Chapter 2

Hypothesis tests

2.1 Introduction

We have seen that confidence intervals can be used to make inferences about population parameters. Sometimes, you may be asked to assess whether or not a parameter takes a specific value. For example, whether the population mean $\mu = 5$. One way of re-expressing this question is to ask whether the parameter value is plausible in light of the data. A simple check to see whether the value is contained in a 95% confidence interval will provide an answer. An alternative method, called a *hypothesis test*, is available. It is used extensively in reporting experimental results.

A hypothesis test is a rule for establishing whether or not a set of data is consistent with a hypothesis about a parameter of interest. The *null hypothesis* is a statement that a parameter has a certain value, and is usually written as H_0. For example, $H_0 : \mu = 2.7$, or $H_0 : \sigma^2 = 12.4$. If the null hypothesis is not true, what alternatives are there? Usually, the *alternative hypothesis* is written as H_1. Examples include $H_1 : \mu \neq 2.7$ and $H_1 : \sigma^2 \neq 12.4$.

Based on the information we have in our sample, we'd like to go with either the null hypothesis or the alternative hypothesis. We might use our sample as evidence to suggest that, for example, the population mean could well be equal to 2.7; alternatively, the sample might give evidence to the contrary and suggest that the population mean is not equal to 2.7, or for that matter the sample might suggest that the population mean is less than 2.7!

Suppose you are going on holiday to Sicily in March; a friend tells you that in March, Sicily has an average of 10 hours sunshine a day. On the first three days of your holiday there are 7, 8 and 9 hours of sunshine respectively. You consider that this is evidence that your friend is wrong. Thus, the null hypothesis would state that the average sunshine hours per day is 10 (as suggested by your friend), and your alternative hypothesis might state that the average sunshine hours per day *is less than* 10. You might be tempted to go with the alternative hypothesis. However, this sample of three days could be a fluke result – you might have chosen the most miserable period in March for years for your holiday. *Your test results are not conclusive; they only give you evidence for or against a particular belief.*

2.2 Methodology for hypothesis testing

All hypothesis tests follow the same basic methodology, although the actual calculations may vary depending on the data available.

1. **State the null hypothesis (H_0)**
   We use a hypothesis test to throw light on whether or not this statement is true. For example, you might ask "is the population mean equal to 10?", or "are the two population means equal?"; such hypotheses are expressed in the following way:

$$\begin{aligned} H_0 &: \mu = 10, \qquad \text{and} \\ H_0 &: \mu_1 = \mu_2; \qquad \text{or maybe} \\ H_0 &: \mu = c, \end{aligned}$$

   Where c could be any constant.

2. **State the alternative hypothesis (H_1)**
   This is the conclusion to be reached if the null hypothesis is found to be false. For example, "the population mean does not equal 10", or even "the population mean is less than 10"; to test for two different populations, we might say "the two population means are different". Remember, we can never reject the null hypothesis with certainty; the most we can say is that there is evidence against the null hypothesis, and so evidence in favour of the alternative. Such alternative hypotheses are expressed in the following way:

$$\begin{aligned} H_1 &: \mu \neq 10 \qquad \text{or maybe} \\ H_1 &: \mu < 10. \end{aligned}$$

   To test for two different populations, we might use

$$H_1 : \mu_1 \neq \mu_2.$$

3. **Calculate the test statistic**
   The value calculated from the sample which is used to perform the test is called the *test statistic*. It usually has a similar nature to the population value mentioned in the null hypothesis.

4. **Find the p–value of the test**
   The probability that such an extreme test statistic occurs, *assuming that H_0 is true*, is called the *p–value*, and can be found by comparing the test statistic to values from statistical tables (the tables used will depend on the nature of the test) or using a computer package such as `Minitab` (see chapter 8).

