

## 37

---

### *Bayesian Inference and Sampling Theory*

There are two schools of statistics. Sampling theorists concentrate on having methods guaranteed to work most of the time, given minimal assumptions. Bayesians try to make inferences that take into account all available information and answer the question of interest given the particular data set. As you have probably gathered, I strongly recommend the use of Bayesian methods.

Sampling theory is the widely used approach to statistics, and most papers in most journals report their experiments using quantities like confidence intervals, significance levels, and  $p$ -values. A  $p$ -value (e.g.  $p = 0.05$ ) is the probability, given a null hypothesis for the probability distribution of the data, that the outcome would be as extreme as, or more extreme than, the observed outcome. Untrained readers – and perhaps, more worryingly, the authors of many papers – usually interpret such a  $p$ -value as if it is a Bayesian probability (for example, the posterior probability of the null hypothesis), an interpretation that both sampling theorists and Bayesians would agree is incorrect.

In this chapter we study a couple of simple inference problems in order to compare these two approaches to statistics.

While in some cases, the answers from a Bayesian approach and from sampling theory are very similar, we can also find cases where there are significant differences. We have already seen such an example in exercise 3.15 (p.59), where a sampling theorist got a  $p$ -value smaller than 7%, and viewed this as strong evidence *against* the null hypothesis, whereas the data actually *favoured* the null hypothesis over the simplest alternative. On p.64, another example was given where the  $p$ -value was smaller than the mystical value of 5%, yet the data again favoured the null hypothesis. Thus in some cases, sampling theory can be trigger-happy, declaring results to be ‘sufficiently improbable that the null hypothesis should be rejected’, when those results actually weakly support the null hypothesis. As we will now see, there are also inference problems where sampling theory fails to detect ‘significant’ evidence where a Bayesian approach and everyday intuition agree that the evidence is strong. Most telling of all are the inference problems where the ‘significance’ assigned by sampling theory changes depending on irrelevant factors concerned with the design of the experiment.

This chapter is only provided for those readers who are curious about the sampling theory / Bayesian methods debate. If you find any of this chapter tough to understand, please skip it. There is no point trying to understand the debate. Just use Bayesian methods – they are much easier to understand than the debate itself!

### ► 37.1 A medical example

We are trying to reduce the incidence of an unpleasant disease called *microsoftus*. Two vaccinations,  $A$  and  $B$ , are tested on a group of volunteers. Vaccination  $B$  is a control treatment, a placebo treatment with no active ingredients. Of the 40 subjects, 30 are randomly assigned to have treatment  $A$  and the other 10 are given the control treatment  $B$ . We observe the subjects for one year after their vaccinations. Of the 30 in group  $A$ , one contracts *microsoftus*. Of the 10 in group  $B$ , three contract *microsoftus*.

Is treatment  $A$  better than treatment  $B$ ?

*Sampling theory has a go*

The standard sampling theory approach to the question ‘is  $A$  better than  $B$ ?’ is to construct a *statistical test*. The test usually compares a hypothesis such as

$\mathcal{H}_1$ : ‘ $A$  and  $B$  have different effectivenesses’

with a null hypothesis such as

$\mathcal{H}_0$ : ‘ $A$  and  $B$  have exactly the same effectivenesses as each other’.

A novice might object ‘no, no, I want to compare the hypothesis “ $A$  is better than  $B$ ” with the alternative “ $B$  is better than  $A$ ”!’ but such objections are not welcome in sampling theory.

