcal{C}_\beta^\psi(\boldsymbol{y})$ is a level β confidence set for ψ exactly when

$$\Pr_t \left\{ g(t) \in \mathcal{C}_\beta^\psi(\boldsymbol{Y}) \right\} \geq \beta \quad \text{for all } t \in \Omega.$$

If ψ is one-dimensional and $\mathcal{C}_\beta^\psi(y)$ is convex for every $\boldsymbol{y}$, then $\mathcal{C}_\beta^\psi$ is termed a *confidence interval*.

These confidence sets can be contructed directly from a confidence set for θ.

Theorem 6.7 (Marginal confidence sets)**.**
If $\mathcal{C}_\beta$ is a level β confidence set for θ, and $g : \theta \mapsto \psi$, then $g\mathcal{C}_\beta$ is a level β confidence set for ψ.[25]

[25] If A is a set in $\mathcal{A}$, and g is a function with domain $\mathcal{A}$, then gA is the set $\{b : b = g(a) \text{ for some } a \in A\}$.

Proof. This follows immediately from

$$t \in \mathcal{C}_\beta(\boldsymbol{y}) \implies g(t) \in g\mathcal{C}_\beta(\boldsymbol{y})$$

for each $\boldsymbol{y}$. Hence

$$\beta \leq \Pr_t \left\{ t \in \mathcal{C}_\beta(\boldsymbol{Y}) \right\} \leq \Pr_t \left\{ g(t) \in g\mathcal{C}_\beta(\boldsymbol{Y}) \right\}$$

as required. □

If g is one-to-one and $\mathcal{C}_\beta$ is an exact level β confidence set for θ, then the proof shows that $g\mathcal{C}_\beta$ is an exact level β confidence set for ψ. Otherwise, though, the coverage of $g\mathcal{C}_\beta$ might be much larger than β for all θ.

As mentioned at the end of Sec. 6.3.4, confidence intervals are an alternative to P-values for composite hypotheses involving nuisance parameters. Consider, for example, the null hypothesis $H_0 : \mu = \mu_0$ in the presence of additional nuisance parameters. The nominal 95% confidence set for all of the parameters can be marginalised to derive a nominal 95% confidence set for μ, according to Thm 6.7. If the resulting confidence interval does not contain μ_0 then the P-value for H_0 is less than 0.05, by Thm 6.6.

There are merits to presenting H_0 in terms of a confidence interval for μ, rather than a P-value for $\mu = \mu_0$, because there is a difference in interpretation between a narrow confidence interval which just misses μ_0 and a wide interval that just includes it. In the former case, μ may be significantly different from μ_0, but not enough to worry about. In the latter case it μ has the potential to be very different from μ_0, even if it is not significantly different.

In medicine, for example, there is the notion of a *minimally clinically important difference (MCID)*, say $\delta > 0$. In an ideal world a hypothesis test would compare $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu \geq \mu_0 + \delta$. A quick-and-dirty approximation would compute a 95% confidence interval for μ, for which one of the following outcomes is possible (writing $\mu_1 := \mu_0 + \delta$):

1. $\mathcal{C}^{\mu}_{0.95}(\boldsymbol{y}) < \mu_0, \mu_1$
2. $\mu_0 \in \mathcal{C}^{\mu}_{0.95}(\boldsymbol{y})$ and $\mathcal{C}^{\mu}_{0.95}(\boldsymbol{y}) < \mu_1$
3. $\mu_0, \mu_1 \in \mathcal{C}^{\mu}_{0.95}(\boldsymbol{y})$
4. $\mu_0 < \mathcal{C}^{\mu}_{0.95}(\boldsymbol{y})$ and $\mathcal{C}^{\mu}_{0.95}(\boldsymbol{y}) < \mu_1$
5. $\mu_0 < \mathcal{C}^{\mu}_{0.95}(\boldsymbol{y})$ and $\mu_1 \in \mathcal{C}^{\mu}_{0.95}(\boldsymbol{y})$
6. $\mu_0, \mu_1 < \mathcal{C}^{\mu}_{0.95}(\boldsymbol{y})$.

We could imagine, depending on the application, that each one of these outcomes could have a different interpretation. Hence a confidence interval for μ provides a much richer source of information than a P-value for $H_0 : \mu = \mu_0$.

Level error and calibration. Level error is the different between the nominal coverage β and the actual coverage, both of which depend on θ. Level error typically arises when large-sample theory is used to construct a confidence set which is asymptotically exact. But, for finite n, the coverage is only approximately β, and can vary by θ.

Due to the duality of P-values and confidence sets, level error in a confidence set is equivalent to error in the calculation of a P-value. This error cannot be assumed to be in the direction of larger confidence sets and P-values, which would be consistent with their definitions. It may be that confidence sets are too small, and P-values likewise. For people who make the mistake of confusing P-values and hypothesis tests (see Sec. 6.3.2) this would lead to too many H_0's rejected. This was the basis of John Ioannidis's controversial paper 'Why most published research findings are false' (Ioannidis, 2005). He noted that ambitious scientists had an incentive not to correct this bias, because a small P-value increased their chance of publication in a prestigious journal.

If it is possible to sample cheaply from the model for $\boldsymbol{Y}$ then there is a simple way to correct for level error, to first order, which is to adjust the nominal coverage β until the actual coverage at the MLE $\hat{\theta}(\boldsymbol{y})$ is equal to the desired coverage. This is *bootstrap calibration*, see DiCiccio and Efron (1996). I find it surprising that this sensible precaution is not better known (but perhaps I should be more cynical!). My advice would be not to trust a confidence set or P-value unless it has been calibrated to have the desired coverage at the MLE.

In related research the Observational Medical Outcomes Partnership (OMOP) have studied confidence intervals and P-values from observational studies based on medical databases (Schuemie *et al.*, 2014; Madigan *et al.*, 2014). The very alarming conclusion is

> Empirical calibration was found to reduce spurious results to the desired 5% level. Applying these adjustments to literature suggests that at least 54% of findings with $p < 0.05$ are not actually statistically significant and should be reevaluated. (Schuemie *et al.*, 2014, abstract)

Commentators are talking about a 'crisis of reproducibility' in science, and it looks as though uncorrected level error in confidence intervals and *P*-values is partly to blame.