Testing hypotheses

Contents

Chapter outline

Hypothesis testing starts with a statement describing some aspect of a population – giving the hypothesis to be tested. Then it examines a sample from the population and sees if there is evidence to support the hypothesis. Either the evidence supports the hypothesis or it does not support it and, by implication, supports some alternative hypothesis. You can use this general approach to hypothesis testing in many circumstances.

After finishing this chapter you should be able to:

- understand the purpose of hypothesis testing
- list the steps involved in hypothesis testing
- understand the errors involved and the use of significance levels
- test hypotheses about population means
- use one- and two-tail tests
- extend these tests to deal with small samples
- use the tests for a variety of problems
- consider non-parametric tests, particularly the chi-squared test.

Aim of hypothesis testing

In Chapter 14 we saw how statistical inference uses data from a sample to estimate values for a population. In this chapter we extend this idea by testing if a belief about a population is supported by the evidence from a sample. This is the basis of **hypothesis testing**.

Suppose you have some preconceived idea about the value taken by a population variable. For instance, you might believe that domestic telephone bills have fallen by 10% in the past year. This is a hypothesis that you want to test. So you take a sample from the population and see whether or not the results support your hypothesis. The formal procedure for this is:

- Define a simple, precise statement about a population (the hypothesis).
- Take a sample from the population.
- Test this sample to see whether it supports the hypothesis, or makes the hypothesis highly improbable.
- If the hypothesis is highly improbable reject it, otherwise accept it.

This seems a reasonable approach – but it needs a small adjustment. Statisticians are more cautious than this, and they do not talk about 'accepting' a hypothesis. Instead, they say that they 'can reject the hypothesis' if it is highly unlikely, or they 'cannot reject the hypothesis' if it is more likely.

WORKED EXAMPLE 15.1

Aceituna GmbH fills bottles with a nominal 400 ml of olive oil. There are small variations around this nominal amount and the actual contents are Normally distributed with a standard deviation of 20 ml. The company takes periodic samples to make sure that they are filling the bottles properly. If it is found that a sample bottle contains 446 ml, are the bottles being overfilled?

Solution

We start with an initial hypothesis that the bottles still contain 400 ml. We have a limited amount of data from a single sample but can use this to test the hypothesis. If the hypothesis is correct then we can find the probability of finding a bottle containing 446 ml by calculating the number of standard deviations this point is from the mean:

$$Z = (446 - 400) / 20 = 2.3$$

which corresponds to a probability of 0.01. If our hypothesis that the bottles contain 400 ml is correct, finding a bottle with 446 ml is highly improbable, occurring on only 1% of occasions. So we can reasonably reject the initial hypothesis that the bottles contain 400 ml.

The original statement is called the **null hypothesis** – usually called H_0. The name 'null' implies there has been no change in the value being tested since the hypothesis was formulated. If we reject the null hypothesis then we implicitly accept an alternative. In Worked example 15.1 we reject the null hypothesis that the bottles contain 400 ml, so we accept the **alternative hypothesis** that they do not contain 400 ml. For each null hypothesis there is always an alternative hypothesis – usually called H_1. If the null hypothesis, H_0, is that domestic telephone bills have fallen by 10% in the last year, the alternative hypothesis, H_1, is that they have not fallen by 10%.

The null hypothesis must be a simple, specific statement – while the alternative hypothesis is less precise and suggests only that some statement other than the null hypothesis is true. In practice, the null hypothesis is usually phrased in terms of one thing equalling another, while the alternative hypothesis is that the equality is not true. So a null hypothesis, H_0, says that the average salary in an office is €50,000 and the alternative hypothesis, H_1, is that the average salary is not €50,000.
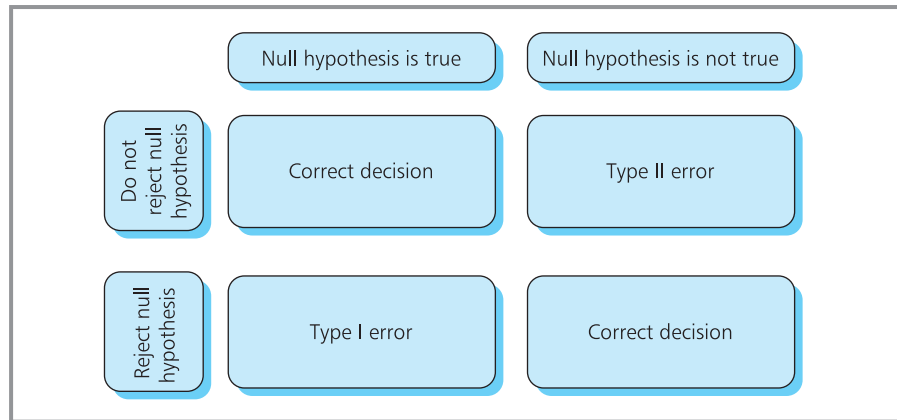
Figure 15.1 Errors in hypothesis testing

Errors in hypothesis testing

Sampling always contains uncertainty, so we can never be certain of the results from a hypothesis test. In Worked example 15.1, the result was said to be unlikely – occurring only 1 time in 100 – so we rejected the null hypothesis. But if the null hypothesis is actually true, we would still get this result in 1% of samples, and we would be rejecting a true hypothesis. Conversely, a sample might give evidence to support a null hypothesis, even when it is not true. So there are two ways of getting the wrong answer (shown in Figure 15.1):

■ **Type I error** – when we reject a null hypothesis that is actually true
■ **Type II error** – when we do not reject a null hypothesis that is actually false.