Once the two hypotheses have been defined, the first hypothesis is scarcely mentioned again – attention focuses solely on the null hypothesis. It makes me laugh to write this, but it’s true! The null hypothesis is accepted or rejected purely on the basis of how unexpected the data were to  $\mathcal{H}_0$ , not on how much better  $\mathcal{H}_1$  predicted the data. One chooses a *statistic* which measures how much a data set deviates from the null hypothesis. In the example here, the standard statistic to use would be one called  $\chi^2$  (chi-squared). To compute  $\chi^2$ , we take the difference between each data measurement and its *expected* value *assuming the null hypothesis to be true*, and divide the square of that difference by the *variance* of the measurement, *assuming the null hypothesis to be true*. In the present problem, the four data measurements are the integers  $F_{A+}$ ,  $F_{A-}$ ,  $F_{B+}$ , and  $F_{B-}$ , that is, the number of subjects given treatment  $A$  who contracted *microsoftus* ( $F_{A+}$ ), the number of subjects given treatment  $A$  who didn’t ( $F_{A-}$ ), and so forth. The definition of  $\chi^2$  is:

$$\chi^2 = \sum_i \frac{(F_i - \langle F_i \rangle)^2}{\langle F_i \rangle}. \quad (37.1)$$

Actually, in my elementary statistics book (Spiegel, 1988) I find Yates’s correction is recommended:

$$\chi^2 = \sum_i \frac{(|F_i - \langle F_i \rangle| - 0.5)^2}{\langle F_i \rangle}. \quad (37.2)$$

In this case, given the null hypothesis that treatments  $A$  and  $B$  are equally effective, and have rates  $f_+$  and  $f_-$  for the two outcomes, the expected counts are:

$$\begin{aligned} \langle F_{A+} \rangle &= f_+ N_A & \langle F_{A-} \rangle &= f_- N_A \\ \langle F_{B+} \rangle &= f_+ N_B & \langle F_{B-} \rangle &= f_- N_B. \end{aligned} \quad (37.3)$$

If you want to know about Yates’s correction, read a sampling theory textbook. The point of this chapter is not to teach sampling theory; I merely mention Yates’s correction because it is what a professional sampling theorist might use.

The test accepts or rejects the null hypothesis on the basis of how big  $\chi^2$  is. To make this test precise, and give it a ‘significance level’, we have to work out what the *sampling distribution* of  $\chi^2$  is, taking into account the fact that the four data points are not independent (they satisfy the two constraints  $F_{A+} + F_{A-} = N_A$  and  $F_{B+} + F_{B-} = N_B$ ) and the fact that the parameters  $f_{\pm}$  are not known. These three constraints reduce the *number of degrees of freedom* in the data from four to one. [If you want to learn more about computing the ‘number of degrees of freedom’, read a sampling theory book; in Bayesian methods we don’t need to know all that, and quantities equivalent to the number of degrees of freedom pop straight out of a Bayesian analysis when they are appropriate.] These sampling distributions are tabulated by sampling theory gnomes and come accompanied by warnings about the conditions under which they are accurate. For example, standard tabulated distributions for  $\chi^2$  are only accurate if the expected numbers  $F_i$  are about 5 or more.

Once the data arrive, sampling theorists estimate the unknown parameters  $f_{\pm}$  of the null hypothesis from the data:

$$\hat{f}_+ = \frac{F_{A+} + F_{B+}}{N_A + N_B}, \quad \hat{f}_- = \frac{F_{A-} + F_{B-}}{N_A + N_B}, \quad (37.4)$$

and evaluate  $\chi^2$ . At this point, the sampling theory school divides itself into two camps. One camp uses the following protocol: first, before looking at the data, pick the significance level of the test (e.g. 5%), and determine the critical value of  $\chi^2$  above which the null hypothesis will be rejected. (The significance level is the fraction of times that the statistic  $\chi^2$  would exceed the critical value, if the null hypothesis were true.) Then evaluate  $\chi^2$ , compare with the critical value, and declare the outcome of the test, and its significance level (which was fixed beforehand).

The second camp looks at the data, finds  $\chi^2$ , then looks in the table of  $\chi^2$ -distributions for the significance level,  $p$ , for which the observed value of  $\chi^2$  would be the critical value. The result of the test is then reported by giving this value of  $p$ , which is the fraction of times that a result as extreme as the one observed, or more extreme, would be expected to arise if the null hypothesis were true.