5. **Reach a conclusion**
   A small p–value suggests that our test statistic is unlikely to occur if H_0 is true, and so we reject H_0 in favour of the alternative H_1. But what constitutes a "small" p–value? One commonly used yardstick, or *significance level*, is 5%, or 0.05, though

others can be used. The *smaller* the p–value, the *more* evidence there is to reject the null hypothesis; conversely, the *larger* the p–value, the *less* evidence we have to reject H_0 and so in this case we are more likely to retain the null hypothesis. The following table gives some guidelines on how to interpret your p–value:

p–value	Interpretation
p is bigger than 10%	no evidence against the null hypothesis: stick with H_0
p lies between 5% and 10%	*slight* evidence against H_0, but not enough to reject it
p lies between 1% and 5%	moderate evidence against H_0: reject it, and go with H_1
p is smaller than 1%	strong evidence against H_0: reject it, and go with H_1

Table 2.1: Conventional interpretation of p–values

The above five steps are universal to all hypothesis testing.

2.3 Testing one mean

Here, from a *single* population, we draw a *single* sample, and we estimate the population mean μ with the sample mean $\bar{x}$. We'd then like to see how convincing a proposal (say c) for the population mean is, based on the information in our sample. As with the construction of confidence intervals, the choice of test statistic in step 3 above depends on whether or not the population variance (or standard deviation) is known. We will now demonstrate a test for one mean using two examples: one in which the population variance is known (case 1), and one in which it is unknown (case 2).

2.3.1 Example: Known population variance σ^2

A chain of shops believes that the average size of transactions is £130, and the population variance is known to be £900. The takings of one branch were analysed and it was found that the mean transaction size was £123 over the 100 transactions in one day. Based on this sample, test the null hypothesis that the true mean is equal to £130.

Since σ^2 is known (we are given that $\sigma^2 = 900$), this corresponds to case 1: population variance known (think back to confidence intervals). We now proceed with the five steps outlined in the previous section.

Steps 1 and 2 *(hypotheses)*
Here, we state our null and alternative hypotheses. The null hypothesis is given in the question – i.e.

$$H_0 \ : \ \mu = £130.$$

We could test against a general alternative, i.e.

$$H_1 \ : \ \mu \neq £130.$$

Step 3 *(calculating the test statistic)*
When σ^2 is known, we use following test statistic

$$\begin{aligned} z &= \frac{|\bar{x} - \mu|}{\sqrt{\sigma^2/n}}, \qquad \text{i.e.} \\ z &= \frac{|123 - 130|}{\sqrt{900/100}} \\ &= \frac{7}{\sqrt{9}} \\ &= 2.33. \end{aligned}$$

Step 4 *(finding the p–value)*
Recall the Central Limit Theorem from Chapter 1; this tells us that the quantity

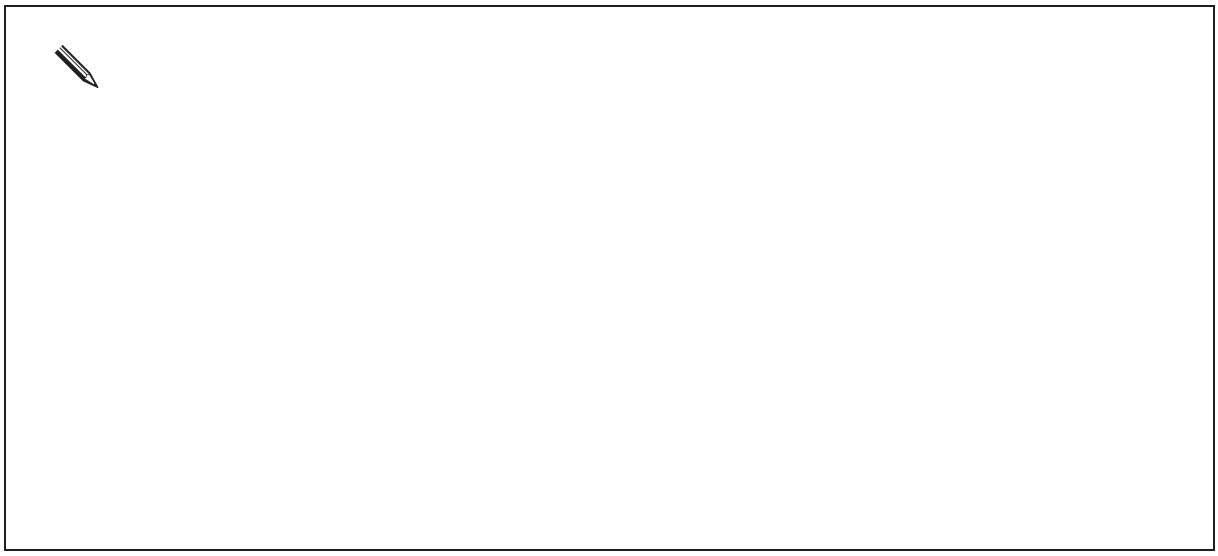
$$\frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$$