We want the probabilities of both Type I and Type II errors to be as close to zero as possible – and the only way of ensuring this is to use large samples. Otherwise, any adjustments to reduce the probability of Type I errors inevitably increase the probability of Type II errors, and vice versa. With a limited sample size, we have to accept a compromise between the two errors.

WORKED EXAMPLE 15.2

The city of Halifax, Nova Scotia, takes a survey of monthly costs of food and housing for a particular type of family. They think that the mean cost is $1,600 with a standard deviation of $489. A sample of 100 families has an average expenditure of $1,712.50. Does this support their initial views?

Solution

We start by defining the null hypothesis, H_0, that the monthly cost of food and housing is $1,600; while the alternative hypothesis, H_1, is that it does not equal $1,600.

Hypothesis tests assume that the null hypothesis is true for the test, so we assume that the population has a mean of $1,600 and a standard deviation of $489. Then we find the probability that a sample with a mean of $1,712.50 comes from this population. With a large sample of 100 the sampling distribution of the mean is Normal, with standard error $\sigma/\sqrt{n} = 489/\sqrt{100} = 48.9$. Then:

$$Z = (1,712.5 - 1,600) / 48.9$$
$$= 2.3$$

Worked example 15.2 continued

This corresponds to a probability of 0.0107. If the null hypothesis is true, there is a probability of 0.0107 that the monthly cost of a sample is $1,712.50. This is very unlikely, so the evidence does not support the null hypothesis. We can reject this and accept the alternative hypothesis. But remember that in about 1% of cases we are making a Type I error and rejecting a hypothesis that is actually true.

WORKED EXAMPLE 15.3

Marcia Lopez says that the mean wage of her employees is €300 a week with a standard deviation of €60. She checks a random sample of 36 wages and decides to revise her views if the mean of the sample is outside the range €270 to €330. What is the probability she makes a Type I error?

Solution

The null hypothesis, H_0, is that the mean wage is €300, and the alternative hypothesis, H_1, is that the mean wage is not €300.

Assuming that the population has a mean of €300 and standard deviation of €60, with a sample of 36 the sampling distribution of the mean is Normal with standard error $\sigma / \sqrt{n} = 60 / \sqrt{36} = 10$. Then:

$$Z = (330 - 300) / 10$$
$$= 3$$

which corresponds to a probability of 0.0013. By symmetry, the probability of a sample mean being less than €270 is also 0.0013. So when the null hypothesis is true, there is a probability of $0.0013 + 0.0013 = 0.0026$ that the sample mean will be outside the acceptable range. This is the probability that Marcia rejects the null hypothesis when it is actually true.

Review questions

15.1 What is the purpose of hypothesis testing?

15.2 Which is a more precise statement, H_0 or H_1?

15.3 What are Type I and Type II errors?

IDEAS IN PRACTICE Politics and hypotheses

Most – but not all – legal systems work with the belief that someone accused of a crime is innocent until proven guilty. In other words, they start with a null hypothesis that someone is innocent, and if this is extremely unlikely the system rejects this hypothesis and accepts the alternative hypothesis that they are guilty.

No justice system is infallible, and a Type I error occurs when an innocent person is punished; a Type II error occurs when a guilty person is not punished. There is always a balance between these errors, and reducing the chance of one inevitably increases the chance of the other. People have different views about achieving a reasonable balance. Some emphasise the importance of punishing the guilty; others emphasise the injustice of condemning the innocent.

This effect can be broadened into other areas, such as public welfare payments – where a Type I error gives welfare payments to someone who does not deserve it and a Type II error fails to give payments to someone who really needs it.

In the USA, Hooke has suggested that liberals and conservatives have fundamentally different views of these errors. For example, with justice, liberals avoid Type I errors, while conservatives avoid Type II errors. With welfare payments it is the other way around – conservatives avoid Type I errors, while liberals avoid Type II errors. As it is impossible to eliminate one type of error without increasing the other type of error, the two political philosophies will never reach agreement.

Significance levels

So far we have rejected a null hypothesis if we consider the result from the sample to be unlikely. However, our judgement of what is 'unlikely' has been purely subjective. We can formalise this judgement in a **significance level**.

> ■ A **significance level** is the minimum acceptable probability that a value actually comes from the hypothesised population.
> ■ When the probability is less than this, we reject the null hypothesis; when the probability is more than this we do not reject it.

With a 5% significance level, we reject the null hypothesis when there is a probability of less than 0.05 that a sample value comes from the population – but we do not reject a null hypothesis if there is a probability greater than 0.05 that a sample value comes from the population. Of course, this still gives 5% of tests where values fall outside the acceptance range when the null hypothesis is actually true, and we reject a null hypothesis that is true. As this is a Type I error, you can see that the significance level is the maximum acceptable probability of making a Type I error.