Let’s apply these two methods. First camp: let’s pick 5% as our significance level. The critical value for  $\chi^2$  with one degree of freedom is  $\chi_{0.05}^2 = 3.84$ . The estimated values of  $f_{\pm}$  are

$$f_+ = 1/10, \quad f_- = 9/10. \quad (37.5)$$

The expected values of the four measurements are

$$\langle F_{A+} \rangle = 3 \quad (37.6)$$

$$\langle F_{A-} \rangle = 27 \quad (37.7)$$

$$\langle F_{B+} \rangle = 1 \quad (37.8)$$

$$\langle F_{B-} \rangle = 9 \quad (37.9)$$

and  $\chi^2$  (as defined in equation (37.1)) is

$$\chi^2 = 5.93. \quad (37.10)$$

Since this value exceeds 3.84, we reject the null hypothesis that the two treatments are equivalent at the 0.05 significance level. However, if we use Yates’s correction, we find  $\chi^2 = 3.33$ , and therefore accept the null hypothesis.

The sampling distribution of a statistic is the probability distribution of its value under repetitions of the experiment, assuming that the null hypothesis is true.

Camp two runs a finger across the  $\chi^2$  table found at the back of any good sampling theory book and finds  $\chi_{.10}^2 = 2.71$ . Interpolating between  $\chi_{.10}^2$  and  $\chi_{.05}^2$ , camp two reports ‘the  $p$ -value is  $p = 0.07$ ’.

Notice that this answer does not say how much more effective  $A$  is than  $B$ , it simply says that  $A$  is ‘significantly’ different from  $B$ . And here, ‘significant’ means only ‘statistically significant’, not practically significant.

The man in the street, reading the statement that ‘the treatment was significantly different from the control ( $p = 0.07$ )’, might come to the conclusion that ‘there is a 93% chance that the treatments differ in effectiveness’. But what ‘ $p = 0.07$ ’ actually means is ‘if you did this experiment many times, and the two treatments *had* equal effectiveness, then 7% of the time you would find a value of  $\chi^2$  more extreme than the one that happened here’. This has almost nothing to do with what we want to know, which is how likely it is that treatment  $A$  is better than  $B$ .

*Let me through, I’m a Bayesian*

OK, now let’s *infer* what we really want to know. We scrap the hypothesis that the two treatments have exactly equal effectivenesses, since we do not believe it. There are two unknown parameters,  $p_{A+}$  and  $p_{B+}$ , which are the probabilities that people given treatments  $A$  and  $B$ , respectively, contract the disease.

Given the data, we can infer these two probabilities, and we can answer questions of interest by examining the posterior distribution.

The posterior distribution is

$$P(p_{A+}, p_{B+} | \{F_i\}) = \frac{P(\{F_i\} | p_{A+}, p_{B+}) P(p_{A+}, p_{B+})}{P(\{F_i\})}. \quad (37.11)$$

The likelihood function is

$$P(\{F_i\} | p_{A+}, p_{B+}) = \binom{N_A}{F_{A+}} p_{A+}^{F_{A+}} p_{A-}^{F_{A-}} \binom{N_B}{F_{B+}} p_{B+}^{F_{B+}} p_{B-}^{F_{B-}} \quad (37.12)$$

$$= \binom{30}{1} p_{A+}^1 p_{A-}^{29} \binom{10}{3} p_{B+}^3 p_{B-}^7. \quad (37.13)$$

What prior distribution should we use? The prior distribution gives us the opportunity to include knowledge from other experiments, or a prior belief that the two parameters  $p_{A+}$  and  $p_{B+}$ , while different from each other, are expected to have similar values.

