10 Frequentist hypothesis testing

In this chapter we look at frequentist hypothesis testing and the resulting concept of p values. I will introduce the Gaussian z test and the (Student's) t test, as well as the χ^2 distribution as a means of assessing goodness of fit. I conclude the chapter with a discussion of the problems with this approach to assessing models. Although this is a book on Bayesian inference, and I believe frequentist hypothesis testing should often be avoided, a basic understanding of this topic is necessary.

10.1 The principles of frequentist statistics and hypothesis testing

In previous chapters we learned how to infer the posterior probability distribution over parameters, $P(\theta \,|\, D, M)$, as well the posterior probability of the model, $P(M \,|\, D)$. With this probabilistic (Bayesian) method we take a direct approach: we infer the probabilities of things we don't know (parameters, models), given the things we do (data, priors). Data are noisy and samples are incomplete, so our inferences cannot be 100% certain. Our uncertainties in both the data and the parameters are quantified by probabilities.

The frequentist approach is different. It assumes that the model parameters are fixed, and it is only the data which are variable. Parameters do not have probability distributions: we can only derive fixed values of them from the data. The idea is to establish how the data might have varied in imaginary repetitions of the measurements, in order to assess our confidence in parameter values. As the name suggests, in the frequentist approach probabilities are defined in terms of frequencies of occurrences over these imagined sets of data.

In frequentist hypothesis testing a hypothesis (or model) is either true or false: it is not possible to attach a probability to it. To decide whether or not it is true, we first assume it is, then compute a statistic (from the data). On the basis of its value we either (a) reject the hypothesis, or (b) do not reject the hypothesis. You would be right to think that this is a rather indirect approach to assessing the plausibility of a model. From this perspective it is perhaps surprising that it was standard practice in science for much of the twentieth century, and was widely taught, often as the only approach for model assessment. Part of the reason for this may have been its apparent simplicity. In addition, the probabilistic alternative using the Bayesian evidence (chapter 11) often requires numerical integration that was often beyond the computational means of the time.

The general procedure for frequentist hypothesis testing is as follows. We do an experiment, gather the data, and then compute a statistic η which summarizes these data. This might be the ratio of the mean to the standard deviation, for example. The statistic is chosen to be one which, when assuming a particular hypothesis – call it H_0 – to be true, has a predictable, one-dimensional PDF $P(\eta \,|\, H_0)$. This is called the *sampling distribution*. If we imagine repeating the experiment a large number of times and plotting the distribution of this statistic over these imaginary data sets, the statistic would follow the distribution $P(\eta \,|\, H_0)$, assuming H_0 (which is a model for the data) is true. We then use $P(\eta \,|\, H_0)$ to calculate the probability that we could have obtained (in the imaginary repetitions) a value of the statistic which is *more extreme* than the one we actually observed. This probability is called the *p value*. If the p value is sufficiently small then we say the data are unlikely to have been so extreme, and so we reject the hypothesis. If the p value is not small then we cannot reject the hypothesis. Note that this is not the same as accepting the hypothesis. In frequentist hypothesis testing we can never accept a hypothesis. The best we can do is fail to reject it, and simultaneously reject all other rival hypotheses. What counts as "sufficiently small" is a matter of choice (and will be discussed later).

The reason for the curious procedure of imagining "data more extreme than those observed" is that the probability of obtaining exactly any specific value of the statistic is zero. This is because the probability *density* function is a continuous, finite function and so has zero probability at any point (an infinitesimally narrow range)

$$\text{probability} = P(\eta)\,\delta\eta \;\; \rightarrow \;\; 0 \quad \text{as} \quad \delta\eta \rightarrow 0. \tag{10.1}$$

We must instead consider this concept of "more extreme", namely

$$p = \int_{\eta}^{\infty} P(\eta')\,d\eta' \tag{10.2}$$

which is an actual probability, the p value (for a one-sided test: more on this later). If we imagine repeating the experiment many times, then the p value is the fraction of experiments in which the measured statistic would exceed the actual measured value of η. Note that the p value is *not* the probability that the hypothesis is true. This cannot be calculated, because in frequentist hypothesis testing hypotheses do not have probabilities.

Bayesian model comparison avoids reference to "more extreme data", because in comparing models it uses the *ratio* of two probability densities, which is well defined at a point. It therefore treats competing models symmetrically, so it can, within the set of models considered, accept as well as reject hypotheses (at some quantified probability level). Frequentist hypothesis testing instead compares the observed data (via the statistic) to hypothetical "more extreme" data sets.

We normally use frequentist hypothesis testing to try to reject a so-called *null hypothesis*. The null hypothesis is generally chosen to be something for which the statistic has a simple distribution (so in practice the statistic and the null hypothesis are chosen together). The null hypothesis is usually a model which shows "no interesting signal", making our attempt to reject the null hypothesis something like proof by contradiction.

If we manage to reject the null hypothesis, then this suggests that some alternative hypothesis may be true. But it does not tell us anything about what that alternative hypothesis is. The only time this is possible is when the alternative is the complement of the null, i.e. the two models are mutually exclusive and exhaustive.

In discussions of frequentist hypothesis testing you will sometimes encounter type I and type II errors, which are two distinct types of error one can make.[1]
Type I error: probability of rejecting the hypothesis when it is true.
Type II error: probability of not rejecting the hypothesis when it is false.
These errors trade off against each other. If we set our confidence threshold sufficiently high to greatly reduce the chance of making type II errors, then we will naturally increase the chance of making type I errors and vice versa. I shall not use these terms, however.

Because the concept of hypothesis testing relies on coming up with a distribution of the test statistic, there is a plethora of hypothesis tests designed for different types of data (t test, F test, KS test, signed rank test, U test, etc.). There are statistical recipe books dedicated to providing a suitable test for given ingredients. Here I will just look at a few widely-used examples which assume Gaussian-distributed data. They differ according to what properties of the parent distribution are known and what null hypothesis we are trying to test. After we have seen how hypothesis testing works, I will return at the end of this chapter to discuss some of the conceptual implications and problems of hypothesis testing.

10.2 One-sample hypothesis tests

One-sample hypothesis tests are used when we have a single sample of data and typically want to ask whether the mean of the sample differs from some specified value. There are two widely used tests, both of which assume Gaussian distributions for the data. They differ in what information we are given. The z test assumes the standard deviation is known. The t test assumes it is not.

10.2.1 Gaussian z test

A company has long produced batteries which, due to inevitable variability in production, have an energy storage capacity that can be described by a Gaussian distribution with a mean of 200 MJ and a standard deviation of 20 MJ. The company is now selling a battery which it claims has a higher mean capacity (but same standard deviation), based on new technology. One of these new batteries is measured to have a capacity of 230 MJ. Is this new technology really better?