You can use any value for a significance level, but the most common is 5%, followed by 1% and occasionally 0.1%. With a large sample, the sampling distribution is Normal, and 95% of observations are within 1.96 standard deviations of the mean – so this defines the acceptance level (shown in Figure 15.2). With a 1% significance level, you cannot reject the null hypothesis if the observation is within 2.58 standard deviations of the mean. This is clearly a less stringent test – and it shows that lower significance levels need stronger evidence to reject the null hypothesis.

Of course, a particular test might have you rejecting a hypothesis at a 5% level, but not rejecting it at a 1% level – or even a 4% level. So this might seem like a rather arbitrary border between reject and cannot reject decisions. It might be better to simply calculate the exact probability that a hypothesis is
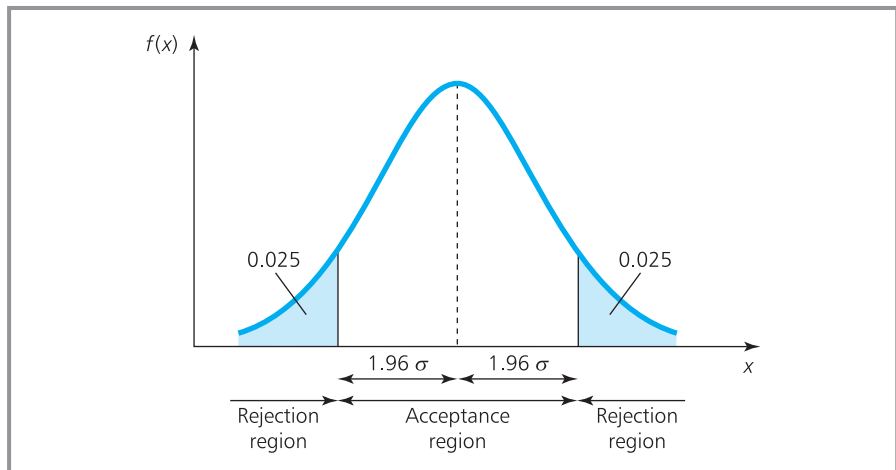

Figure 15.2 Acceptance and rejection regions for 5% significance level

true, and you could say, 'there is a probability of 0.002 that this sample comes from the hypothesised population'. In principle this would be better – but the standard format has been used for many years and is unlikely to change.

WORKED EXAMPLE 15.4

John Lo thinks that the mean value of orders received by his firm is €260. He checks a sample of 36 accounts and finds a mean of €240 and standard deviation of €45. Does this evidence support his belief?

Solution

The null hypothesis is that the mean value of accounts is €260, and the alternative hypothesis is that the mean is not €260. Then:

$$H_0 : \mu = 260 \quad H_1 : \mu \neq 260$$

We do not know the population standard deviation, but can estimate it from the sample standard deviation using $\sigma = s / \sqrt{n}$. Then with a sample of

36, the sampling distribution of the mean is Normal with mean 260 and standard error 45 / $\sqrt{36}$ = 7.5. With a significance level of 5% we do not reject values that are within 1.96 standard deviations of the mean. So the acceptance range is:

$$260 - 1.96 \times 7.5 \quad \text{to} \quad 260 + 1.96 \times 7.5$$

or

$$245.3 \quad \text{to} \quad 274.4$$

The actual observation of €240 is outside this range, so we reject the null hypothesis, and accept the alternative hypothesis that the mean value of orders is not equal to €260.

Worked example 15.4 illustrates the steps in the formal procedure for hypothesis testing.

---

1 State the null and alternative hypotheses.
2 Specify the significance level.
3 Calculate the acceptance range for the variable tested.
4 Find the actual value for the variable tested.
5 Decide whether or not to reject the null hypothesis.
6 State the conclusion.

WORKED EXAMPLE 15.5

The Central Tax Office says that the average income in Port Elizabeth is $15,000. A sample of 45 people found their mean income to be $14,300 with a standard deviation of $2,000. Use a 5% significance level to check the claim. What is the effect of using a 1% significance level?

Solution

For this, you use the standard six-step procedure.

1 *State the null and alternative hypotheses:*

$$H_0 : \mu = 15,000 \quad H_1 : \mu \neq 15,000$$

2 *Specify the level of significance*. This is given as 5%.

3 *Calculate the acceptance range for the variable tested*. With a sample of 45, the sampling distribution of the mean is Normal with mean 15,000 and standard error approximated by $s / \sqrt{n}$ = 2,000 / $\sqrt{45}$ = 298.14. For a 5% significance level you cannot reject points that are within 1.96 standard deviations of the mean. So the acceptance range is:

$$15,000 - 1.96 \times 298.14 \quad \text{to}$$
$$15,000 + 1.96 \times 298.14$$

$$\text{or} \qquad 14,415.65 \quad \text{to} \quad 15,584.35$$

4 *Find the actual value for the variable tested*. This is $14,300.