follows a standard Normal distribution. Thus, the value we obtain from our test statistic formula above will be from the positive half of the standard Normal distribution. We can therefore compare our test statistic to critical values from the standard Normal distribution to find our p–value, or at least a range for our p–value. Remember, this is the probability of observing our data, or anything more extreme than this, if the null hypothesis is true; thus, the smaller the p–value, the more evidence there is *against* H_0.

Our alternative hypothesis is *two–tailed* (i.e. $\neq$ rather than $<$ or $>$), and so our values are:

Significance level	10%	5%	1%
Critical value	1.645	1.96	2.576

Our test statistic $z = 2.33$ lies between the critical values of 1.96 and 2.576, and so our p–value lies between 1% and 5%. We can see this more clearly on a diagram:



Step 5 *(conclusion)*
Using table 2.1 to interpret our p–value, we see that there is moderate evidence against H_0. Thus, we should reject H_0 in favour of the alternative hypothesis H_1; it appears that the population mean transaction size *is not equal to* £130.

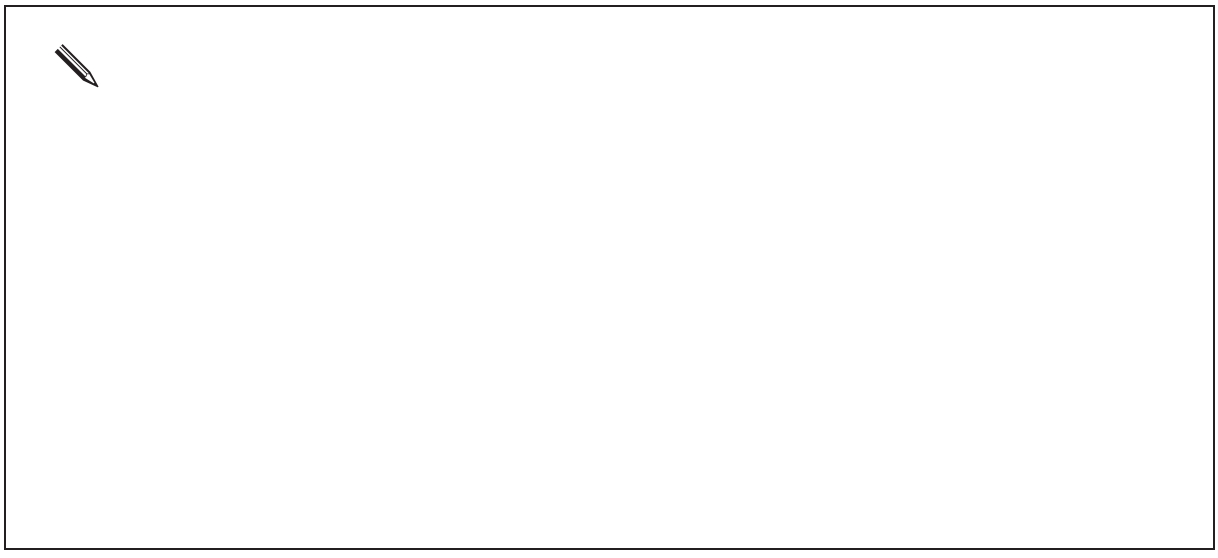
Alternatively, since our sample mean $\bar{x} = £123$ is smaller than the proposed value of £130, we could have set up a *one–tailed* alternative hypothesis in step 2, i.e. we could have tested

$$\begin{aligned} H_0 &: \mu = £130 \qquad \text{against} \\ H_1 &: \mu < £130. \end{aligned}$$

This is now a one–tailed test and the critical values from table 2.2 are

Significance level	10%	5%	1%
Critical value	1.282	1.645	2.326

The test statistic is (as before) 2.33, which now lies "to the right" of the last critical value in the table (2.326). Thus, our p–value is now smaller than 1%, and so, using table 2.1, we see that in this more specific test there is *strong* evidence against H_0. Again, this can be seen more clearly with a diagram:



Notice that this one–tailed test is more specific than the two–tailed test previously carried out. If you're not sure whether you should perform a one–tailed test or a two–tailed test, (i.e. there might not be much difference between the proposed mean and the sample mean), it's usually safer to test against the more general two–tailed alternative.