Here we will use the simplest vanilla prior distribution, a uniform distribution over each parameter.

$$P(p_{A+}, p_{B+}) = 1. \quad (37.14)$$

We can now plot the posterior distribution. Given the assumption of a separable prior on  $p_{A+}$  and  $p_{B+}$ , the posterior distribution is also separable:

$$P(p_{A+}, p_{B+} | \{F_i\}) = P(p_{A+} | F_{A+}, F_{A-}) P(p_{B+} | F_{B+}, F_{B-}). \quad (37.15)$$

The two posterior distributions are shown in figure 37.1 (except the graphs are not normalized) and the joint posterior probability is shown in figure 37.2.

If we want to know the answer to the question ‘how probable is it that  $p_{A+}$  is smaller than  $p_{B+}$ ’, we can answer exactly that question by computing the posterior probability

$$P(p_{A+} < p_{B+} | \text{Data}), \quad (37.16)$$

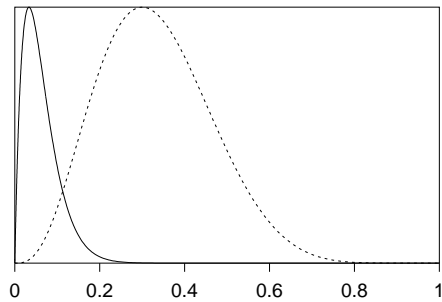


Figure 37.1. Posterior probabilities of the two effectivenesses. Treatment A – solid line; B – dotted line.

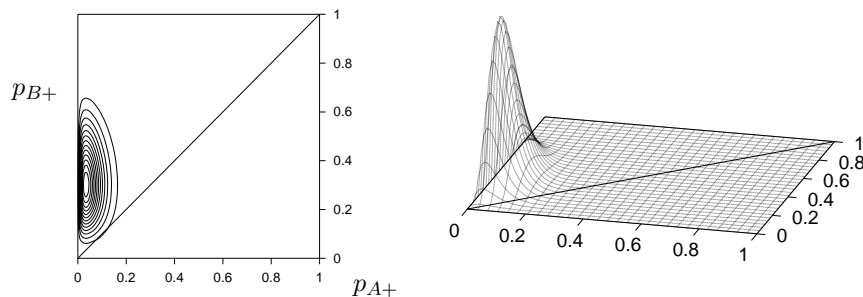


Figure 37.2. Joint posterior probability of the two effectivenesses – contour plot and surface plot.

which is the integral of the joint posterior probability  $P(p_{A+}, p_{B+} | \text{Data})$  shown in figure 37.2 over the region in which  $p_{A+} < p_{B+}$ , i.e., the shaded triangle in figure 37.3. The value of this integral (obtained by a straightforward numerical integration of the likelihood function (37.13) over the relevant region) is  $P(p_{A+} < p_{B+} | \text{Data}) = 0.990$ .

Thus there is a 99% chance, given the data and our prior assumptions, that treatment A is superior to treatment B. In conclusion, according to our Bayesian model, the data (1 out of 30 contracted the disease after vaccination A, and 3 out of 10 contracted the disease after vaccination B) give very strong evidence – about 99 to one – that treatment A is superior to treatment B.

In the Bayesian approach, it is also easy to answer other relevant questions. For example, if we want to know ‘how likely is it that treatment A is ten times more effective than treatment B?’, we can integrate the joint posterior probability  $P(p_{A+}, p_{B+} | \text{Data})$  over the region in which  $p_{A+} < 10 p_{B+}$  (figure 37.4).

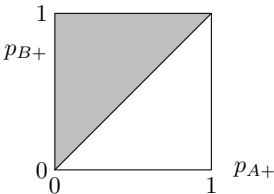


Figure 37.3. The proposition  $p_{A+} < p_{B+}$  is true for all points in the shaded triangle. To find the probability of this proposition we integrate the joint posterior probability  $P(p_{A+}, p_{B+} | \text{Data})$  (figure 37.2) over this region.

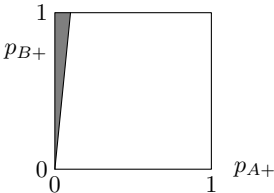


Figure 37.4. The proposition  $p_{A+} < 10 p_{B+}$  is true for all points in the shaded triangle.

