Chapter 6

Hypothesis tests

Contents

6.1 Introduction

So far, we have looked at estimating population parameters (means, variances, proportions), using both single point estimates and (confidence) interval estimates. Sometimes, there is special interest in whether or not a population parameter takes a particular value.

In this chapter, we will mainly consider the problems of:

- testing whether the mean μ of some population is equal to a specific value μ_0, given some observations $x_1, \ldots, x_n$ from that population;
- testing whether the means μ_X and μ_Y of two separate populations are equal to each other, given observations $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ from each population.

Motivating example

Consider again the cholesterol example in Exercise 5.3. To recap, we suppose a new cholesterol lowering drug has been developed. It is first tested in a small study: 10 patients are given the drug (with a daily dose over twelve weeks), and 10 patients are given a placebo (with patients unaware of whether they are receiving the active drug or the placebo or not). The outcome measure for each patient is the percentage reduction in LDL cholesterol, from the beginning to the end of the twelve week period.

We will ignore the placebo group for now. The data for the active drug are given below.

```
drug

##  [1] -25.7 -11.6   6.6 -28.7 -15.0 -12.3  -4.8 -17.1  11.8 -15.8
```

so our observations for the 10 patients given the active drug are $x_1 = -25.7\%, \ldots, x_{10} = -15.8\%$. Now let's suppose that for the population of patients taking this drug, the percentage reductions will follow a $N(\mu, \sigma^2)$ distribution. Of particular interest is whether $\mu = 0$ or not: if $\mu = 0$, there would be no change, *on average*, in patient's cholesterol levels when taking the drug, suggesting that the drug has no effect. Changes in individual patients' levels could be thought of as 'natural variation over time', that would have occurred *even if the drug had not been taken.*

So how do we decide whether $\mu = 0$, based on our observed data $x_1, \ldots, x_{10}$?

6.2 Setting up the null and alternative hypotheses

In a hypothesis test, we will specify a **null hypothesis**, denoted by H_0. In the cholesterol example, we state

$$H_0 : \mu = 0. \tag{6.1}$$

We will also specify an **alternative hypothesis** denoted by H_A. In the cholesterol example, one possibility would be a **one-sided alternative hypothesis** $H_A : \mu < 0$, implying that the drug *does* lower cholesterol levels on average. However, if the drug turned out to be harmful,

and actually *increased* cholestorol levels on average, we would want to know, so we will instead consider a **two-sided alternative hypothesis**

$$H_A : \mu \neq 0, \tag{6.2}$$

i.e. the alternative hypothesis is simply that the drug does have *some* effect, on average.

Our goal is to draw a conclusion about which hypothesis, the null H_0 or the alternative H_A, we think is correct, using the data we have observed.

6.3 A first attempt: hypothesis testing 'by eye'

Can we guess which hypothesis is correct, simply by plotting the data? A problem is that if we do not the population variance σ^2, then the variance of $\bar{X}$ is also unknown, and we won't know how far the sample mean *could* be from its expected value: the population mean μ. But perhaps there are other clues in the data that can point to which hypothesis is correct. We will consider the following:

1. What patterns might we see in a sample of ten random variables $X_1, \ldots, X_{10}$, drawn from the $N(0, \sigma^2)$ distribution (so that H_0 is true)?

2. What patterns might we see in a sample of ten random variables $X_1, \ldots, X_{10}$, drawn from the $N(\mu, \sigma^2)$ distribution, with $\mu \neq 0$ (so that H_A is true)?

3. Since we don't know what σ^2 is, are there *differences* between what we might see in (1) and (2), that *do not* depend on the value of σ^2?

In the next two Figures, we will plot samples of size 10 drawn from normal distributions, where in the first four plots, $\mu = 0$, and the second four $\mu \neq 0$. We will also change σ^2 in each case. For each random sample, we will compute the sample mean $\bar{x}$, the sample variance s^2, and the ratio $\bar{x}/s$, where, to recap,