Worked example 15.5 continued

5   *Decide whether or not to reject the null hypo-thesis.* The actual value is outside the accept-ance range, so you reject the null hypothesis.
6   *State the conclusion.* At a 5% significance level, the evidence from the sample does not support the claim that the average income per capita in Port Elizabeth is $15,000. Instead, it supports the alternative hypothesis that the average income is not $15,000.

With a 1% significance level, the acceptance range is within 2.58 standard deviations of the mean, or:

$$15,000 - 2.58 \times 298.14 \quad \text{to} \quad 15,000 = 2.58 \times 298.14$$

or   14230.79   to   15,769.21

The actual observation of $14,300 is within this range, so you cannot reject the null hypothesis (as shown in Figure 15.3).
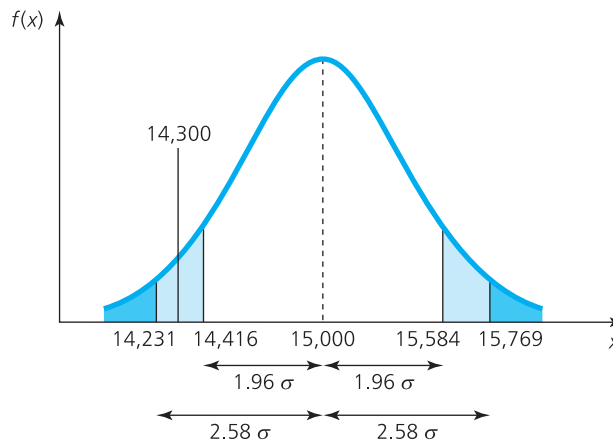
Figure 15.3 Acceptance range for incomes in Port Elizabeth (Worked example 15.3)

One-sided tests

In the problems we have looked at so far, we have stated a null hypothesis of the form:

$$H_0 : \mu = 10$$

and an alternative hypothesis in the form:

$$H_1 : \mu \neq 10$$

In practice, we often want to test whether a value is above or below some claimed value. If we buy a bottle of whisky, we want to be sure that the volume is not below the specified value; and if we are delivering parcels, we want to know that their weight is not above the claimed weight. We can tackle problems of this type by using the standard procedure, but with an adjustment to the phrasing of the alternative hypothesis.

If you are buying boxes of chocolates with a specified weight of 500 g, you want to be sure that the actual weight is not below this and use:

Null hypothesis, H_0 : $\mu = 500$ g
Alternative hypothesis, H_1 : $\mu < 500$ g
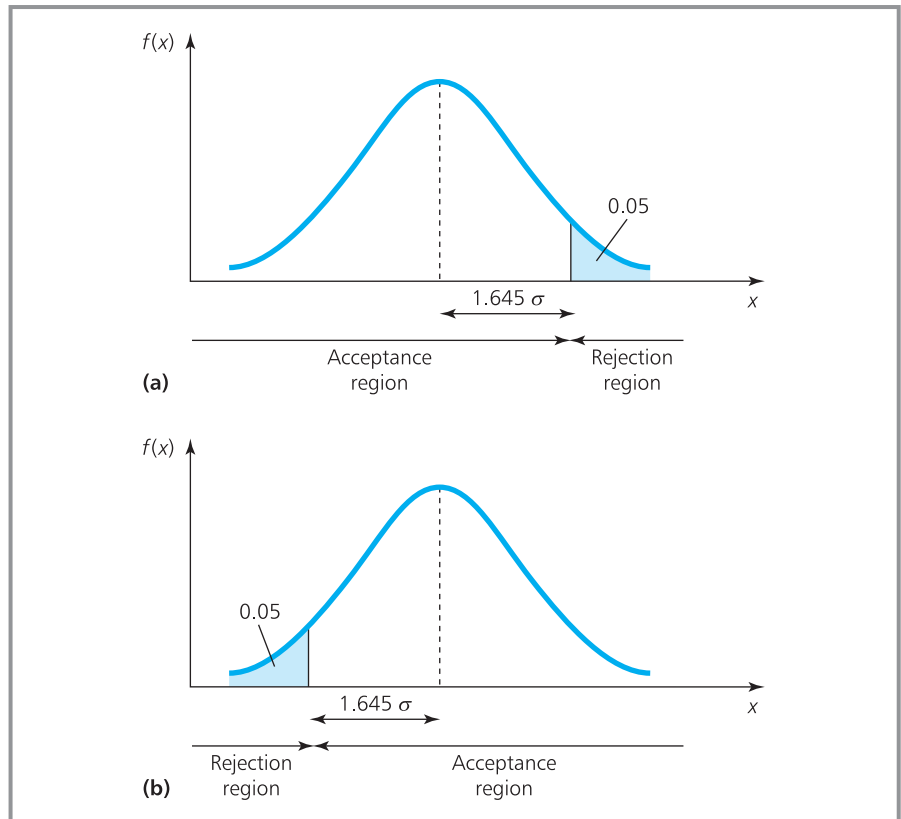
Figure 15.4 One-sided test for 5% significance level: (a) when concerned with a maximum value; (b) when concerned with a minimum value

If you are delivering parcels with a claimed weight of 25 kg, you want to be sure the actual weight is not above this and use:

Null hypothesis, H_0 : $\mu = 25$ kg
Alternative hypothesis, H_1 : $\mu > 25$ kg

In both examples you are interested in only one tail of the sampling distribution, so the acceptance range is altered. In particular, a 5% significance level has the 5% area of rejection in one tail of the distribution. In a Normal distribution this point is 1.645 standard deviations from the mean, as shown in Figure 15.4.