To formulate this question statistically we ask "what is the probability that this capacity could be achieved by the old technology?". We need to find the probability that the old process would produce a battery with a capacity as high as 230 MJ. We therefore test the following null hypothesis:

[1] Sometimes these are referred to as false positive and false negative probabilities, but which is which depends on whether you consider rejecting a hypothesis as positive or negative.

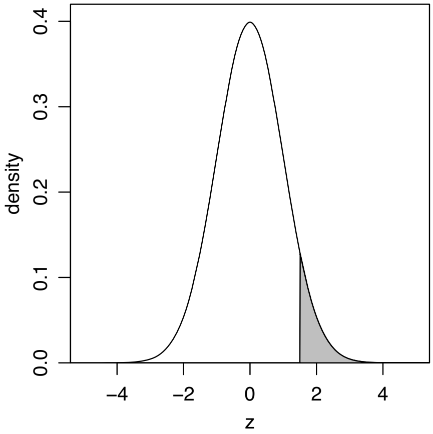

Fig. 10.1 The standardized Gaussian distribution as used in the Gaussian z test. The grey region indicates $z > 1.5$, which has a probability of 0.067.

H_0: the battery was drawn from a Gaussian with mean $\mu = 200\,\mathrm{MJ}$ and standard deviation $\sigma = 20\,\mathrm{MJ}$.

As we're interested in a battery that has a capacity of $x = 230\,\mathrm{MJ}$ *or even higher*, we carry out a so-called one-sided test. We want to find the probability that we would get a battery with as much capacity as the one observed, assuming the null hypothesis to be true. We do this by calculating the z value or z statistic

$$z = \frac{x - \mu}{\sigma}. \tag{10.3}$$

As μ and σ are fixed and x has a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, the statistic z is the standardized Gaussian distribution $\mathcal{N}(0, 1)$, which has no free parameters. In the present example

$$z = \frac{230 - 200}{20} = 1.5. \tag{10.4}$$

The probability of getting this value of z or higher, the p value, is the area under the distribution with $z > 1.5$, shown graphically in figure 10.1. We compute the area by integration using the cumulative distribution function (equation 1.64), for which R provides functions for standard distributions. For the Gaussian distribution, this area is[2]

```
1 - pnorm(1.5) = 0.067
```

This is of course equivalent to considering the probability of drawing a value of 230 or higher from a $\mathcal{N}(200, 20)$ distribution:

```
1 - pnorm(230, mean=200, sd=20) = 0.067
```

We interpret this result to mean that there is a probability of 0.067 that we would get a battery with at least the measured capacity from the old technology (the null hypothesis).

[2] As the Gaussian is symmetric, `1 - pnorm(q) = pnorm(-q)`.

As this probability is quite small, it suggests that the old technology was not the one used, i.e. the null hypothesis may not be true, so we may want to reject it. We may then take this to imply that the company's claim is true. Remember, however, that by rejecting a null we do not automatically accept some alternative, even if it is the complement of the null.[3] Nonetheless, an even smaller p value would give you more confidence that the null hypothesis is not true.

Strictly speaking we should not be talking about the null hypothesis being true or not, because we are doing a one-sided test. We are only interested in whether the battery has a *larger* capacity than before. That is, we would not reject the null hypothesis if we measured a very small capacity (a large negative value of z), even though getting such extreme values is also unlikely under the null hypothesis. In the current one-sided test, the implicit *alternative hypothesis* H_1 is therefore

H_1: the battery was drawn from a Gaussian distribution with mean $\mu > 200\,\text{MJ}$.

This is only the complement of H_0 for a one-sided test. A two-sided test would be to test whether the capacity is significantly higher or lower than 200 MJ (we'll do this later). For such a two-sided test we might reject H_0 on the grounds that z is large and negative, but this would obviously not lead us to think H_1 can be favoured.

Some practitioners decide to reject hypotheses at a fixed confidence level. If we reject at the 90% confidence level, this means we reject the null hypothesis if $p < 0.1$. So in the above example we would. If we choose the 95% confidence level we need $p < 0.05$ to reject, which is not met here. At what level one chooses to reject a hypothesis is a matter of choice and convention, and varies between disciplines. Often $p < 0.01$ or smaller is used. The p value is sometimes reported in terms of the equivalent number of sigma for a Gaussian distribution. We just saw that $p = 0.067$ corresponds to 1.5σ for a one-sided test. Note that 2, 3, 4, and 5 times σ correspond to p values of 2.3×10^{-2}, 1.3×10^{-3}, 3.2×10^{-5}, and 2.9×10^{-7} respectively, for a one-sided test. A level of 5σ or higher is often used in physics. The computed value of p should always be reported in addition to a reject/no-reject decision, so others can make their own decision.

We return to the example. Instead of one battery we are now given $N = 4$ batteries and measure their mean capacity to be $\overline{x} = 230\,\text{MJ}$ (same value as before). What do we now say about the company's claim?

As we have more batteries the given mean capacity must be more secure, i.e. it must have a standard deviation less than 20 MJ. Given what we learned in section 2.3 from the central limit theorem about repeated measurements, we see that the standard deviation in $\overline{x}$ is a factor of $\sqrt{N}$ smaller than the single battery standard deviation (which is given to us; we don't have to infer it from the data). So the z value is now

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{N}} = \frac{230 - 200}{20/\sqrt{4}} = 3 \tag{10.5}$$

and we can once again calculate the p value – the probability that a sample of four batteries drawn from the original distribution would have a capacity of 230 MJ or more – which is

```
1 - pnorm(3) = 0.0013
```

[3] Would you accept a hypothesis that you haven't actually tested?

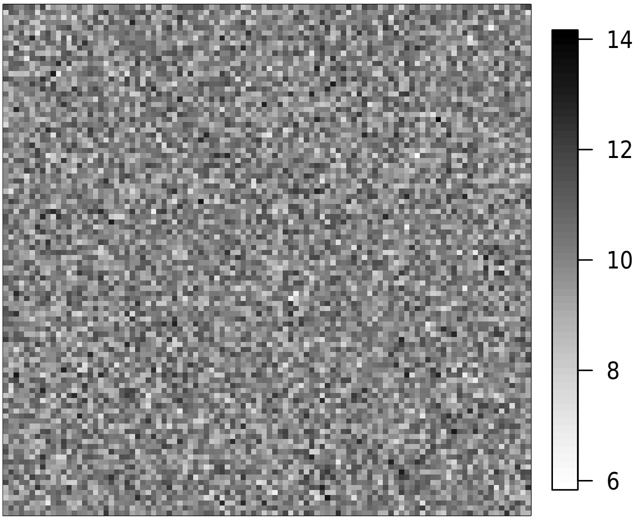

Fig. 10.2 An image with 100×100 pixels with a flux level in each pixel generated from a standardized Gaussian distribution $\mathcal{N}(10, 1)$.