Model comparison

If there were a situation in which we really did want to compare the two hypotheses  $\mathcal{H}_0: p_{A+} = p_{B+}$  and  $\mathcal{H}_1: p_{A+} \neq p_{B+}$ , we can of course do this directly with Bayesian methods also.

As an example, consider the data set:

D: One subject, given treatment A, subsequently contracted *microsofts*.  
One subject, given treatment B, did not.

| Treatment     | A | B |
|---------------|---|---|
| Got disease   | 1 | 0 |
| Did not       | 0 | 1 |
| Total treated | 1 | 1 |

How strongly does this data set favour  $\mathcal{H}_1$  over  $\mathcal{H}_0$ ?

We answer this question by computing the evidence for each hypothesis. Let's assume uniform priors over the unknown parameters of the models. The first hypothesis  $\mathcal{H}_0$ :  $p_{A+} = p_{B+}$  has just one unknown parameter, let's call it  $p$ .

$$P(p | \mathcal{H}_0) = 1 \quad p \in (0, 1). \quad (37.17)$$

We'll use the uniform prior over the two parameters of model  $\mathcal{H}_1$  that we used before:

$$P(p_{A+}, p_{B+} | \mathcal{H}_1) = 1 \quad p_{A+} \in (0, 1), p_{B+} \in (0, 1). \quad (37.18)$$

Now, the probability of the data  $D$  under model  $\mathcal{H}_0$  is the normalizing constant from the inference of  $p$  given  $D$ :

$$P(D | \mathcal{H}_0) = \int dp P(D | p) P(p | \mathcal{H}_0) \quad (37.19)$$

$$= \int dp p(1-p) \times 1 \quad (37.20)$$

$$= 1/6. \quad (37.21)$$

The probability of the data  $D$  under model  $\mathcal{H}_1$  is given by a simple two-dimensional integral:

$$P(D | \mathcal{H}_1) = \int \int dp_{A+} dp_{B+} P(D | p_{A+}, p_{B+}) P(p_{A+}, p_{B+} | \mathcal{H}_1) \quad (37.22)$$

$$= \int dp_{A+} p_{A+} \int dp_{B+} (1 - p_{B+}) \quad (37.23)$$

$$= 1/2 \times 1/2 \quad (37.24)$$

$$= 1/4. \quad (37.25)$$

Thus the evidence ratio in favour of model  $\mathcal{H}_1$ , which asserts that the two effectivenesses are unequal, is

$$\frac{P(D | \mathcal{H}_1)}{P(D | \mathcal{H}_0)} = \frac{1/4}{1/6} = \frac{0.6}{0.4}. \quad (37.26)$$

So if the prior probability over the two hypotheses was 50:50, the posterior probability is 60:40 in favour of  $\mathcal{H}_1$ .  $\square$

Is it not easy to get sensible answers to well-posed questions using Bayesian methods?

[The sampling theory answer to this question would involve the identical significance test that was used in the preceding problem; that test would yield a 'not significant' result. I think it is greatly preferable to acknowledge what is obvious to the intuition, namely that the data  $D$  do give *weak* evidence in favour of  $\mathcal{H}_1$ . Bayesian methods quantify how weak the evidence is.]

## ► 37.2 Dependence of $p$ -values on irrelevant information

In an expensive laboratory, Dr. Bloggs tosses a coin labelled  $a$  and  $b$  twelve times and the outcome is the string

$$aaabaaaaabaab,$$

which contains three  $b$ s and nine  $a$ s.

What evidence do these data give that the coin is biased in favour of  $a$ ?