WORKED EXAMPLE 15.6

BookCheck Mail-Order charge customers a flat rate for delivery based on a mean weight for packages of 1.75 kg with a standard deviation of 0.5 kg. Postage costs have risen and it seems likely that the mean weight is greater than 1.75 kg. The company checked a random sample of 100 packages and found a mean weight of 1.86 kg. Does this support the view that the mean weight is more than 1.75 kg?

Worked example 15.6 continued

Solution
We use the standard procedure.

1 *State the null and alternative hypotheses*. We want to test that the mean weight is not above 1.75 kg, so we have:

$$H_0 : \mu = 1.75 \text{ kg} \quad H_1 : \mu > 1.75 \text{ kg}$$

2 *Specify the level of significance*. This is not given, so we assume 5%.

3 *Calculate the acceptance range for the variable tested*. With a sample of 100, the sampling distribution of the mean is Normal with mean of 1.75 kg and standard deviation $\sigma / \sqrt{n} = 0.5 / \sqrt{100} = 0.050$ kg. For a 5% significance level and

a one-sided test, we reject points that are more than 1.645 standard deviations above the mean. The acceptance range is below $1.75 + 1.645 \times 0.05 = 1.83$ kg.

4 *Find the actual value for the variable tested*. The observed weight of parcels is 1.86 kg.

5 *Decide whether or not to reject the null hypothesis*. The actual value is outside the acceptance range, so we reject the null hypothesis.

6 *State the conclusion*. The evidence from the sample does not support the view that the mean weight of packages is 1.75 kg. The evidence supports the alternative hypothesis, that the mean weight is more 1.75 kg. This is illustrated in Figure 15.5.
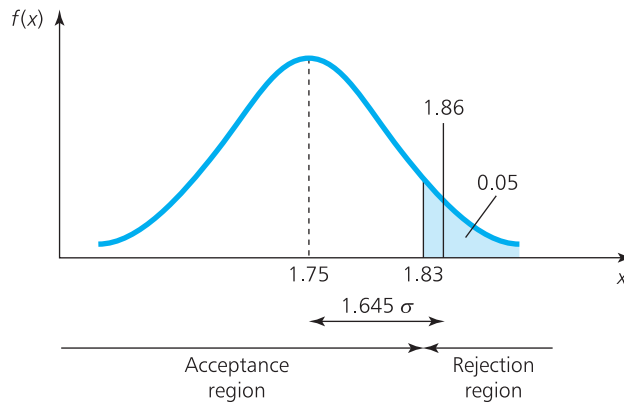
Figure 15.5 Acceptance region for BookCheck Mail-Order (Worked example 15.6)

WORKED EXAMPLE 15.7

Elisabeta Horst is a management consultant who has recently introduced new procedures to a reception office. The receptionist should do at least 10 minutes of paperwork in each hour. Elisabeta made a check on 40 random hours of work and found that the mean time spent on paperwork is 8.95 minutes with a standard deviation of 3 minutes. Can she reject the hypothesis that the new procedures meet specifications at a 1% level of significance?

Solution

1 *State the null and alternative hypotheses*. Elisabeta wants to check that the time spent on paperwork is at least 10 minutes in an hour. So:

$$H_0 : \mu = 10 \text{ minutes} \quad H_1 : \mu < 10 \text{ minutes}$$

2 *Specify the level of significance*. This is given as 1%.

3 *Calculate the acceptance range for the variable tested*. With a sample of 40, the sampling

Worked example 15.7 continued

distribution of the mean is Normal with a mean of 10 minutes and standard deviation $s/\sqrt{n} = 3/\sqrt{40} = 0.474$ minutes. For a 1% significance level and a one-sided test, Elisabeta should reject values that are below 2.33 standard deviations below the mean. Then the acceptance range is above $10 - 2.33 \times 0.474 = 8.89$ minutes.

4 *Find the actual value for the variable tested*. The observed number of minutes spent on paperwork in each hour is 8.95.

5 *Decide whether or not to reject the null hypothesis*. The actual value is inside the acceptance range and Elisabeta cannot reject the null hypothesis.

6 *State the conclusion*. The evidence from the sample supports the view that the mean time spent on paperwork is at least 10 minutes an hour.

Review questions

15.4   What is a significance level?

15.5   Is the probability of a Type II error lower with a 5% significance level or a 1% significance level?

15.6   If a value is in the acceptance range, does this prove that the null hypothesis is true?

15.7   When would you use a one-sided hypothesis test?

Tests with small samples

In Chapter 14 we noted that sampling distributions are Normal only when the population is Normal or the sample size is more than 30. When this is not true, the sampling distribution follows a *t*-distribution. Remember that the shape of a *t*-distribution depends on the degrees of freedom, which is the sample size minus one.