As this is smaller than the value of 0.067 from before, we can now be more confident about rejecting the null hypothesis.

To use this test it's actually not necessary for the parent distribution of the data $\{x\}$ to be Gaussian. Given the definition of z, it is sufficient if the sample mean $\overline{x}$ has a Gaussian distribution. Provided the central limit theorem applies to the parent distribution, then in the limit of large N its sample mean will have a Gaussian distribution.

Example: detecting a signal in an image

Suppose we want to look for a signal (a single bright pixel) in a digital image in the presence of a noisy but otherwise uniform background illumination. An example is shown in figure 10.2. In the absence of a signal, the flux in the pixels will vary around the mean background level due to noise. Given a model for this background+noise – here a Gaussian with mean $\mu = 10$ and standard deviation $\sigma = 1$ – we want to set a threshold above which it is unlikely that the noise would give rise to such a large flux. A 5σ detection, for example, means that the probability of the noise producing a flux of 15 (or brighter) in any one pixel is $p = 2.9 \times 10^{-7}$ (`pnorm(-5)`). In an image with $N = 100 \times 100$ pixels, the probability that at one or more of the pixels would be this bright due to noise is $1 - (1-p)^N = 0.003$. So if we did measure a pixel flux at 5σ in an image of this size, the standard response is to say that this is unlikely to be a result of noise (or more correctly, of *this* noise model). We may then either reject such a measurement by claiming it to be an outlier – an implicit admission that our noise model is inferior[4] – or claim that it is due to something of interest. One scientist's noise is another scientist's signal!

The following code makes figure 10.2. It looks better if you plot it with a diverging colour scale, which you can achieve by replacing `Greys` with `BrBG` in the definition of `mypalette`.

R file: `noisy_image.R`

```
##### Plot a Gaussian noise image

library(fields) # for image.plot
library(RColorBrewer) # for colorRampPalette
mypalette <- colorRampPalette(brewer.pal(9, "Greys"), space="rgb",
                              interpolate="linear", bias=1)
mycols <- mypalette(64)
pdf("noisy_image.pdf", 5, 4)
par(mfrow=c(1,1), mgp=c(2.0,0.8,0), mar=c(1,1,1,1), oma=0.1*c(1,1,1,1))
set.seed(100)
x <- 1:1e2
y <- 1:1e2
z <- matrix(data=rnorm(length(x)*length(y))+10, nrow=length(y),
            ncol=length(x))
image.plot(z=z, x=x, y=y, xaxt="n", xlab="", yaxt="n", ylab="", nlevel=1024,
           zlim=c(-4.1,4.1)+10, col=mycols)
dev.off()
```

10.2.2 Student's t test

One-sided test

Consider a subtle but important variation of the battery problem variation on page 208.

We now measure the capacity of four batteries (which were all produced by the same technology) to have a mean of $\overline{x} = 230\,\text{MJ}$ and a standard deviation of $\hat{\sigma} = 16\,\text{MJ}$. How likely is it that they come from a technology which produces batteries with a mean capacity of $\mu = 200\,\text{MJ}$?

This appears to be the same problem: we take the same null and alternative hypotheses and also want a one-sided test. But now we have estimated the standard deviation from the sample itself, by calculating the sample standard deviation, $\hat{\sigma}$ (equation 2.12). Before we just took the manufacturer's quoted value, whereas now we aren't given one. We can still write down something which looks like the z value, but it is now called the t value or t statistic

$$t = \frac{\overline{x} - \mu}{\hat{s}} \tag{10.6}$$

because $\hat{s}$ is now an estimate *obtained from the data* of the standard deviation in $\overline{x}$. This estimate is the standard error in the mean, $\hat{s} = \hat{\sigma}/\sqrt{N}$ (equation 2.13). Because there is

[4] A Gaussian noise model is unlikely to apply to this kind of problem in reality. There are invariably additional noise sources which occur only rarely but can have a large impact. Examples are hot pixels in the detector, scattered light, and interference in the read-out electronics. In practice we would want to use a noise model with heavier tails, or a mixture model to handle large, occasional noise contributions (cf. section 9.3).

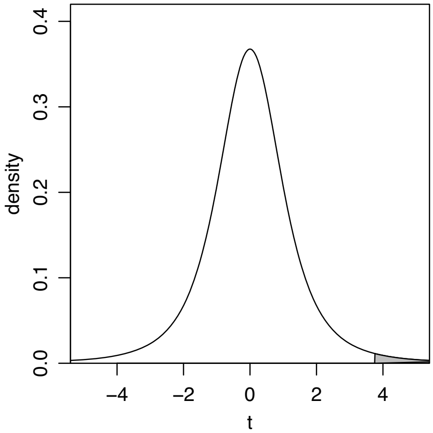

Fig. 10.3 The t distribution with three degrees of freedom. The grey region indicates $t > 3.75$, which has a probability of 0.017.

uncertainty in the value of $\hat{s}$ itself, the significance of a given deviation between $\overline{x}$ and μ is now less than when s was known. The degree of uncertainty in $\hat{s}$ depends on the degrees of freedom (see section 4.3). Whereas the z value in equations 10.3 and 10.5 followed a standardized Gaussian distribution, which has no free parameters, it can be shown that t follows a Student's t distribution. This has one free parameter, the degrees of freedom ν. This is equal to $N - 1$ here, since one degree of freedom is "used up" to estimate $\hat{s}$.

We encountered the Student's t distribution in section 6.2: it is the marginal posterior PDF for the mean of a Gaussian distribution given N data points when the standard deviation is unknown (and assigned a Jeffreys prior). Its form is given in equation 6.12 and shown in figure 6.2. It is similar in shape to the standardized Gaussian but has more probability in the tails; this reflects the extra uncertainty arising from having to estimate the standard deviation from the data. As the number of degrees of freedom tends toward infinity the distribution becomes a Gaussian. Equivalently, the standard deviation of the data would then be estimated exactly, so the conditions for the z test are fulfilled. Note, therefore, that if we have enough data we can just use the Gaussian z test rather than the t test.

Let me be more precise about the definition of the t statistic in equation 10.6. The quantity $\hat{s}$ is actually the standard deviation of $\overline{x} - \mu$. But if μ is given (a constant) then its standard deviation is zero,[5] so $\hat{s}$ is the standard deviation of $\overline{x}$. This is the case when we test the hypothesis that the data were drawn from a Gaussian with some specified mean μ. The Student's t distribution is then the distribution of the sample mean relative to the specified mean, scaled by the standard error in the sample mean.