2.3.2 Example: Unknown population variance σ^2

The batteries for a fire alarm system are required to last for 20000 hours before they need replacing. 16 batteries were tested; they were found to have an average life of 19500 hours and a standard deviation of 1200 hours. Perform a hypothesis test to see if the batteries do, on average, last for 20000 hours.

Steps 1 and 2 *(hypotheses)*
Using a one–tailed test, our null and alternative hypotheses are:

$$\begin{aligned} H_0 &: \mu = 20000 \qquad \text{versus} \\ H_1 &: \mu < 20000. \end{aligned}$$

We use a one–tailed test because we are interested in whether the batteries are effective or not; there is no problem if they last longer than 20000 hours.

Step 3 *(calculating the test statistic)*
Unlike the previous example, the population variance σ^2 is unknown (i.e. the question does not say "the population variance is ...", or "the population standard deviation is ...", for example). However, the *sample* standard deviation is given, based on a sample of size 16, and so we need to use a slightly different test statistic. In fact, we do what we did last week when we were constructing confidence intervals – i.e. we replace σ^2 with s^2 and then use tables of values from Student's t distribution instead of the standard Normal distribution. Thus, the test statistic is given by

$$\begin{aligned} t &= \frac{|\bar{x} - \mu|}{\sqrt{s^2/n}} \\ &= \frac{|19500 - 20000|}{\sqrt{1200^2/16}} \\ &= \frac{500}{\sqrt{1440000/16}} \\ &= 1.667. \end{aligned}$$

Step 4 *(finding the p–value)*
Since σ^2 is unknown, we use t–distribution tables (table 2.3) to obtain a range for our p–value. The degrees of freedom, $\nu = n - 1 = 16 - 1 = 15$, and under a one–tailed test this gives the following critical values:

Significance level	10%	5%	1%
Critical value	1.341	1.753	2.602

Our test statistic of $t = 1.667$ lies between the critical values of 1.341 and 1.753, and so the corresponding p–value lies between 5% and 10%.

Step 5 *(conclusion)*
Using table 2.1 to interpret our p–value, we see that there is only *slight* evidence against the null hypothesis and certainly not enough grounds to reject it, so we retain H_0. There is insufficient evidence to suggest there is a problem with these batteries.

Exercises

1. A machine for filling cans of Coke has a process variance of 400ml. A sample of 100 cans is taken and it is found that the average contents are 240ml. Is this consistent with the cans containing the stated weight of 250ml? Use a two–tailed test.

2. A company manufactures bolts. One production line produces bolts that are designed to be 100mm long. A sample of 50 bolts is taken and measured; they are found to have an average length of 97.5mm with variance 150mm. Are the bolts being made too short?

3. A company is in dispute with its work force. The workers claim that under a new flexitime system they are working longer than the standard 37.5 hour week. The time cards of 10 workers were selected at random and these showed the following hours worked:

   35 40 45 41 36 37 39 38 42 32

   Perform a hypothesis test to see if staff *are* working more than a standard week.

One–tailed test	10%	5%	2.5%	1%	0.5%
Two–tailed test	20%	10%	5%	2%	1%
Critical value	1.282	1.645	1.96	2.326	2.576

Table 2.2: Tabulated values of z for which $\Pr(Z > z) = p$, where Z has a standard normal distribution

	One–tailed test	10%	5%	2.5%	1%	0.5%
	Two–tailed test	20%	10%	5%	2%	1%
	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.449
ν	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	21	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763
	29	1.311	1.699	2.045	2.462	2.756
	⋮	⋮	⋮	⋮	⋮	⋮
	∞	1.282	1.645	1.960	2.326	2.576

Table 2.3: Tabulated values of t for which $\Pr(|T| > t) = p$, where T has a t–distribution with ν degrees of freedom