Dr. Bloggs consults his sampling theory friend who says ‘let  $r$  be the number of  $bs$  and  $n = 12$  be the total number of tosses; I view  $r$  as the random variable and find the probability of  $r$  taking on the value  $r = 3$  or a more extreme value, assuming the null hypothesis  $p_a = 0.5$  to be true’. He thus computes

$$\begin{aligned} P(r \leq 3 | n=12, \mathcal{H}_0) &= \sum_{r=0}^3 \binom{n}{r} 1/2^n = ((\binom{12}{0}) + (\binom{12}{1}) + (\binom{12}{2}) + (\binom{12}{3})) 1/2^{12} \\ &= 0.07, \end{aligned} \quad (37.27)$$

and reports ‘at the significance level of 5%, there is not significant evidence of bias in favour of  $a$ ’. Or, if the friend prefers to report  $p$ -values rather than simply compare  $p$  with 5%, he would report ‘the  $p$ -value is 7%, which is not conventionally viewed as significantly small’. If a two-tailed test seemed more appropriate, he might compute the two-tailed area, which is twice the above probability, and report ‘the  $p$ -value is 15%, which is not significantly small’. We won’t focus on the issue of the choice between the one-tailed and two-tailed tests, as we have bigger fish to catch.

Dr. Bloggs pays careful attention to the calculation (37.27), and responds ‘no, no, the random variable in the experiment was not  $r$ : I decided before running the experiment that I would keep tossing the coin until I saw three  $bs$ ; the random variable is thus  $n$ ’.

Such experimental designs are not unusual. In my experiments on error-correcting codes I often simulate the decoding of a code until a chosen number  $r$  of block errors ( $bs$ ) has occurred, since the error on the inferred value of  $\log p_b$  goes roughly as  $\sqrt{r}$ , independent of  $n$ .



**Exercise 37.1.**<sup>[2]</sup> Find the Bayesian inference about the bias  $p_a$  of the coin given the data, and determine whether a Bayesian’s inferences depend on what stopping rule was in force.

According to sampling theory, a different calculation is required in order to assess the ‘significance’ of the result  $n = 12$ . The probability distribution of  $n$  given  $\mathcal{H}_0$  is the probability that the first  $n-1$  tosses contain exactly  $r-1$   $bs$  and then the  $n$ th toss is a  $b$ .

$$P(n | \mathcal{H}_0, r) = \binom{n-1}{r-1} 1/2^n. \quad (37.28)$$

The sampling theorist thus computes

$$P(n \geq 12 | r=3, \mathcal{H}_0) = 0.03. \quad (37.29)$$

He reports back to Dr. Bloggs, ‘the  $p$ -value is 3% – there *is* significant evidence of bias after all!’

What do you think Dr. Bloggs should do? Should he publish the result, with this marvellous  $p$ -value, in one of the journals that insists that all experimental results have their ‘significance’ assessed using sampling theory? Or should he boot the sampling theorist out of the door and seek a coherent method of assessing significance, one that does not depend on the stopping rule?

At this point the audience divides in two. Half the audience intuitively feel that the stopping rule is irrelevant, and don’t need any convincing that the answer to exercise 37.1 (p.463) is ‘the inferences about  $p_a$  do not depend on the stopping rule’. The other half, perhaps on account of a thorough

training in sampling theory, intuitively feel that Dr. Bloggs's stopping rule, which stopped tossing the moment the third  $b$  appeared, may have biased the experiment somehow. If you are in the second group, I encourage you to reflect on the situation, and hope you'll eventually come round to the view that is consistent with the likelihood principle, which is that the stopping rule is *not* relevant to what we have learned about  $p_a$ .

As a thought experiment, consider some onlookers who (in order to save money) are spying on Dr. Bloggs's experiments: each time he tosses the coin, the spies update the values of  $r$  and  $n$ . The spies are eager to make inferences from the data as soon as each new result occurs. Should the spies' beliefs about the bias of the coin depend on Dr. Bloggs's intentions regarding the continuation of the experiment?

The fact that the  $p$ -values of sampling theory *do* depend on the stopping rule (indeed, whole volumes of the sampling theory literature are concerned with the task of assessing 'significance' when a complicated stopping rule is required – 'sequential probability ratio tests', for example) seems to me a compelling argument for having nothing to do with  $p$ -values at all. A Bayesian solution to this inference problem was given in sections 3.2 and 3.3 and exercise 3.15 (p.59).