Let us return to the battery question. $N = 4$, so the relevant t statistic is

$$t = \frac{\overline{x} - \mu}{\hat{\sigma}/\sqrt{N}} = \frac{230 - 200}{16/\sqrt{4}} = 3.75 \tag{10.7}$$

[5] In section 10.3.2 we will compare two samples, in which case μ is not constant and the denominator includes the standard deviation in both terms in the numerator.

and this follows a t distribution with $N - 1 = 3$ degrees of freedom. This is plotted in figure 10.3. The probability of getting a t value this large or larger – the p value – is $P(t \geq 3.75 \,|\, H_0)$. This is given by integrating the t distribution. This is done in R with the function `pt`, which is analogous to `pnorm` for the Gaussian distribution. The area under the curve is

```
1 - pt(3.75, df=3) = 0.017
```

Thus the probability that we would get a mean battery capacity as large as we did (or larger), given these four samples, is quite small ($<$2%) when the null hypothesis is true. So we can reject the null hypothesis at a 98% confidence level (but not at a 99% level). Compare this with `1 - pnorm(3.75)` $= 8.8 \times 10^{-5}$, which is much smaller. That is, if we *knew* that $s = 16$ MJ, rather than having to measure it from the data, we could be much more confident about rejecting the null hypothesis, because the same measured deviation from the expected mean is then more significant.

Two-sided test

Remaining with the same data, we now ask not whether the mean of the sample of four batteries has a larger capacity than that claimed by the old technology, but whether this mean is *different*, i.e. larger *or* smaller. Thus we test the null hypothesis
$H_0 : \mu = 200$ MJ
for which the alternative hypothesis is now
$H_1 : \mu \neq 200$ MJ
i.e. we use a two-sided test. The alternative hypothesis is now the complement of the null hypothesis.

The test statistic is the same as before, but we now calculate the probability that t is outside of the interval -3.75 to $+3.75$ (assuming the null hypothesis to be true). As the t distribution is symmetric, the probability is

```
2*(1-pt(3.75, df=3)) = 0.033
```

This is the probability that the data are consistent with the null hypothesis (more strictly: the probability of getting the measured capacity or something more extreme, assuming the null hypothesis is true). The probability with the two-sided test is twice as high as with the one-sided test, because we ask whether the mean differs from the specified value, rather than just whether it is larger. So for given data, the result of a two-sided test will be a larger p value. It is a weaker test.

10.2.3 General t testing in R and frequentist confidence intervals

The above is all you need for doing t tests with a single sample. Yet R provides a more convenient interface via the function `t.test`, which works directly on the original set of data. Suppose we have the following data set on batteries

```
x <- c(177, 194, 209, 228, 229, 235, 241, 244, 244, 287)
```

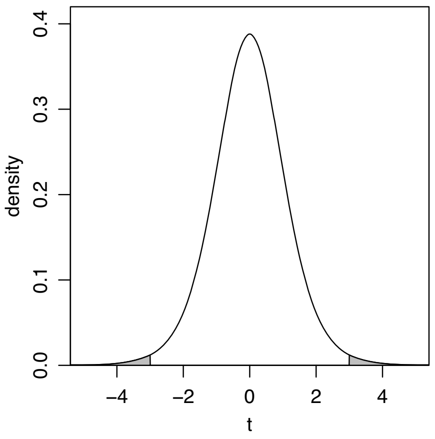

Fig. 10.4 The t distribution with nine degrees of freedom. The grey regions are those with $t < -2.99$ and $t > 2.99$, which together have a probability of 0.015.

We want to know whether their mean deviates from 200 (a two-sided test). We do

```
t.test(x, mu=200)
```

which produces

```
      One Sample t-test

  data:  x
  t = 2.9947, df = 9, p-value = 0.01508
  alternative hypothesis: true mean is not equal to 200
  95 percent confidence interval:
   207.0451 250.5549
  sample estimates:
  mean of x
      228.8
```

The output is largely self-explanatory. The t value is

$$t = \frac{\overline{x} - 200}{\hat{\sigma}/\sqrt{N}} \tag{10.8}$$

which you can check with R

```
(mean(x)-200)/(sd(x)/sqrt(length(x)))
```

The number of degrees of freedom ν is $N-1 = 9$. The p value is `2*pt(-2.9947, df=9)`, and is the probability of getting a t value of this magnitude $|t|$, or more, assuming the null hypothesis to be true. The fact that it's quite small (see figure 10.4) suggests that the sample mean is significantly different from the mean specified by the null hypothesis.

The frequentist 95% confidence interval (quoted in the output) means that if we repeated this experiment many times, 95% of these confidence intervals (which would differ each time) would contain the true mean.[6] This interval is symmetric about the estimated mean because we're doing a two-sided test. We can calculate the interval using the quantile function for the t distribution $Q(p;\nu)$, which is the inverse of the cumulative distribution function (see section 1.5). $Q(p;\nu)$ gives the value of t below which the distribution contains probability p. In R it can be computed by the function `qt(p,df)`, where `df` $= \nu$. By setting $p = 0.025$ we get the lower bound on the t distribution corresponding to the lower end of the 95% confidence interval. To scale this into a deviation from $\overline{x}$, it follows from equation 10.6 that we must multiply $Q(p;\nu)$ by $\hat{\sigma}/\sqrt{N}$. Thus the lower bound on the 95% confidence interval relative to $\overline{x}$ is

```
qt(p=0.025,df=9)*sd(x)/sqrt(length(x))
```

which gives -21.8. As the t distribution is symmetric (so `qt(1-p,df) = -qt(p,df)`), the upper bound on the 95% confidence limit (at $p = 0.975$) is at $+21.8$ relative to $\overline{x}$. Thus the estimate of the mean and its 95% confidence interval is

$$\overline{x} \,\pm\, Q(p = 0.975, \nu{=}N{-}1)\,\frac{\hat{\sigma}}{\sqrt{N}} \tag{10.9}$$

which gives 228.8 ± 21.8, which is the range $(207.0, 250.6)$. This is the output from `t.test` above.

The default for `t.test` is a two-sided test. To do a one-sided test, we specify

```
t.test(x, mu=200, alternative="greater")

   One Sample t-test

  data:  x
  t = 2.9947, df = 9, p-value = 0.007542
  alternative hypothesis: true mean is greater than 200
  95 percent confidence interval:
   211.1712      Inf
  sample estimates:
  mean of x
      228.8
```

The parameter `alternative` in the function call defines the direction of the one-sided test. The p value here is the probability that we get a value of t equal to or greater than the one measured when the null hypothesis is true. This defines the alternative hypothesis: the mean is more than 200. As we get a sample with mean 228.8 and a standard error in this mean of 9.6 (`sd(x)/sqrt(length(x))`), then it's more likely that these data are consistent with the alternative hypothesis.