WORKED EXAMPLE 15.8

A coffee machine is set to fill cups with 200 ml of coffee. A sample of 10 cups contained 210 ml with a standard deviation of 10 ml. Is the machine working properly?

Solution

We can use the standard approach for hypothesis testing, with a two-tail test, but because the sample size is small we have to use a *t*-distribution.

1 *State the null and alternative hypotheses*. The null hypothesis is that the dispenser is filling cups with 200 ml, while the alternative hypothesis is that it is not filling cups with 200 ml.

$H_0 : \mu = 200$ ml $H_1 : \mu \neq 200$ ml

2 *Specify the level of significance*. We can use the standard 5%.

3 *Calculate the acceptance range for the variable tested*. With a sample of size 10, the sampling distribution of the mean follows a *t*-distribution with $n - 1 = 10 - 1 = 9$ degrees of freedom, a mean of 200 ml and standard deviation $s/\sqrt{(n-1)} = 10/\sqrt{9} = 3.33$ ml. With the small sample we have again used Bessel's correction for small samples, dividing the sample standard deviation by $n - 1$ rather than n. For a 5% significance level and a two-sided test, we look up

➤

Worked example 15.8 continued

(either in tables or using the TINV function in a spreadsheet) a probability of 0.025 in each tail, and with 9 degrees of freedom the value is 2.262. Then the acceptance range is:

$$200 - 2.262 \times 3.33 \quad \text{to} \quad 200 + 2.262 \times 3.33$$
$$192.47 \quad \text{to} \quad 207.53$$

4 *Find the actual value for the variable tested.* The actual mean of the sample was 210 ml.
5 *Decide whether or not to reject the null hypothesis.* The actual value is outside the acceptance range, so we reject the null hypothesis.
6 *State the conclusion.* The evidence from the sample does not support the view that the machine is filling cups with 200 ml of coffee.

WORKED EXAMPLE 15.9

A supermarket is getting complaints that its tins of strawberries contain a lot of juice, but few strawberries. A team from the supermarket make a surprise visit to the supplier who is about to deliver another batch. Each tin in this batch is claimed to have a minimum of 300 g of fruit, but a random sample of 15 tins found only 287 g with a standard deviation of 18 g. What conclusion can the supermarket make?

Solution

1 *State the null and alternative hypotheses.* The null hypothesis is that the mean weight of the fruit is 300 g, while the alternative hypothesis is that the weight is less than this.

$$H_0 : \mu = 300 \text{ g} \quad H_1 : \mu < 300 \text{ g}$$

2 *Specify the level of significance.* We can use the standard 5%.
3 *Calculate the acceptance range for the variable tested.* With a sample of size 15, the sampling

distribution of the mean follows a *t*-distribution with $n - 1 = 15 - 1 = 14$ degrees of freedom and a mean of 300 g. The estimated standard error is $s / \sqrt{(n - 1)} = 18 / \sqrt{14} = 4.81$ g. For a 5% significance level and a one-sided test, we look up (either in tables or using the TINV function in a spreadsheet) a probability of 0.05. With 14 degrees of freedom the value is 1.761. Then the acceptance range is above $300 - 1.761 \times 4.81 = 291.53$ g.

4 *Find the actual value for the variable tested.* The actual mean of the sample was 287 g.
5 *Decide whether or not to reject the null hypothesis.* The actual value is outside the acceptance range, so we reject the null hypothesis.
6 *State the conclusion.* The evidence from the sample does not support the view that the mean weight is 300 g. It supports the alternative hypothesis that the mean weight of fruit is less than 300 g.

Review questions

15.8 Why do we not use the Normal distribution for small samples?

15.9 What shape is the *t*-distribution for large samples?

15.10 When would you use the approximation $s / \sqrt{(n - 1)}$ for $\sigma / \sqrt{n}$?

Testing other hypotheses

So far we have focused on hypothesis tests for population means, but we can use the same approach for a variety of other problems.

Population proportions

In Chapter 14 we mentioned that sampling could test the proportion of a population that shared some common feature. In particular, we used the standard result that when the proportion in the population is π, the sampling distribution of the proportion has a mean of π and standard deviation of $\sqrt{(\pi(1 - \pi) / n)}$. Now we can use this result to test hypotheses about proportions.

WORKED EXAMPLE 15.10

High street banks claim that they lend the money for 20% of all house purchases. To test this, a sample of 100 people with mortgages was interviewed, 15 of whom arranged their loan through a bank. Does this support the original claim?

Solution

Hypothesis tests always use the same procedure, and the only difference with this problem is that we are interested in a proportion, π, rather than a mean.

1 *State the null and alternative hypotheses*. The null hypothesis is that banks lend 20% of funds for mortgages, so using proportions we have:

$$H_0 : \pi = 0.2 \quad H_1 : \pi \neq 0.2$$

2 *Specify the level of significance*. This is not given, so we assume 5%.