Would it help clarify this issue if I added one more scene to the story? The janitor, who's been eavesdropping on Dr. Bloggs's conversation, comes in and says 'I happened to notice that just after you stopped doing the experiments on the coin, the Officer for Whimsical Departmental Rules ordered the immediate destruction of all such coins. Your coin was therefore destroyed by the departmental safety officer. There is no way you could have continued the experiment much beyond  $n = 12$  tosses. Seems to me, you need to recompute your  $p$ -value?'

### ► 37.3 Confidence intervals

In an experiment in which data  $D$  are obtained from a system with an unknown parameter  $\theta$ , a standard concept in sampling theory is the idea of a *confidence interval* for  $\theta$ . Such an interval  $(\theta_{\min}(D), \theta_{\max}(D))$  has associated with it a *confidence level* such as 95% which is informally interpreted as 'the probability that  $\theta$  lies in the confidence interval'.

Let's make precise what the confidence level really means, then give an example. A confidence interval is a function  $(\theta_{\min}(D), \theta_{\max}(D))$  of the data set  $D$ . The confidence level of the confidence interval is a property that we can compute before the data arrive. We imagine generating many data sets from a particular true value of  $\theta$ , and calculating the interval  $(\theta_{\min}(D), \theta_{\max}(D))$ , and then checking whether the true value of  $\theta$  lies in that interval. If, averaging over all these imagined repetitions of the experiment, the true value of  $\theta$  lies in the confidence interval a fraction  $f$  of the time, and this property holds for all true values of  $\theta$ , then the confidence level of the confidence interval is  $f$ .

For example, if  $\theta$  is the mean of a Gaussian distribution which is known to have standard deviation 1, and  $D$  is a sample from that Gaussian, then  $(\theta_{\min}(D), \theta_{\max}(D)) = (D-2, D+2)$  is a 95% confidence interval for  $\theta$ .

Let us now look at a simple example where the meaning of the confidence level becomes clearer. Let the parameter  $\theta$  be an integer, and let the data be a pair of points  $x_1, x_2$ , drawn independently from the following distribution:

$$P(x|\theta) = \begin{cases} 1/2 & x = \theta \\ 1/2 & x = \theta + 1 \\ 0 & \text{for other values of } x. \end{cases} \quad (37.30)$$



For example, if  $\theta$  were 39, then we could expect the following data sets:

$$\begin{aligned} D = (x_1, x_2) = (39, 39) & \text{ with probability } 1/4; \\ (x_1, x_2) = (39, 40) & \text{ with probability } 1/4; \\ (x_1, x_2) = (40, 39) & \text{ with probability } 1/4; \\ (x_1, x_2) = (40, 40) & \text{ with probability } 1/4. \end{aligned} \quad (37.31)$$

We now consider the following confidence interval:

$$[\theta_{\min}(D), \theta_{\max}(D)] = [\min(x_1, x_2), \min(x_1, x_2)]. \quad (37.32)$$

For example, if  $(x_1, x_2) = (40, 39)$ , then the confidence interval for  $\theta$  would be  $[\theta_{\min}(D), \theta_{\max}(D)] = [39, 39]$ .

Let's think about this confidence interval. What is its confidence level? By considering the four possibilities shown in (37.31), we can see that there is a 75% chance that the confidence interval will contain the true value. The confidence interval therefore has a confidence level of 75%, by definition.

Now, what if the data we acquire are  $(x_1, x_2) = (29, 29)$ ? Well, we can compute the confidence interval, and it is  $[29, 29]$ . So shall we report this interval, and its associated confidence level, 75%? This would be correct by the rules of sampling theory. But does this make sense? What do we actually know in this case? Intuitively, or by Bayes' theorem, it is clear that  $\theta$  could either be 29 or 28, and both possibilities are equally likely (if the prior probabilities of 28 and 29 were equal). The posterior probability of  $\theta$  is 50% on 29 and 50% on 28.