$$s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2.$$

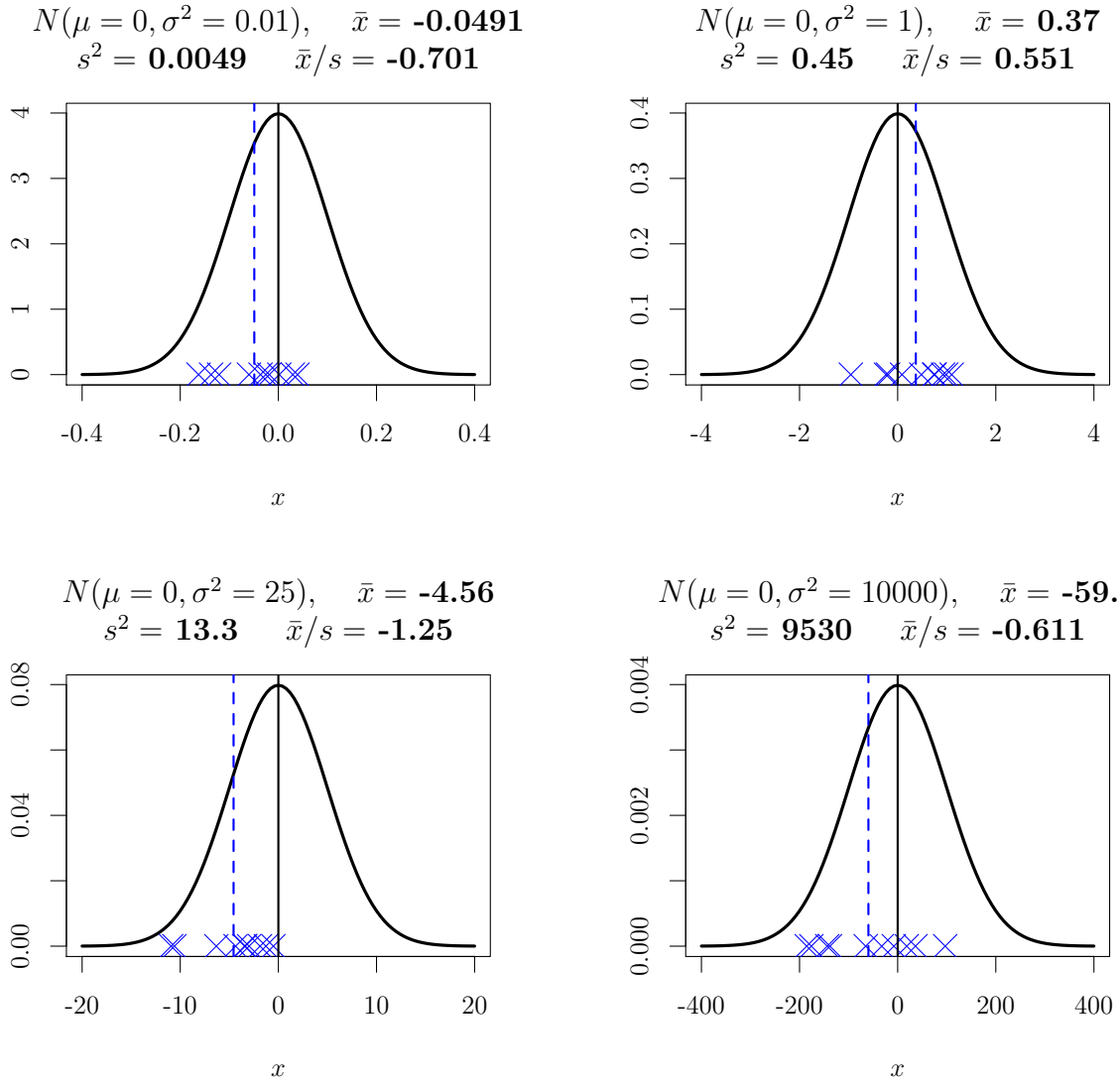


Figure 6.1: Random samples of size 10 (shown by the blue crosses) drawn from four different $N(0, \sigma^2)$ distributions, where the null hypothesis $H_0 : \mu = 0$ is **true**. The solid vertical line shows the population mean 0, and the dashed line shows the sample mean.

Patterns to spot:

1. In each case, we don't have $\bar{x} = 0$, even though $\mu = 0$, so simply observing $\bar{x} \neq 0$ isn't enough to tell us whether the null hypothesis is false or not.

2. Changing σ^2 has the effect of 'rescaling' the plots (changing the numbers on the axes), *but not the general appearance of each plot*: in each case we see $\bar{x}$ fairly close to 0, relative to the 'scatter' in the observed values.

3. The values of $\bar{x}$ and s^2 all quite different in the four plots, but the ratios $\bar{x}/s$ are all relatively similar.

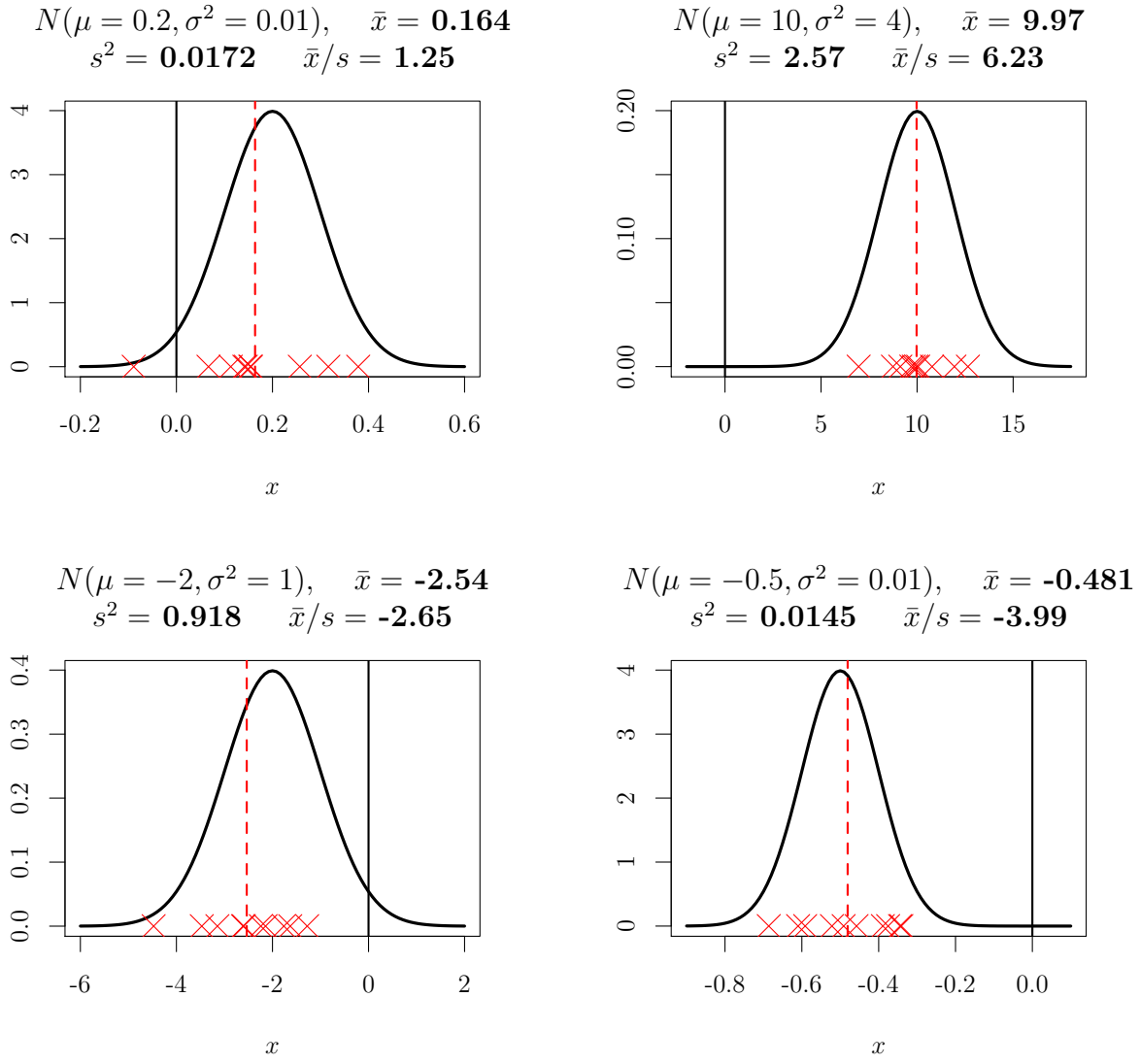


Figure 6.2: Random samples of size 10 (shown by the red crosses) drawn from four different $N(\mu, \sigma^2)$ distributions, where the null hypothesis $H_0 : \mu = 0$ is **false**. The solid vertical line indicates $\mu = 0$ under the null hypothesis, and the dashed line shows the sample mean.

Patterns to spot:

1. Even though we have $\mu \neq 0$ in each case, we don't necessarily have $\bar{x}$ further from 0 than we did in Figure 6.1. (The distance we get depends on σ^2.)

2. Again, changing σ^2 has the effect of 'rescaling' the plots, but not their appearance: we now see $\bar{x}$ further from 0, relative to the 'scatter' in the observed values; the observations can be **clustered** away from 0.

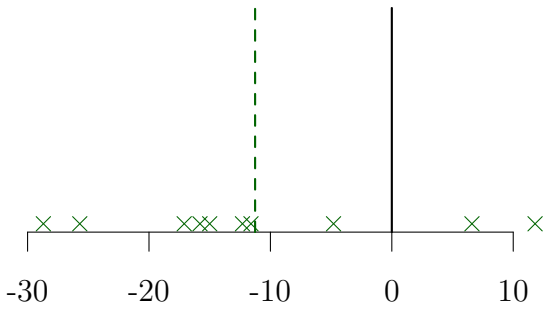
3. The ratios x/s are typically larger (in absolute value) compared with those in the previous figure.