3 *Calculate the acceptance range for the variable tested*. With a sample of 100, the sampling distribution is Normal with mean 0.2 and standard deviation $\sqrt{(\pi(1 - \pi) / n)} = \sqrt{(0.2 \times 0.8) / 100}$ = 0.04. For a 5% significance level we want points that are within 1.96 standard deviations of the mean. Then the acceptance range is:

$$0.2 - 1.96 \times 0.04 \quad \text{to} \quad 0.2 + 1.96 \times 0.04$$

or

$$0.122 \quad \text{to} \quad 0.278$$

4 *Find the actual value for the variable tested*. The sample had a proportion of 15 / 100 = 0.15.

5 *Decide whether or not to reject the null hypothesis*. The actual value is within the acceptance range, so we cannot reject the null hypothesis.

6 *State the conclusion*. We cannot reject the claim that banks lend money for 20% of mortgages.

Testing for differences in means

Managers often want to compare two populations, to see if there are significant differences. For example, they might have two shops and want to know whether each has the same profitability or not, or they might want to check sales before and after an advertising campaign.

We can use hypothesis testing to see if the means of two populations are the same. For this we take a sample from each population, and if the sample means are fairly close we can assume that the population means are the same, but if there is a large difference in the sample means we have to assume that the population means are different. So the procedure is to take the means of two samples, $\bar{x}_1$ and $\bar{x}_2$, and find the difference, $\bar{x}_1 - \bar{x}_2$. Then we use a standard result that for large samples the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is Normal with:

$$\text{mean} = 0 \quad \text{and} \quad \text{standard error} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where: n_1 = sample size from population 1
n_2 = sample size from population 2
s_1 = standard deviation of sample 1
s_2 = standard deviation of sample 2.

WORKED EXAMPLE 15.11

Krinkle Kut Krisps uses two machines to fill packets of crisps. A sample of 30 packets from the first machine has a mean weight of 180 g and a standard deviation of 40 g. A sample of 40 packets from the second machine has a mean weight of 170 g and a standard deviation of 10 g. Are the two machines putting the same amount in packets?

Solution

1 *State the null and alternative hypotheses.* We want to check that the two machines are putting the same amounts in packets, so the null hypothesis is that the means from each machine are the same. The alternative hypothesis is that the means are not the same.

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

2 *Specify the level of significance.* We can use the standard 5%.
3 *Calculate the acceptance range for the variable tested.* We are looking at the sampling distribution of $\bar{x}_1 - \bar{x}_2$, with sample sizes $n_1 = 30$ and $n_2 = 40$, and standard deviations $s_1 = 14$ and $s_2 = 10$. This sampling distribution is Normal with:

mean = 0   and

$$\text{standard error} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= \sqrt{\frac{14^2}{30} + \frac{10^2}{40}}$$

$$= 3.01$$

For a 5% significance level and a two-sided test, the acceptance range is within 1.96 standard deviations of the mean. This defines the range:

$$0 - 1.96 \times 3.01 \quad \text{to} \quad 0 + 1.96 \times 3.01$$

or

$$-5.90 \qquad \text{to} \quad +5.90$$

4 *Find the actual value for the variable tested.* The observed difference in samples is $\bar{x}_1 - \bar{x}_2$ = 180 − 170 = 10.
5 *Decide whether or not to reject the null hypothesis.* The actual value is outside the acceptance range, so we reject the null hypothesis.
6 *State the conclusion.* The evidence from the samples does not support the view that the mean weight put into packets is the same from each machine.

Paired tests

If you want to see whether a diet works or not, then you will weigh a set of people before the diet, and weigh them again after the diet. This gives a set of paired data – two weights for each person in your test – and you want to see if there is a difference between the two. This is the kind of problem that managers meet when, for example, they interview people before and after an advertising campaign, or to see if two people interviewing candidates for a job give different opinions.

To test for differences between paired observations, we find the difference between each pair. If the two sets of observations are similar, the mean difference should be around zero – but if there is a real distinction between the observations the mean difference becomes bigger. So we use hypothesis testing to see whether or not the differences between samples are small enough to suggest that the two samples are the same, or are big enough to suggest that they are different.

WORKED EXAMPLE 15.12

Amethyst Interviews counted the number of interviews that a sample of eight of their staff did in a day. Then they adjusted the way the questions were presented, and again counted the number of interviews the eight staff did. From the following results, can you say whether or not the adjustments had any effect?

Interviewer	1	2	3	4	5	6	7	8
Original interviews	10	11	9	6	8	10	7	8
Later interviews	10	9	11	10	9	12	9	11

Solution

Here we subtract the number of original interviews from the number of later interviews to get:

Interviewer	1	2	3	4	5	6	7	8
Difference	0	−2	2	4	1	2	2	3

Now we can use the standard approach on the sample differences.

1 *State the null and alternative hypotheses.* We want to test the null hypothesis that the mean difference is zero, and the alternative hypothesis that the mean difference is not zero.

$$H_0 : \mu = 0 \quad H_1 : \mu \neq 0$$

2 *Specify the level of significance.* We use the standard 5%.