We can also do the one-sided test the other way. The p value is now the probability that we get a value of t equal to or less than the one measured when the null hypothesis (which

[6] This does *not* mean that there is a 95% chance that the true mean falls within the specific confidence interval measured. That is given by the Bayesian confidence interval, sometimes also called the credible interval (section 5.5). The distinction may appear subtle, but it is both substantial and important. A frequentist analysis makes statements about possible data. In this paradigm unobserved parameters do not have uncertainty (a distribution), so the frequentist cannot say anything about the probability that the true value lies within a specified range.

says the mean is 200) is true. This is of course quite high. The new alternative hypothesis is the complement of the previous one, so the p value is one minus what we got before

```
t.test(x, mu=200, alternative="less")

One Sample t-test

  data:  x
  t = 2.9947, df = 9, p-value = 0.9925
  alternative hypothesis: true mean is less than 200
  95 percent confidence interval:
       -Inf 246.4288
  sample estimates:
  mean of x
      228.8
```

to within the rounding error, at least.

Note that one of the confidence limit bounds in the one-sided test will always be infinite, because by construction we allow arbitrarily extreme values in one direction.

10.2.4 Summary: z test vs t test

In both tests we have a set of N measurements and compute their mean to be $\overline{x}$. We want to compare this to a mean μ specified by the null hypothesis.

We use the z test when we additionally know that the true standard deviation of each measurement is σ. Thus the standard deviation in the mean is $\sigma/\sqrt{N}$, and the resulting z statistic, $(\overline{x} - \mu)/(\sigma/\sqrt{N})$ has a standardized Gaussian distribution.

In the t test we do not know the standard deviation. We therefore estimate it from the data to be $\hat{\sigma}$, so the standard deviation in the mean is $\hat{\sigma}/\sqrt{N}$. The resulting t statistic $(\overline{x} - \mu)/(\hat{\sigma}/\sqrt{N})$ has a Student's t distribution with $N - 1$ degrees of freedom.

10.3 Two-sample hypothesis tests

10.3.1 Gaussian z test

In the previous section we compared statistics estimated from a sample of data to fixed values. Sometimes we want to compare two samples, for example to find out whether two samples have the same mean, or whether two samples with the same mean have different variances. In each case we must identify what knowledge we have a priori and what statistic we can and want to determine from the data. We can then establish the distribution of the statistic and identify the appropriate test to apply.

Let's take an example. We measure the boiling points of two liquids to be $127 \pm 5\,\mathrm{K}$ and $141 \pm 3\,\mathrm{K}$. Are they significantly different?

The null hypothesis is that they are the same. We interpret the first measurement as saying that the boiling temperature is a Gaussian random variable with mean $\overline{x}_1$ and standard deviation σ_1, and similarly for the second liquid. Our task is to assess whether $\overline{x}_1 - \overline{x}_2$ is significantly different from zero. The difference of two Gaussian variables is another Gaussian variable with mean $\overline{x}_1 - \overline{x}_2$ and variance $\sigma_d^2 = \sigma_1^2 + \sigma_2^2$. The standard deviations are given, and not estimated from the data, so the z test is appropriate, with statistic

$$z = \frac{141 - 127}{\sqrt{3^2 + 5^2}} = \frac{14}{5.83} = 2.40. \tag{10.10}$$

The boiling points differ by $2.4\sigma_d$, which corresponds to the following p value (two-sided test).

```
2*(1-pnorm(2.4)) = 0.016
```

There is (just) a 1.6% chance that we would observe these two boiling points if they came from the same distribution. The complementary alternative hypothesis is that the boiling points differ, for which there is reasonable support. This conclusion assumes that the errors are indeed Gaussian. If they are not, then the z test does not apply.

10.3.2 Student's t test

We now turn to the more common case where we have to estimate the mean and standard deviation of each distribution from the data themselves. Suppose we have two samples, $\{x_1\}$ and $\{x_2\}$, with N_1 and N_2 members each. We first estimate the means of the two samples, $\overline{x}_1$ and $\overline{x}_2$, as well as their (sample) standard deviations, $\hat{\sigma}_1$ and $\hat{\sigma}_2$. In analogy to what we did earlier, the test statistic is a standardized difference between the two means

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\hat{\sigma}_c\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \tag{10.11}$$

where $\hat{\sigma}_c$ is the standard deviation of the combined samples. This is given by a weighted mean of the individual sample variances

$$\begin{aligned} \hat{\sigma}_c^2 &= \frac{\nu_1\hat{\sigma}_1^2 + \nu_2\hat{\sigma}_2^2}{\nu_1 + \nu_2} \\ &= \frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 - 2} \end{aligned} \tag{10.12}$$

where $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$ are the number of degrees of freedom in the samples (each minus one is there because we are estimating the standard deviations from the data). This t statistic is appropriate if the samples have been drawn from two parent populations which have a common (but unknown) standard deviation, which we are estimating with $\hat{\sigma}_c$. (We don't assume the means are the same, of course). The denominator in equation 10.11 can be thought of as a generalization of the standard error in the mean. The above t statistic has a t distribution with $N_1 + N_2 - 2$ degrees of freedom. The -2 is there because we are calculating two standard deviations from the data.

If the two samples do not have a common standard deviation (or rather, if we relax this assumption), then the statistic is different

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}. \tag{10.13}$$

This does not have a t distribution, but it can be approximated by one with a number of degrees of freedom given by

$$\nu = \frac{\left(\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}\right)^2}{\frac{(\hat{\sigma}_1^2/N_1)^2}{N_1-1} + \frac{(\hat{\sigma}_2^2/N_2)^2}{N_2-1}} \tag{10.14}$$

(you'll need to consult advanced classical texts for a proof of this). The statistical test for two samples using this statistic known as the *Welch two-sample t test*.

Needless to say, all of this can be done directly in R using `t.test`.

10.4 Hypothesis testing in linear modelling

We now return to chapter 4, and in particular to section 4.1.5, to see how the t test can be used in linear regression to assess the quality of a model fit.

We saw from equation 4.10 that the least squares estimate of the gradient of the line $\hat{b}$ (I now use the "hat" symbol to indicate it is an estimate) could be written as the weighted sum of several random variables, namely the $\{y_i\}$. The same can be shown for the estimate of the intercept, $\hat{a}$. If we assume – as we did – that each y_i has a Gaussian distribution with a common standard deviation, then the variable

$$t = \frac{\hat{b}}{\sigma_{\hat{b}}} \tag{10.15}$$

has a Student's t distribution with $N-2$ degrees of freedom, where $\sigma_{\hat{b}}^2 = \text{Var}(\hat{b})$ (given by equation 4.15). In fact, the definition of the t statistic (equation 10.6) is $t = (\hat{b} - b_{\text{hyp}})/\sigma_{\hat{b}}$, where b_{hyp} is the value of the gradient under the hypothesis we want to test. So the expression in equation 10.15 only has a t distribution when $b_{\text{hyp}} = 0$. Thus by applying a t test to the statistic in equation 10.15, we are testing the null hypothesis that the data have zero gradient. This may be used to assess how important the parameter is for the model. This test has $N-2$ degrees of freedom because we have "used up" two degrees of freedom in estimating a and b.