Conclusions:

- From plotting the data, if the sample mean $\bar{x}$ is far from 0, relative to the variation in the data, i.e. the observations are "clustered" away from 0, that may suggest the null hypothesis $\mu = 0$ is false.
- Calculating $\bar{x}/s$ is way of quantifying "far from 0, relative to the variation in the data."

Now let's try a similar plot for the 10 cholesterol reduction observations. (We won't draw a density function, as we don't know what the true mean and variance are.)

$\bar{x} =$ **-11.3** $\bar{x}/s =$ **-0.88**



percentage change in LDL cholesterol

What should we conclude? It's not obvious either way, though perhaps the sample mean is a little far from 0, relative to the scatter in the data. We will need something more formal.

6.4 Formal hypothesis testing: two approaches

There are two main approaches to hypothesis testing[1] which we will refer to as

1. The Neyman-Pearson framework;
2. Fisher's p-value method.

They share much in common, but differ in how the results are reported. We will work through each in turn, illustrating them in the case of the cholesterol example.

6.5 The Neyman-Pearson framework

Here we suppose that a *decision* has to be taken, between the options of

- act as if H_0 is true - do not "reject" H_0;

[1] devised by Jerzy Neyman (1894-1981), Egon Pearson (1895-1980) and Sir Ronald Fisher (1890-1962): three of the most influential statisticians of the 20th century.

- act as if H_A is true - "reject" H_0.

In the example, to "act as if H_0 is true" may mean to declare that the drug does not work, perhaps abandoning further development, or not licensing it for further use. To "act as if H_A is true" may mean to declare that the drug *does* work and adopting for use in healthcare (although we need to check the drug effect is in the right direction: lowering rather than raising cholesterol.)

6.5.1 The Neyman-Pearson framework: step-by-step

We illustrate the Neyman-Pearson framework for the Cholesterol example (which will result in what is known as a **one-sample t-test**), but the framework can be applied for any hypothesis testing problem. Before collecting the data, we suppose we are going to observe random variables $X_1, \ldots, X_n$ (with $n = 10$ in our example).

1. State the null and alternative hypotheses H_0 and H_A. In our example, we have

$$H_0 : \mu = \mu_c, \tag{6.3}$$
$$H_A : \mu \neq \mu_c, \tag{6.4}$$

   with $\mu_c = 0$.

2. Choose the size of the test, denoted by α.

   The **size** of the test, α, is the probability of rejecting H_0 *if* H_0 *were true*, i.e., *mistakenly* rejecting H_0. Making this mistake is known as a **Type I error**. The size of the test is also referred to as the **level of significance**.

   We can choose α to be anything we like, but not too small: $\alpha = 0.05$ is fairly common in practice. Significance levels are typically given as percentages: $100\alpha\%$. (For more complex null hypotheses of the form $\mu \in R$ for some set R, the definitions of size and significance level are slightly different, but we won't consider this case in this module.)

   The aim of the Neyman-Pearson framework is to 'control' the Type I error rate. For example, if we (always) use $\alpha = 0.05$, then out of all occasions in which H_0 is true, we will make the wrong decision 5% of the time.

3. Choose a **test statistic**. This will be a function of the random variables $X_1, \ldots, X_n$.

   For a suitable **test statistic** we want increasingly large (absolute) values to be increasingly **unlikely** under H_0: we want large values to make us think the null hypothesis H_0 is **false**.

What function of the data should we use? In Section 6.3, we saw that calculating the sample mean $\bar{x}$ scaled by the sample standard deviation s could suggest whether the population mean was 0 or not, with large (absolute) values favouring the alternative. Here, we will use something similar:

$$T = \frac{\bar{X} - \mu_c}{S/\sqrt{n}}. \tag{6.5}$$

(With $\mu_c = 0$, we are almost using the same function of the data we did in Section 6.3, but now with an extra factor to allow for the sample size n.)

> With this choice of test statistic, we are conducting what is known as a **one sample t-test**.

4. Assume H_0 is true, and derive the distribution of the test statistic under H_0. With the choice of T above, if H_0 is true, then we have $X_1, \ldots, X_n \sim N(\mu_c, \sigma^2)$ and so by Theorem 5.4 in Chapter 5

$$T \sim t_{n-1}. \tag{6.6}$$

> Hence we have a **second requirement of a test statistic**: we need to know what probability distribution it will have when H_0 is true.

5. Identify the **critical region**:

> - The probability of the test statistic lying in this region, assuming H_0 is true, must be exactly α.
> - Observing the test statistic in the critical region should make you think H_0 is less likely to be true: the critical region should be in the 'tails' of the distribution under H_0.

In the example, we set the critical region to be above $t_{n-1;\alpha/2}$ and below $t_{n-1;1-\alpha/2}$. For example, with $\alpha = 0.05$ the **critical values** are 2.262 and -2.262 respectively:

```
qt(c(0.025, 0.975), 9)

## [1] -2.262157  2.262157
```

and so the critical region is $(-\infty, -2.262) \cup (2.262, \infty)$.

6. Now we calculate the observed test statistic t_{obs}:

$$t_{obs} = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}. \tag{6.7}$$

- If t_{obs} lies in the critical region, we declare that "we **reject** H_0 at the $100\alpha\%$ significance level."
- If t_{obs} does *not* lie in the critical region, we declare that "we **do not reject** H_0 at the $100\alpha\%$ significance level."

In the latter case, although don't use the phrase "we accept H_0", the implication is that we will behave as if H_0 is true.

In our example, we calculate

```
mean(drug) / sqrt(var(drug) / 10)

## [1] -2.78136
```

so we have $t_{obs} = -2.78$. This does lie in the critical region, and hence we reject H_0 at the 5% level of significance. This is illustrated in Figure 6.3.

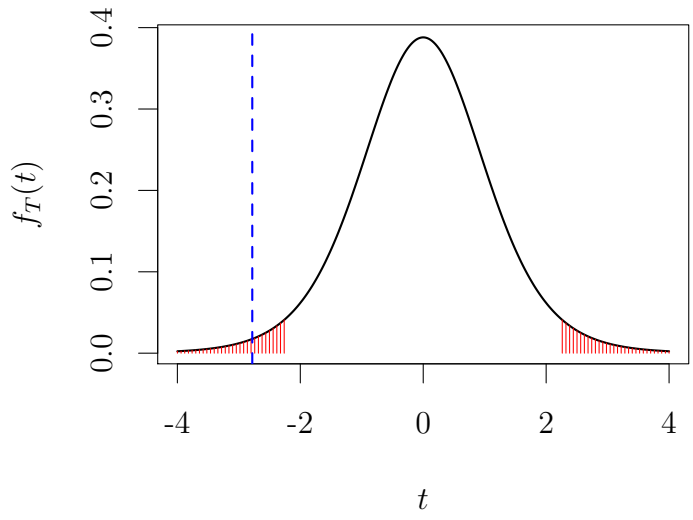


Figure 6.3: The solid line shows the distribution of the random test statistic under H_0. The dashed line shows the value of the observed test statistic (-2.78). The shaded region indicates the 5% critical region. If H_0 were true, the probability of the random test statistic lying in this region would be exactly 0.05. As the observed test statistic does lie in this region, we reject H_0 at the 5% level of significance.