What if the data are  $(x_1, x_2) = (29, 30)$ ? In this case, the confidence interval is still  $[29, 29]$ , and its associated confidence level is 75%. But in this case, by Bayes' theorem, or common sense, we are 100% sure that  $\theta$  is 29.

In neither case is the probability that  $\theta$  lies in the '75% confidence interval' equal to 75%!

Thus

1. the way in which many people interpret the confidence levels of sampling theory is *incorrect*;
2. given some data, what people usually want to know (whether they know it or not) is a Bayesian posterior probability distribution.

Are all these examples contrived? Am I making a fuss about nothing? If you are sceptical about the dogmatic views I have expressed, I encourage you to look at a case study: look in depth at exercise 35.4 (p.446) and the reference (Kepler and Oprea, 2001), in which sampling theory estimates and confidence intervals for a mutation rate are constructed. Try both methods on simulated data – the Bayesian approach based on simply computing the likelihood function, and the confidence interval from sampling theory; and let me know if you don't find that the Bayesian answer is always better than the sampling theory answer; and often much, much better. This suboptimality of sampling theory, achieved with great effort, is why I am passionate about Bayesian methods. Bayesian methods are straightforward, and they optimally use all the information in the data.

## ► 37.4 Some compromise positions

Let's end on a conciliatory note. Many sampling theorists are pragmatic – they are happy to choose from a selection of statistical methods, choosing whichever has the 'best' long-run properties. In contrast, I have no problem

with the idea that there is only *one* answer to a well-posed problem; but it's not essential to convert sampling theorists to this viewpoint: instead, we can offer them Bayesian estimators and Bayesian confidence intervals, and request that the sampling theoretical properties of these methods be evaluated. We don't need to mention that the methods are derived from a Bayesian perspective. If the sampling properties are good then the pragmatic sampling theorist will choose to use the Bayesian methods. It is indeed the case that many Bayesian methods have good sampling-theoretical properties. Perhaps it's not surprising that a method that gives the optimal answer for each individual case should also be good in the long run!

Another piece of common ground can be conceded: while I believe that most well-posed inference problems have a unique correct answer, which can be found by Bayesian methods, not all problems are well-posed. A common question arising in data modelling is 'am I using an appropriate model?' Model criticism, that is, hunting for defects in a current model, is a task that may be aided by sampling theory tests, in which the null hypothesis ('the current model is correct') is well defined, but the alternative model is not specified. One could use sampling theory measures such as  $p$ -values to guide one's search for the aspects of the model most in need of scrutiny.

### Further reading

My favourite reading on this topic includes (Jaynes, 1983; Gull, 1988; Lored, 1990; Berger, 1985; Jaynes, 2003). Treatises on Bayesian statistics from the statistics community include (Box and Tiao, 1973; O'Hagan, 1994).

## ► 37.5 Further exercises

- ▷ Exercise 37.2.<sup>[3C]</sup> A traffic survey records traffic on two successive days. On Friday morning, there are 12 vehicles in one hour. On Saturday morning, there are 9 vehicles in half an hour. Assuming that the vehicles are Poisson distributed with rates  $\lambda_F$  and  $\lambda_S$  (in vehicles per hour) respectively,
  - (a) is  $\lambda_S$  greater than  $\lambda_F$ ?
  - (b) by what factor is  $\lambda_S$  bigger or smaller than  $\lambda_F$ ?
- ▷ Exercise 37.3.<sup>[3C]</sup> Write a program to compare treatments A and B given data  $F_{A+}$ ,  $F_{A-}$ ,  $F_{B+}$ ,  $F_{B-}$  as described in section 37.1. The outputs of the program should be (a) the probability that treatment A is more effective than treatment B; (b) the probability that  $p_{A+} < 10p_{B+}$ ; (c) the probability that  $p_{B+} < 10p_{A+}$ .