The `lm` function in R does this t test for all the parameters in the model. Specifically, it tests the null hypothesis that the particular parameter is zero, with all the other parameters held fixed at their fitted values. It prints the value of t and the resulting p value (see the code outputs in section 4.1.5).

In principle a small p value means the parameter is significant, i.e. we need it in the model to fit the data well. But this is only the case if we keep the other parameters fixed at their fitted values. Look back at the regression problem with nonlinear functions of the data in section 4.6. In the quadratic fit all the parameters were highly significant. This suggests that adding a cubic term would improve the fit. The cubic fit is indeed better – in terms of a lower residual standard error – but we see from the code output that the significance of the quadratic term has now dropped (larger p value). But this does not necessarily imply the quadratic term is not required. (Indeed it cannot mean this, because the cubic term has an even lower significance than the quadratic one.) The reason is that the t test on the quadratic parameter is applied by keeping the cubic parameter fixed at its fitted value. So we cannot fit a multiparameter model and identify the significance of each parameter using just the results of the t tests on this model. To find out if the quadratic or cubic term is better, we need to fit separate models. Specifically, we need to compare the results of `model2`, $y = b_0 + b_1x + b_2x^2$, with $y = b_0 + b_1x + b_3x^3$, which produces

```
lm(formula = y ~ x + I(x^3))
Residuals:
     Min       1Q   Median       3Q      Max
-2.15365 -0.75737 -0.00452  0.85481  2.28427
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.4897920  1.0350118  18.831 2.82e-10 ***
x           -1.8180150  0.2130960  -8.531 1.93e-06 ***
I(x^3)       0.0040257  0.0008941   4.503 0.000723 ***
Residual standard error: 1.355 on 12 degrees of freedom
Multiple R-squared:  0.9265,Adjusted R-squared:  0.9142
```

The quadratic fit is marginally better in terms of a lower residual standard error and a higher significance (smaller p value) on its nonlinear term. So on balance we need a quadratic term more than a cubic one. Having both (`model3`) improves the fit further, but only marginally.

10.5 Goodness of fit and the χ^2 distribution

When we fit a model to data, we don't usually expect (or want) a perfect fit, because data are noisy. But we do want to get an idea of how good the fit is. One measure of this is the χ^2 statistic. It is defined as

$$\chi^2 = \sum_{i=1}^{N} z_i^2 = \sum_{i=1}^{N} \left(\frac{y_i - f_i}{\sigma_i} \right)^2 \qquad (10.16)$$

(which also defines z_i), where $\{y_i\}$ are the observed data values, $\{f_i\}$ are the corresponding values predicted by the model, and $\{\sigma_i\}$ are the standard deviations in the observed values, i.e. the uncertainties in the measurements. Sometimes these are known independently of the data. For example, when weighing objects the precision of the scales is often known in advance. We saw in section 4.1.4 how χ^2 could then be used as an error function to fit linear models, and we saw in section 4.4.2 that this is equivalent to maximizing the likelihood for known Gaussian uncertainties. In other situations the $\{\sigma_i\}$ must be estimated from the data (and we saw in section 9.1 that this can be done by computing the posterior).

The difference $y_i - f_i$ is the residual. If the residuals are typically of the same size as the

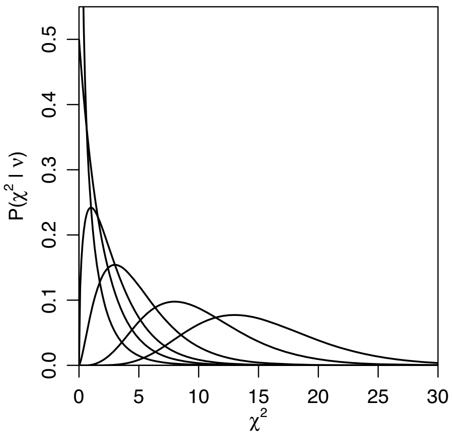

Fig. 10.5 The χ^2 distribution for 1, 2, 3, 5, 10, and 15 degrees of freedom. The higher the number of degrees of freedom, the further to the right is the curve.

expected error – suggesting a reasonable fit to the data – then $z_i \sim 1$ and χ^2 will be about N, the number of data points. If the residuals are much larger than the expected errors, then this suggests a poorer fit, so χ^2 will be much larger than N. If χ^2 is much smaller than N, this means the residuals are smaller than the expected errors. In that case either the fit is "too good" – the model is too finely tuned to the exact data and so is fitting the noise, something known as *overfitting* – or the errors have been overestimated. Clearly there will be a lot of scatter in the calculated value of χ^2 when N is small, so how much χ^2 can tolerably deviate from N depends on both N and the noise model.

The χ^2 statistic is useful for assessing the quality of fit for a given model. But it may not be used to find the best model among a set of models. I will explain why in chapter 11.

10.5.1 The χ^2 distribution

In equation 10.16, χ^2 can be calculated for any data with any distribution. However, if each z_i is an independent, standard Gaussian random variable – $\mathcal{N}(0, 1)$ – then χ^2 follows the χ^2 distribution with integer parameter $\nu = N$ (we shall prove this in section 10.5.3). Its density function is

$$P(\chi^2; \nu) = \frac{2^{-\frac{\nu}{2}}}{\Gamma(\nu/2)} (\chi^2)^{\frac{\nu}{2}-1} \exp(-\chi^2/2) \quad \text{where} \quad \nu \geq 1. \tag{10.17}$$

where $\Gamma()$ is the gamma function (equation 1.62). Examples of the distribution for different ν – the degrees of freedom – are shown in figure 10.5. The distribution has mean ν and variance 2ν. For $\nu = 1$, $P(\chi^2) \to \infty$ as $\chi^2 \to 0$. For $\nu = 2$ the mode is at $\chi^2 = 0$ and the distribution is just a negative exponential. For larger values of ν the distribution has a mode at larger values of χ^2 and becomes increasingly symmetric; from inspection of equation 10.17 we see that it in fact becomes increasingly like a Gaussian. The χ^2 distribution is a special case of the gamma distribution (section 1.4.6) with scale $\theta = 2$ and shape $k = \nu/2$.