Critical regions for one-sided alternative hypotheses

Had we specified the alternative hypthothesis as

$$H_A : \mu < 0,$$

we would only reject the null in favour of the alternative when we also have $\bar{x} < 0$: we wouldn't conclude $\mu < 0$ had we observed $\bar{x} > 0$, no matter how large $\bar{x}$ was. The critical value would be the 5th percentile of the t_9 distribution:

```
qt(0.05, 9)

## [1] -1.833113
```

This gives a larger critical region in the left-hand tail, but no critical region in the right-hand tail. We plot this in Figure 6.4

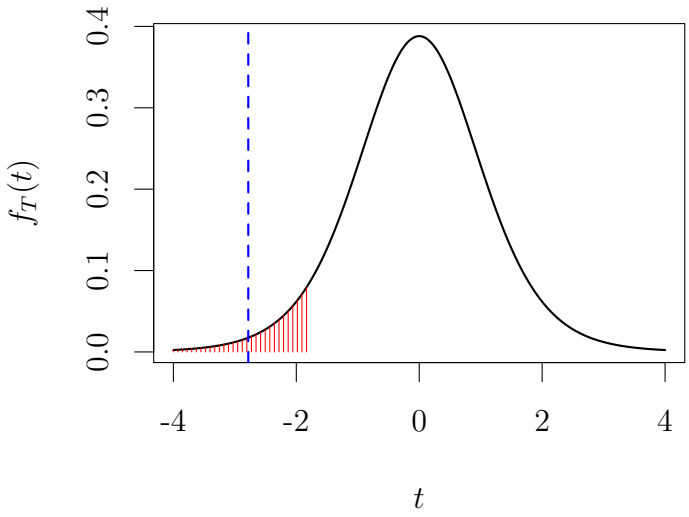


Figure 6.4: The shaded region indicates the 5% critical region for a one-sided alternative $\mu < 0$.

6.5.2 Equivalence of confidence intervals and the Neyman-Pearson framework

A $100(1-\alpha)\%$ confidence interval contains precisely the set of all values μ_c that would *not* be rejected in a Neyman Pearson test of the null hypothesis $H_0 : \mu = \mu_c$ of size α.

> If we have already calculated a $100(1-\alpha)\%$ confidence interval for μ, there is **no need** to do a separate calculation to perform a Neyman Pearson test (of size α) of the hypothesis $H_0 : \mu = \mu_c$: we simply look to see whether the confidence interval contains the value μ_c or not.

Exercise 6.1. *Verify this for the cholesterol example.*

Exercise 6.2. *The 10 percentage changes in cholesterol for patients given the placebo are given below. Using the following R output, conduct a test of size 0.05 that the population mean percentage change is 0%. Can the outcome of the test prove that the population mean change is 0%?*

```
placebo

##  [1]   5.2  13.1  -6.1 -15.2  24.4 -33.0  11.7  -0.1  13.6   5.5

sum(placebo)

## [1] 19.1

sum(placebo^2)

## [1] 2503.37

qt(0.975, 9)

## [1] 2.262157
```

6.6 Fisher's p-value method

In the Neyman-Pearson framework, once the test has been concluded, we decide whether to act as if H_0 is true or not, and we 'consider the matter closed'. In other situations, we may not yet have to make a decision; further investigation may be possible. Using Fisher's p-value method, we simply report the 'strength of evidence' *against* a null hypothesis H_0. To understand the logic of this approach, it is helpful to recap the method of proof by contradiction.

Proof by contradiction

In MAS114 (or perhaps elsewhere), you will have seen the technique of **proof by contradiction**: we assume a statement P is true, show that this leads to a contradiction ($P \to Q$, where we already know that Q is false), from which we conclude that the original statement P must be false. Note that *failing* to find a contradiction doesn't imply P must be true: the argument would reduce to "if we assume P is true, then we conclude P is true".

Fisher's p-value method and proof by contradiction

> Fisher's p-value method can be thought of, informally, as a 'probability version' of proof by contradiction.

1. We assume the null hypothesis H_0 is true.
2. We consider how probable the data would be (in some sense), if the assumption were true.
3. The more unlikely the data are to occur, assuming H_0 is true, the stronger the evidence *against* this assumption; ' highly unlikely' data suggest a 'contradiction' of the assumption that H_0 is true.
4. If the data are *not* unlikely to occur, assuming H_0 is true, then we have simply failed to find evidence against H_0; we have failed to find a 'contradiction'. We do **not** now claim H_0 is true, because that was the assumption we made in (1).

6.6.1 The p-value method step-by-step

We now describe the general procedure for a hypothesis test using Fisher's p-value method, and we illustrate it using the cholesterol example. Before collecting the data, we again suppose we are going to observe random variables $X_1, \ldots, X_n$ (with $n = 10$ in our example).

1. We state the null and alternative hypotheses H_0 and H_1. In the example, we have

$$\begin{aligned} H_0 &: \mu = 0, \\ H_A &: \mu \neq 0. \end{aligned}$$

2. We choose a test statistic: a function of the observations.

> The test statistic in any hypothesis testing problem will be the **same** regardless of whether we are using the Neyman-Pearson framework or Fisher's p-value method.

   For the cholesterol example, we again have

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}, \tag{6.8}$$

   with the corresponding observed value

$$t_{obs} = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}, \tag{6.9}$$

   so $t_{obs} = -2.781$ in this case.

3. We assume that H_0 is true, and derive the distribution of our random test statistic under H_0. Here, we have

$$T \sim t_{n-1}.$$

4. To measure how *unlikely* or 'surprising' the data are under H_0, we calculate what is known as the p-value:

> $p-$value: the probability of obtaining a random test statistic, when H_0 is true, that is at least as 'extreme' as the observed test statistic.

   For our choice of test statistic, and for a two-sided alternative, by 'extreme', we mean large in absolute value, and so the p-value is defined as

$$p := P(T \geq |t_{obs}|) + P(T \leq -|t_{obs}|) = 2 \times P(T \leq -|t_{obs}|). \tag{6.10}$$

   In calculating the p-value, we are calculating how unlikely it is to get an observed test statistic as far away as 2.781 from 0, *if* H_0 were true.

   The observed p-value in our example is obtained from R:

```
2 * pt(-2.781, 9)
```

```
## [1] 0.02136582
```

so our p-value here is 0.02. We illustrate this in Figure 6.5.

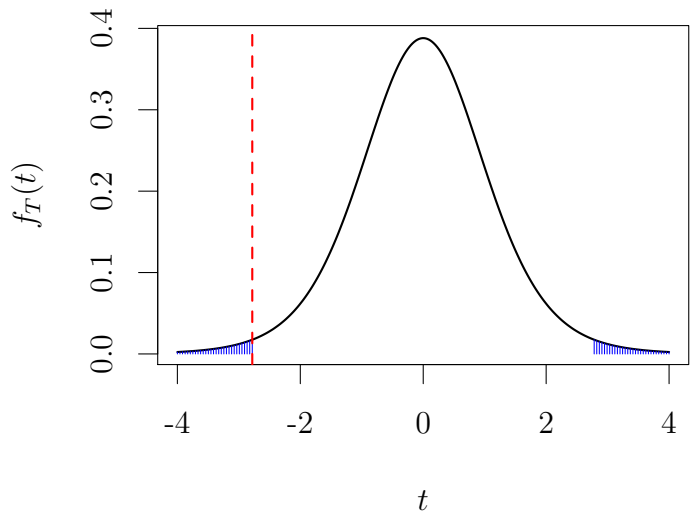


Figure 6.5: The solid line shows the distribution of the random test statistic under H_0. The dashed line shows the value of the observed test statistic. The p-value is the probability of observing a value this 'extreme', in terms of *distance* from 0, and is indicated by the shaded areas.

- The **smaller** the p-value, the more **unlikely** the data are under H_0, and the stronger the evidence **against** H_0 is.
- A **large** p-value simply means we have *failed to find evidence against* that H_0 is true (we have 'failed to find a contradiction'). It is **not** evidence in *in favour of* H_0.

Since our observed p-value was 0.02, we would conclude data as 'extreme' would be fairly unlikely if H_0 were true, so we would count this is 'good' evidence *against* H_0 being true.

What counts as a small p-value?

This varies between scientific fields. In medical research, a p-value of 0.05 or smaller would typically count as 'significant' evidence against the null hypothesis. Hence in our cholesterol example, given the p-value of 0.0214, we would count this as evidence that the mean reduction using the drug is not 0 (and then report the sample mean and a confidence interval to say what we think the mean reduction *is*.)

If a scientist wants to claim a new discovery, and publish the results of his/her experiment in an academic journal, some journals will require a p-value less than 0.05 for the article to be

published, although one journal recently banned this practice[2]. Particle physicists are rather more demanding! They require a p-value smaller than 0.003 for "evidence of a particle", and smaller than 0.0000003 for a "discovery"[3].

Interpreting a p-value: final comments

- Look again at the definition of the $p-$value: it is **not** the probability the null hypothesis is true; it refers to the probability of obtaining data as 'extreme' as those observed *assuming* the null hypothesis is true.
- A large p-value does not indicate strong evidence in favour of the null hypothesis - think of proof by contradiction.
- Although $p < 0.05$ is a commonly used threshold for 'statistically' significant evidence against the null hypothesis, be skeptical if the p-value is only just under 0.05; such evidence should be thought of as 'weak' at best, and the experiment should be repeated to see if the findings can be replicated.

6.7 Equivalence of Fisher's p-value method and the Neyman-Pearson framework

Researchers often blur the distinction between the two approaches

- When using Fisher's p-value method, some researchers choose to state that they will "reject H_0" if the p-value is sufficiently small, for example, if $p < 0.05$. This would be equivalent to conducting a test of size 0.05 in the Neyman Pearson framework.
- When using the Neyman-Pearson framework, some researchers choose to state the smallest significance level at which H_0 would be rejected. This is equivalent to reporting the p-value

Exercise 6.3. *Verify this for the cholesterol example.*

6.8 One sample t-tests and confidence intervals in R

We can use the command `t.test` to perform the test and obtain the confidence interval as follows:

```
t.test(drug)

##
##  One Sample t-test
##
## data:  drug
```

[2] see https://www.statslife.org.uk/news/2116-academic-journal-bans-p-value-significance-test
[3] see https://blogs.scientificamerican.com/observations/five-sigmawhats-that/

```
## t = -2.7814, df = 9, p-value = 0.02135
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -20.418071  -2.101929
## sample estimates:
## mean of x
##    -11.26
```

Note how we can use both the p-value and the confidence interval to conduct a Neyman-Pearson hypothesis test.

6.9 Hypothesis testing methods in general: what you need to know

We have considered a hypothesis test where we are interested in the mean of a single population. There are lots of other different tests for different situations, and we will study some more shortly. The important points to know are

- the basic approach is always the same;
- all that changes is that we use a different test statistic, which will have different distribution under H_0;
- in this module, you will not be expected to construct suitable test statistics for different situations, but you should know how to *recognise* a suitable test statistic: increasing (absolute) values should become less likely under H_0, and more likely under H_A.

6.10 Two-sample t-tests for comparing two population means

Motivation

Continuing with the cholesterol example, we can conduct separate tests to see if the mean reductions are 0 for each of the drug and placebo groups, but what we often want to do is to *compare directly* these two groups. Perhaps taking *any* pill has at least some effect on average, so the question is then whether the average change is *different* for patients taking the drug, compared with those taking the placebo.

Notation and hypotheses

We suppose we have two populations: we observer a random sample $X_1, \ldots, X_n$ from the first, and another random sample $Y_1, \ldots, Y_m$ from the second.

- We suppose the two populations are both normally distributed, denoted by $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ respectively.

- We assume independence: all the random variables $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are independent of each other.

Hence in our example, the population mean cholesterol changes are μ_X for patients on the drug, and μ_Y for patients on the placebo. We write the null and alternative hypotheses as

$$H_0 \ : \ \mu_X - \mu_Y = \delta, \tag{6.11}$$
$$H_A \ : \ \mu_X - \mu_Y \neq \delta \tag{6.12}$$

We will almost always have $\delta = 0$, but the above form is more general. Again, we have used a **two-sided alternative here**: although we may be hoping $\mu_X < \mu_Y$, we would still want to know if in reality $\mu_X > \mu_Y$ (so that the drug is 'harmfull'.)

The test statistic

For a two-sample t-test of the null hypothesis $H_0 : \mu_X - \mu_Y = \delta$ we use the test statistic

$$T = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}, \tag{6.13}$$

where

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{6.14}$$
$$S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2. \tag{6.15}$$

Note:

- The denominator corresponds to the square root of $Var(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$: the variance of the difference between the sample means.
- A large value of $|T|$ would suggest H_A is more likely to be true: we would have $\bar{X} - \bar{Y}$ 'far' from δ, relative to the 'variation' in $\bar{X} - \bar{Y}$.