Note that the χ^2 distribution is a distribution in χ^2, not χ. The distribution for χ is obtained by a change of variables, $P(\chi^2)d\chi^2 = P(\chi)d\chi$, to give $P(\chi) = 2\chi P(\chi^2)$, which is also parametrized by ν. For $\nu = 3$ the χ distribution is the Maxwell–Boltzmann distribution, which is the probability density function used to describe the speed $v = (v_x^2 + v_y^2 + v_z^2)^{1/2}$ of particles in a gas in three-dimensional space: $P(v) \propto v^2 \exp(-av^2/2)$ for some constant a. This follows because the velocity distributions of particles in each spatial dimension are assumed to be independent zero mean Gaussians (with a common standard deviation). In two dimensions ($\nu = 2$) this is the Rayleigh distribution, $P(\chi) \propto \chi \exp(-\chi^2/2)$.

10.5.2 Using χ^2 as a measure of goodness of fit

Suppose we have a linear model for some data which we believe is the true model, and the measurements differ from this only due to zero mean Gaussian noise. For point i the model predicts the value f_i and the expected standard deviation is σ_i. In that case each term $z_i = (y_i - f_i)/\sigma_i$ has a standard Gaussian distribution, and if the N measurements are independent, χ^2 will follow the χ^2 distribution with N degrees of freedom. This is the case when the model has been decided upon independently of these data. If we have instead fit the model to the data, then the residuals in this fit are no longer independent, so the degrees of freedom is reduced. If our (linear) model has p parameters, then χ^2 follows a χ^2 distribution with $\nu = N - p$ degrees of freedom. This last statement involves three important conditions.

(1) The residual are assumed to follow a Gaussian distribution. This is often not achieved in practice, for example due to outliers.
(2) The σ_i values are assumed to be known: they have not been estimated from the data. Estimating them would effectively reduce the number of degrees of freedom further, but the metric χ^2 would no longer have a χ^2 distribution.
(3) The model is linear in the parameters (see section 4.2). If it is not, then the concept of p degrees of freedom being "used up" by estimating p parameters does not apply (section 4.3).

Let's now use the χ^2 distribution to calculate a goodness of fit. We fitted a straight line to some data in 4.1.5. Suppose that the uncertainty in all the y-values is 2. We compute χ^2 as follows.

```
resid <- y - predict(model1)
err   <- 2
chi2  <- sum((resid/err)^2)
df    <- length(x)-2  # lose 2 dof due to the straight line fit
pchisq(chi2, df)
```

This gives $\chi^2 = 16.7$ on eight degrees of freedom. The last statement in the code above returns the probability that χ^2 has the specified value or less. This probability is 0.967. Let the null hypothesis H_0 be the statement that the model is the true one, i.e. the predictions agree with the measured data to within the expected errors. When H_0 is true we expect low χ^2 values (around 1 per degree of freedom). The probability that we would get a χ^2 as high as the one observed (or even higher), the p value, is $1-$`pchisq(chi2,df)` = 0.033. This suggests that it is unlikely we would observe such a large value of χ^2 for this number of degrees of freedom: either the model is a poor fit to the data, or we've underestimated the expected errors, or the data are erroneous (e.g. there are outliers). Note that even if we can rule out the last two alternatives, a low p value still does not "prove" the model to be wrong. All we can do is find an alternative (but still plausible) model which explains the data better (which for a frequentist can only be achieved by failing to reject it).

You will sometimes see reference to the *reduced chi-squared*. This is simply χ^2 divided by its expectation value (the number of degrees of freedom)

$$\chi_r^2 = \frac{\chi^2}{\nu} = \frac{1}{\nu}\sum_{i=1}^{N}\left(\frac{y_i - f_i}{\sigma_i}\right)^2 . \tag{10.18}$$

You may encounter the rule of thumb "a fit is good if the reduced χ^2 is about 1". Although the mean of χ^2/ν for ν degrees of freedom is 1, the shape of the distribution still depends on ν.

We should not forget that χ^2 is only a statement about how likely the data are assuming the model to be true. It tells you nothing directly about how likely the model is given the data. The reason is that the probability of any particular data set – even from the true model – is generally very low (recall the discussion in section 5.2). Thus χ^2 can be large and the p value small even when the null hypothesis is true. Think of N cards drawn (with replacement) at random from a pack of cards. The probability of the data set is $1/52^N$ no matter which cards we draw. If some combinations of cards *seem* less likely (e.g. four aces), then this is only because we attach more meaning to these than to other combinations. We will discuss this further in section 11.1.

10.5.3 Derivation of the χ^2 distribution

I have not yet proven that the χ^2 distribution is the PDF of the sum of the squares of ν independent Gaussian variables $\{z_i\}$ each with zero mean and unit variance. This is quite straightforward.

Think of the variables as defining a point in a ν-dimensional space with coordinates $(z_1, z_2, \ldots, z_\nu)$. The PDF of each variable is proportional to $\exp(-z_i^2/2)$ so their joint PDF is their product

$$P(\{z_i\}) \propto \prod_i \exp\left(-z_i^2/2\right) = \exp\left(-\sum_i z_i^2/2\right) = \exp(-\chi^2/2). \tag{10.19}$$

In this ν-dimensional space a hypersphere with radius χ is a surface of constant probability. The probability that a point lies in a hypershell between χ and $\chi + d\chi$ is the above probability density (which is per unit volume) multiplied by the volume of this hypershell. The hypershell has surface area proportional to $\chi^{\nu-1}$ and thickness $d\chi$, so its volume is proportional to $\chi^{\nu-1}d\chi$. Thus

$$P(\chi)d\chi \propto \chi^{\nu-1}\exp(-\chi^2/2)d\chi. \tag{10.20}$$

To get $P(\chi^2)$ we do a transformation of variables

$$\begin{aligned} P(\chi^2) &= P(\chi)\frac{d\chi}{d\chi^2} \\ &\propto \frac{1}{\chi}\chi^{\nu-1}\exp(-\chi^2/2) \\ &\propto (\chi^2)^{\frac{\nu}{2}-1}\exp(-\chi^2/2) \end{aligned} \tag{10.21}$$

which is the χ^2 distribution (equation 10.17) to within a normalization constant.

A useful property of the χ^2 distribution follows from this derivation. If two variables x_1 and x_2 are described by χ^2 distributions with ν_1 and ν_2 degrees of freedom, then their sum, $x_1 + x_2$, must follow a χ^2 distribution with $\nu_1 + \nu_2$ degrees of freedom.

10.6 Issues with frequentist hypothesis testing

I dedicated this chapter to frequentist hypothesis testing because it is widely used, not because I think it is a sound way to test models. It has some significant problems, some of which I have already alluded to. I will now summarize these, before discussing some more fundamental issues on model inference.