The distribution of the test statistic under H_0

Assuming H_0 is true, then *approximately*, T will have a t_ν distribution. The best approximation is given by

$$\nu = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}}, \tag{6.16}$$

the **Welch approximation**, and this is what R uses.

For moderate sample sizes (typically $n, m > 30$), $T \sim t_\nu$ is a good approximation to the distribution of T, *even if the invididual population distributions are not normal.*

For large sample sizes (typically $n, m > 100$), the Student t and standard normal distributions become indistinguishable, so for calculation 'by hand'[4], you could use the approximation $T \sim N(0, 1)$.

Once we have the distribution of the test statistic, we can calculate the critical values and/or p-value in the usual way. We illustrate this with the cholesterol data.

6.10.1 Two-sample t-test: an example

Setting up the hypotheses

Define μ_X to be the population mean percentage change for patients on the new drug, and μ_Y to be the population mean percentage change for patients on the placebo. The null and alternative hypotheses are

$$H_0 \ : \ \mu_X = \mu_Y, \tag{6.17}$$
$$H_A \ : \ \mu_X \neq \mu_Y. \tag{6.18}$$

We will test the null hypothesis at the 5% level of significance (so the size of the test is 0.05) and also calculate the p-value.

Calculating the test statistic and the degrees of freedom parameter

Denote the 10 observed percentage changes for the patients with the new drug as $x_1, \ldots, x_{10}$, and the 10 observed percentage changes for the patients with the new drug as $y_1, \ldots, y_{10}$. We have

```
drug
```

```
##  [1] -25.7 -11.6   6.6 -28.7 -15.0 -12.3  -4.8 -17.1  11.8 -15.8
```

```
mean(drug)
```

```
## [1] -11.26
```

```
var(drug)
```

```
## [1] 163.8938
```

```
placebo
```

```
##  [1]   5.2  13.1  -6.1 -15.2  24.4 -33.0  11.7  -0.1  13.6   5.5
```

```
mean(placebo)
```

```
## [1] 1.91
```

```
var(placebo)
```

```
## [1] 274.0988
```

[4] which you would only ever need to do in an exam...

and so we have

$$\bar{x} = -11.26, \quad s_X^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2 = 163.8938, \tag{6.19}$$

$$\bar{y} = 1.91, \quad s_Y^2 = \frac{1}{9}\sum_{i=1}^{10}(y_i - \bar{y})^2 = 274.0988. \tag{6.20}$$

Our observed test statistic is

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{10} + \frac{s_Y^2}{10}}} = -1.99. \tag{6.21}$$

The degrees of freedom parameter, to be used in the t-distribution approximation, is

$$\nu = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}} = 16.928. \tag{6.22}$$

Neyman-Pearson test: obtain the critical values

For a test of size 0.05, the required critical values are given by the 2.5th and 97.5th percentiles of the t-distribution with 16.928 degrees of freedom. We obtain these from R:

```
qt(c(0.025, 0.975), 16.928)
```

```
## [1] -2.110499  2.110499
```

and so the critical region is $(-\infty, -2.110) \cup (2.110, \infty)$.

Neyman-Pearson test: conclusion

Since t_{obs} does *not* lie in the critical region, we do *not* reject the null hypothesis at the 5% level of significance: we have *not* found statistically significant evidence that the mean cholesterol change is different depending on whether the new drug or a placebo is taken. We illustrate this in Figure 6.6

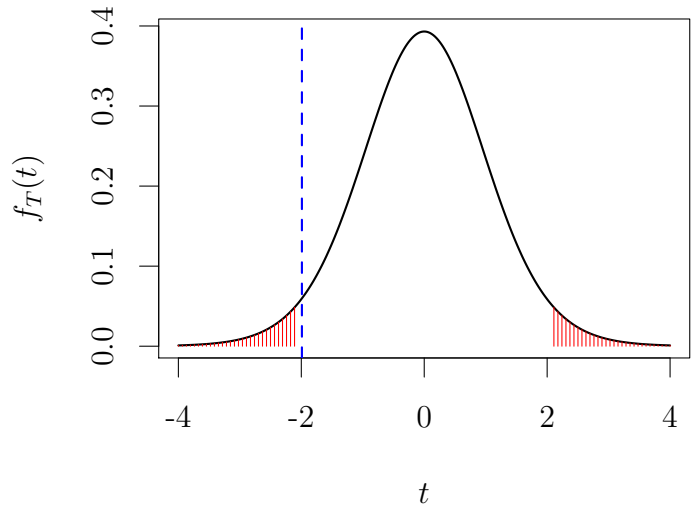


Figure 6.6: The solid line shows the approximate distribution of the random test statistic under H_0. The dashed line shows the value of the observed test statistic. The shaded indicates the 5% critical region. If H_0 were true, the probability of the random test statistic lying in this region would be exactly 0.05. As the observed test statistic does *not* lie in this region, we do *not* reject H_0 at the 5% level of significance.

Calculating the p-value

We need to calculate

$$p = 2 \times P(T > 1.99) = 2 \times P(T < -1.99) \tag{6.23}$$

where $T \sim t_{16.928}$. In R, we do

```
2 * pt(-1.99, 16.928)
```

```
## [1] 0.0629932
```

which gives a p-value of 0.06. If we are working with the threshold of 0.05, we might describe this as 'borderline/weak' evidence against H_0. We illustrate the p-value calculation in Figure 6.7.

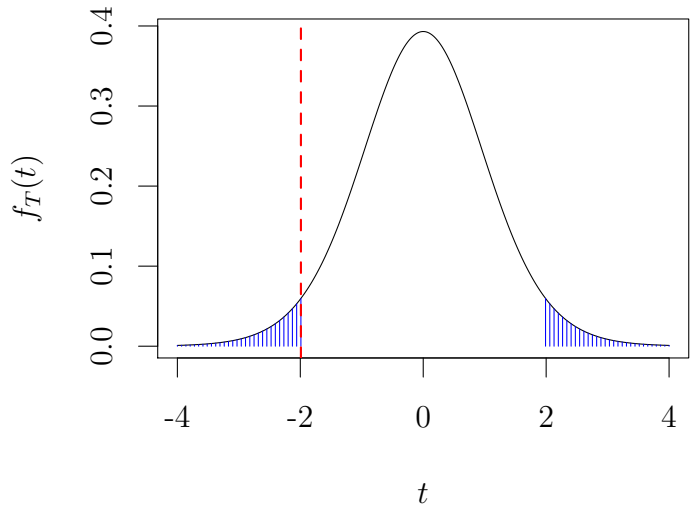


Figure 6.7: The solid line shows the distribution of the random test statistic under H_0. The dashed line shows the value of the observed test statistic. The p-value is the probability of observing a value this 'extreme', in terms of *distance* from 0, and is indicated by the shaded areas. The p-value is calculated to be 0.06.