(1) Frequentist hypothesis testing requires us to summarize the data via a test statistic and then find a convenient probability distribution for the statistic. This often involves significant approximations in practice.

(2) The approach hinges on calculating the probability that the hypothesis could produce a value of the statistic which is more extreme than the observed value of the statistic (this probability is the p value). It is not clear why the concept of imaginary "more extreme" data should be relevant to our inference.

(3) Although not a fault with the approach in itself, p values are frequently interpreted as being the probability that the hypothesis is true. This is wrong. The p value tells you nothing about the probability of the hypothesis; this concept is absent in the frequentist approach.

(4) A hypothesis is rejected on the basis that the data obtained were unlikely. This is odd, because we *did* get these data. By rejecting hypotheses in this way we ignore the possibility that the data may be equally unlikely under other plausible hypotheses.

(5) Rejecting the null hypothesis on the basis of a small p value does not imply that another (plausible) hypothesis exists which could explain the data better. In particular, rejecting the null does not imply we can accept some particular alternative. We should surely only accept a hypothesis if it has been explicitly tested. Yet in the frequentist paradigm we can never accept hypotheses, just fail to reject them (at some confidence level).

(6) The p values are sensitive to the size of a data set. The more data we have, the less probable any particular data set – even under the true model – and so the smaller the p value (this was seen in section 5.2). This makes it increasingly likely that we will reject a given model (null hypothesis) as we get more data. We could remedy this by raising

the rejection threshold as we get more data. But with complex problems in particular we cannot know how to set this threshold, and in practice fixed confidence levels (e.g. 95%, 99%, 3σ, 5σ) are used.

(7) As will be demonstrated in section 11.8, p values can depend on irrelevant information, which can nonetheless make the difference between rejecting and not rejecting a hypothesis.

(8) The relevance of a p value depends on what other tests have been done. Suppose we wanted to test whether there was a significant correlation between the annual per capita deaths by leukaemia, and each of a number of metrics, such as income, mobile phone use, number of doctors, etc. For each individual correlation comparison we might set $p = 0.05$ as the significance level for declaring a significant correlation. But if we did 20 such tests then we would expect to get one[7] significant correlation just by chance – indeed, by definition of the significance level. Yet if we reported just this one positive result it would appear to have statistical significance. This problem is known as *multiple testing*, more informally as *p hacking*. One can try to compensate for it by adjusting the significance level according to the number of tests made (e.g. the Bonferroni correction), or by looking at the fraction of false positive results among all tests made (e.g. with the false discovery rate).

Having highlighted some specific problems, I now turn to a more fundamental issue concerning frequentist hypothesis testing.

A popular impression of science is that it works by falsification. The idea is that a scientist specifies a hypothesis which makes a clear prediction. If an observation or measurement is made which contradicts this prediction, the hypothesis is falsified. A much used analogy is that of the black swan. The hypothesis is that all swans are white; the observation of one black swan falsifies this hypothesis. Null hypothesis testing attempts to put rejection of the hypothesis on a numerical basis by using the p value to specify a rejection confidence.

It's a nice story, but science does not work like this in practice. First, we rarely observe directly what the hypothesis states. In reality we observe something (e.g. the motions of planets) in order to infer the truth (or perhaps just the utility) of the hypothesis (e.g. "the Sun and planets orbit the Earth"). Second, a hypothesis rarely makes an explicit binary prediction. We must instead set up a model which generally predicts the size of an effect, not simply its presence or absence. Third, measurements are noisy (or are a sample from a population), so we cannot say with certainty whether a prediction is correct; we just get a measure of proximity between data and model predictions. The black swan is a poor caricature of science, and by sticking to it we would have to ask what we mean by "black" and indeed what we mean by "swan". What do we conclude if we observe a grey swan, or a bird that looks much like a swan but has small genetic differences? Testing models is not black and white (excuse the pun). Fourth, a model is only ever an approximation to reality. All models are therefore false at some level, so what does it even mean to falsify a model? If we required models to be true, we would end up rejecting them all. Yet science is full of approximate models which are very useful in practice: the model of gases

[7] The probability of getting at least one significant correlation by chance from 20 tests is $1 - (1 - 0.05)^{20} = 0.64$.

as non-interacting particles with zero volume; classical electromagnetism; Newton's law of gravity. All of these are false models of reality, yet they are very accurate in certain domains.

In the real world we cannot falsify a hypothesis or model any more than we "truthify" it. We can only determine a degree (probability) of model accuracy, of which "true" and "false" are the two extremes. Our model predictions will never agree with measurements exactly. We must therefore ask *to what extent* measured data support a specified model. As we have seen in earlier chapters, this is not an absolute measure: even under the true model, the observed data can be very unlikely (e.g. section 5.2). Thus not only are we unable to reject a model in an absolute sense, we cannot even do it probabilistically in any meaningful sense. All we can do is ask which of the available models explains the data best. That is, we must *compare* models. How to do this is the subject of the next chapter.

You may now be wondering why, given all these problems, we ever bother to do frequentist hypothesis testing. As we already saw in section 3.2, Bayesian model comparison is more direct, because it can give you $P(M\,|\,D)$. Actually, it can only give probabilities of the models if we can specify the complete set of models. This is often not possible, although we can still calculate the posterior odds ratio for any two models. But what if we only have one model? Bayesian model comparison cannot then tell you anything. This is, in some senses, logical: if we don't believe there is an alternative model, then the data *must* have come from the one and only model. We may nonetheless have reasons[8] to believe this model is suboptimal, even if we don't yet have a concrete proposal for a better alternative. Frequentist hypothesis testing may be able to help here, because a sufficiently small p value could indicate that the data are too unlikely, thereby triggering the search for a new model. Indeed, such tests seem quite natural in some circumstances, as in the example of looking for a signal on top of a noisy background (page 209). Given the pitfalls of p values, and in particular the fact that even data drawn from the true model (in a simulated set up) can result in arbitrarily small p values, it is difficult to know how low the p value should be before we initiate the search for a new model. Yet it does seem that, in practice, scientists are sometimes motivated to search for new models by low p values. Of course, once we have established a new model, we cannot rely on p values to establish its "truth".

The Bayesian approach to model assessment, discussed in more detail in the next chapter, avoids many of the pitfalls of frequentist hypothesis testing, although it is not without its own limitations, as we shall see. Frequentist hypothesis testing is useful – and used – because it is easy to perform, and does not require us to specify alternative models. It can be used to get an indication of the plausibility of an existing sole model for some data, and thereby help decide whether we should search for a better alternative.

[8] The reasons may be unrelated to specific model predictions and data, and so transcend statistical considerations.