6.10.2 The two-sample t-test in R and a confidence interval for the difference between the means

We can do the above directly with the `t.test()` command. This will give us a p-value, and a confidence interval for the difference between the two means:

```
t.test(drug, placebo)

##
##  Welch Two Sample t-test
##
## data:  drug and placebo
## t = -1.99, df = 16.928, p-value = 0.06299
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27.1374795   0.7974795
## sample estimates:
## mean of x mean of y
##    -11.26      1.91
```

Note that the 95% confidence interval is calculated as

$$\bar{x} - \bar{y} \pm t_{\nu,0.025}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}. \tag{6.24}$$

6.11 Power and type II errors for a Neyman-Pearson hypothesis test

In the cholesterol example, we *did not* find statistically significant evidence at the 5% level that the drug was more (or less) effective on average than a placebo. However, the confidence interval for the difference between the two means was relatively wide: $[-27.1\%, 0.80\%]$. Perhaps the sample size was too small for detecting a difference? If the null hypothesis *were* false, is it possible we could fail to reject it?

An example scenario when the null hypothesis is *false*

Let X_i and Y_i denote a change in cholesterol level for a randomly sampled patient, treated with the drug and placebo respectively. Suppose ten patients were to be treated in each group, and

$$X_1, \ldots, X_{10} \sim N(\mu_X = -10,\, \sigma_X^2 = 100), \tag{6.25}$$
$$Y_1, \ldots, Y_{10} \sim N(\mu_Y = 0,\, \sigma_Y^2 = 100), \tag{6.26}$$

so that the null hypothesis of $\mu_X = \mu_Y$ would be false here; the drug really is more effective on average. We will collect our data, and perform the two-sample t-test using the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{10} + \frac{S_Y^2}{10}}}.$$

An important point here is that

> whether or not the null hypothesis is actually false, *we always assume it to be true* when performing our hypothesis test: we derive the critical region for T, *assuming* H_0 *to be true.*

Hence, we still compare T with the t_ν distribution, with ν defined in equation (6.10), and, derive the appropriate critical region. The actual critical region will depend on the data, but for now, we will consider the critical region we had in the example: $(-\infty, -2.11) \cup (2.11, \infty)$. Given that H_0 is not true, and the true distribution of the data is specified in equations (6.25) and (6.26), how likely is the test statistic to fall in the critical region: how likely are we to correctly reject H_0?

Power of a hypothesis test and type II errors

> The power of a hypothesis test is the probability of rejecting the null hypothesis when the null hypothesis is **false**. Failing to reject the null hypothesis when the null hypothesis is false is known as a **Type II** error.

To simplify the analysis, we will define

$$\tilde{T} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}}}.$$

and suppose that $T \simeq \tilde{T}$, so that they have approximately the same distribution. We then have

$$\tilde{T} \sim N\left(\frac{-10}{\sqrt{20}}, 1\right)$$

Exercise 6.4. *Verify the above distribution for* $\tilde{T}$.

From this, we can compute (using R)

$$P(\tilde{T} \in (-\infty, -2.11) \cup (2.11, \infty)) = 0.55$$

```
pnorm(-2.11, -10/sqrt(20), 1) + (1 - pnorm(2.11, -10/sqrt(20), 1))

## [1] 0.5501679
```

So, *even though the null hypothesis is false*, there is only (approximately) a 55% chance the test statistic will lie in the critical region, making us reject H_0: in this scenario, *it is still quite likely that we would not 'make the right decision' and reject* H_0. We visualise this in Figure 6.8

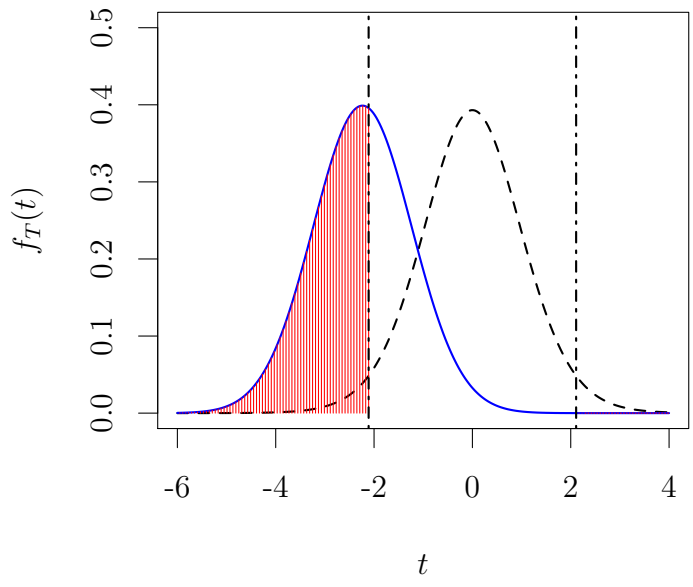


Figure 6.8: The assumed distribution of the test statistic under H_0 is shown as the dashed line, with the dot-dashed lines marking off the critical region. The *true* distribution of the test statistic is shown (approximately) by the solid line, with the shaded region showing the true probability the test statistic will lie in the critical region, and hence the true probability of rejecting H_0. Even though H_0 is false, this probability is not very high: only about 55%.

What determines the power? We will investigate with four further scenarios, where the null hypothesis is false in each case.

Scenario 2: sample size of $n = 10$ patients per group.

$$\begin{aligned} X_1, \ldots, X_{10} &\sim N(-5, 100), && (6.27)\\ Y_1, \ldots, Y_{10} &\sim N(0, 100), && (6.28)\\ \tilde{T} &\sim N\left(\frac{-5}{\sqrt{20}}, 1\right). && (6.29) \end{aligned}$$

Scenario 3: sample size of $n = 10$ patients per group.

$$\begin{aligned} X_1, \ldots, X_{10} &\sim N(-20, 100), && (6.30)\\ Y_1, \ldots, Y_{10} &\sim N(0, 100), && (6.31)\\ \tilde{T} &\sim N\left(\frac{-20}{\sqrt{20}}, 1\right). && (6.32) \end{aligned}$$

Scenario 4: sample size of $n = 10$ patients per group.

$$\begin{aligned} X_1, \ldots, X_{10} &\sim N(-20, 200), && (6.33)\\ Y_1, \ldots, Y_{10} &\sim N(0, 200), && (6.34)\\ \tilde{T} &\sim N\left(\frac{-10}{\sqrt{40}}, 1\right). && (6.35) \end{aligned}$$

Scenario 5: sample size of $n = 100$ patients per group.

$$\begin{aligned} X_1, \ldots, X_{10} &\sim N(-5, 100), && (6.36)\\ Y_1, \ldots, Y_{10} &\sim N(0, 100), && (6.37)\\ \tilde{T} &\sim N\left(\frac{-5}{\sqrt{2}}, 1\right). && (6.38) \end{aligned}$$

Without looking at the next page, what do you think will happen to the power in each case, compared with Scenario 1?

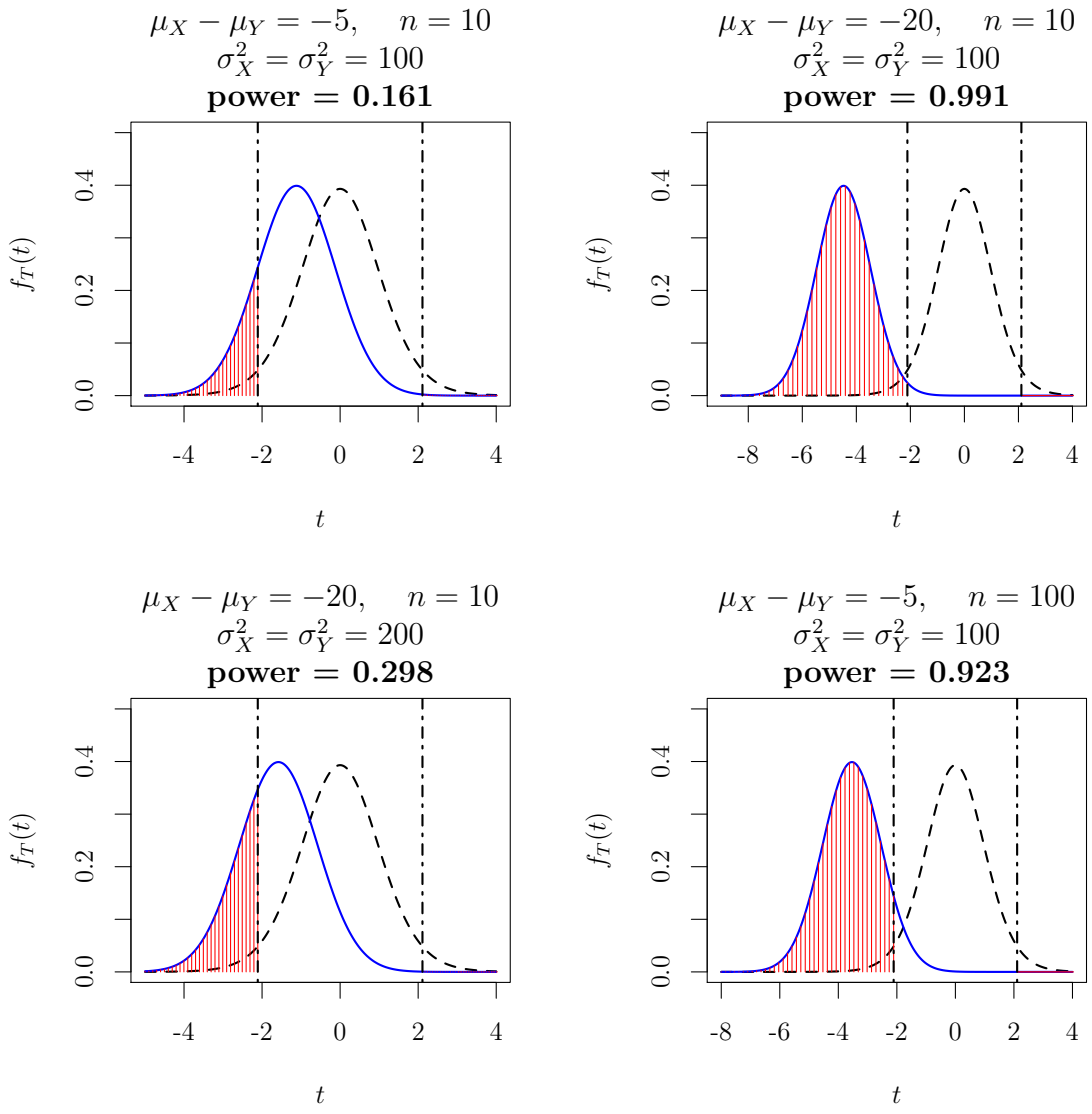


Figure 6.9: The assumed distribution of the test statistic under H_0 is shown as the dashed line, with the dot-dashed lines marking off the critical region. This is the same under each scenario. For Scenarios 2-5, the *true* distribution of the test statistic is shown (approximately) by the solid line, with the shaded region showing the true probability the test statistic will lie in the critical region, and hence the true probability of rejecting H_0.

From inspecting Figure 6.9, we see that the power of the test will increase as any one of

1. the sample size increases (compare top left and bottom right plots);
2. the population variances decrease (compare bottom left and top right plots);
3. the value of $|\mu_X - \mu_Y - \delta|$ increases (compare top left and top right plots).

The effect of (1) and (2) is to reduce the 'noise' or variability in the data: either will reduce the variance of the sample means. With (3), if we are only attempting to detect a large difference between $(\mu_X - \mu_Y)$ and δ, we can be more confident of doing so with a smaller dataset.

6.12 Exercise: testing for differences between two population proportions

We finish this chapter with an exercise to illustrate hypothesis testing in different situation: to compare two proportions. The test statistic and its distribution under the null hypothesis will be different, but the general approach will be the same.

A market research company is trying to improve the response rates (proportion of willing respondents) in its door-to-door surveys. Two surveys are to be conducted. In each survey, $n_1 = n_2 = 1000$ addresses are selected, with different addresses chosen in the two surveys. A company representative visits each address, and asks a household occupant to take part in the survey. Let X_1 and X_2 denote the number of willing respondents in the two surveys.

In the second survey, prospective participants are told that by taking part, they will be entered into a draw to win a £200 cash prize. The company wishes to know whether the offer of a possible cash prize makes participation in the survey more likely.

1. Defining your notation carefully, state the null and alternative hypotheses.

2. A proposed test statistic is

$$Z = \frac{\frac{X_2}{n_2} - \frac{X_1}{n_1}}{\sqrt{P^*(1-P^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{6.39}$$

   where

$$P^* = \frac{X_1 + X_2}{n_1 + n_2} \tag{6.40}$$

   Why might this be an appropriate choice of test statistic?

3. Using a normal approximation to the binomial distribution, derive an approximate distribution of Z assuming the null hypothesis is true.

4. The observed values of X_1 and X_2 are 111 and 168 respectively. Test your null hypothesis at the 5% level of significance. What R command would you use to find the p-value?

5. Suppose, instead, the null hypothesis to be tested was that the two response probabilities were both equal to 0.1.

   (a) By considering an approximate normal distribution for $X_2/n_2 - X_1/n_1$, suggest a modification of the test statistic to be used. Hint: standardise the distribution of $X_2/n_2 - X_1/n_1$.

   (b) Suppose the true response probabilities were 0.1 and 0.15 respectively. For $n_1 = n_2 = 500$, what would the power of the test of size 0.05 be, using your modified test statistic? Give the R command used to find the power.