3 *Calculate the acceptance range for the variable tested.* Using the basic definitions, the mean of the differences is $(0 − 2 + 2 + 4 + 1 + 2 + 2 + 3) / 8 = 1.5$, the variance is $(1.5^2 + 3.5^2 + 0.5^2 + 2.5^2 + 0.5^2 + 0.5^2 + 0.5^2 + 1.5^2) / 8 = 3.0$, and the standard deviation is $\sqrt{3.0} = 1.732$. With a small sample of eight pairs of observations, the sampling distribution is a t-distribution with $8 − 1 = 7$ degrees of freedom and standard error $s / \sqrt{(n − 1)} = 1.732 / \sqrt{7} = 0.655$. For a two-tail 5% significance level, the t-distribution with 7 degrees of freedom is 2.365. So the acceptance range is:

$$0 − 2.365 \times 0.655 \quad \text{to} \quad 0 + 2.365 \times 0.655$$

or

$$−1.548 \quad \text{to} \quad 1.548$$

4 *Find the actual value for the tested variable.* The mean of the differences is 1.5.

5 *Decide whether or not to reject the null hypothesis.* The actual value is within the acceptable range, so we cannot reject the null hypothesis.

6 *State the conclusion.* The evidence says that we cannot reject the view that there is no difference between the number of interviews before and after the adjustment. This is an interesting result, as it seems fairly clear that the adjustments have made a difference – but the explanation is that the sample size is very small.

We have done the calculations for hypothesis testing by hand but, as always, we could have used a computer. Figure 15.6 shows results from putting data from Worked example 15.12 into Excel's data analysis option which does calculations for paired samples. Here the data is on the left, and the analysis is on the right. You can see that the computer presents the results in a slightly different way. We have calculated the limits within which we accept the null hypothesis, and then we see if the actual value lies within these limits. An alternative is to state the number of standard errors the acceptable range is from the mean, and then find how many standard errors the actual value is away from the mean. In this example, the 5% significance level sets the acceptable range for a two-tail test as within 2.3646 standard errors of the mean (called the critical t value), while the actual value is $1.5 / 0.655 = 2.29$ standard errors from the mean. As the actual value is within the acceptable range, we cannot reject the null hypothesis. However, you can see

	A	B	C	D	E	F
1	**Paired tests**					
2						
3	**Amethyst Interviews**					
4						
5	**Data**			**t-test: paired two sample for means**		
6	**Original interviews**	**Later interviews**			**Original interviews**	**Later interviews**
7	10	10		Mean	8.625	10.125
8	11	9		Variance	2.8393	1.2679
9	9	11		Observations	8	8
10	6	10		Hypothesised mean difference	0	
11	8	9		Degrees of freedom	7	
12	10	12		t Statistic	2.291	
13	7	9		$P(T<=t)$ one-tail	0.0279	
14	8	11		t Critical one-tail	1.8946	
15				$P(T<=t)$ two-tail	0.0557	
16				t Critical two-tail	2.3646	

Figure 15.6 Spreadsheet of paired test for Amethyst Interviews (Worked example 15.12)

that the probability that the value is within the acceptable range is only 0.0557, so it only just passes at this significance level.

Review questions

15.11 In what circumstances would you use hypothesis tests?

15.12 Is the significance level the same as a confidence interval?

IDEAS IN PRACTICE **Automatic testing systems**

Many automatic systems are programmed to make decisions under conditions of uncertainty. Then they check a response and follow prescribed rules to reach a decision. Because there is uncertainty, there is always the chance that they make a Type I or Type II error.

You can imagine this with an airport security system that is designed to detect passengers carrying weapons. When you walk through an airport metal detector, the detector has a null hypothesis that you are carrying a weapon, and it takes electromagnetic measurements to test this hypothesis. When it does not have enough evidence to support the null hypothesis it lets you through; when the evidence supports the null hypothesis it stops you. But there is uncertainty in its readings, caused by other things that you might be carrying,

and it makes two kinds of error. Either it stops people who are not carrying a weapon, or it does not stop people who are.

Another example is e-mail spam filters, which block junk messages or those that are somehow objectionable. The filter has a null hypothesis that a message is spam, and then examines the contents to test this hypothesis. If it finds evidence to support the hypothesis that the message is spam (the presence of key words, patterns in senders' address, multiple copies transmitted, types of content etc.), the filter blocks the message.

You can find many other examples of such systems, including Internet search engines, automatic recorders of earthquakes, burglar alarms, roadside speed cameras and so on.

Chi-squared test for goodness of fit

Variables that take a specific value during an investigation are called parameters, and hypothesis tests that concern the value of a parameter are **parametric tests**. Often we want to test a hypothesis, but there is no appropriate variable – or parameter – to measure. You can imagine this with nominal data like the type of industry, value, quality, colour and so on. For instance, you might suggest a hypothesis that one type of product offers better value than another – but there is no convincing parameter you can use to measure value. When you cannot use parametric hypothesis tests, you have to use the alternative **non-parametric**, or **distribution-free, tests**. These have the major benefit of making no assumptions about the distribution of the population – and the weakness of being somewhat less specific.

The most important non-parametric test is the **chi-squared** or χ^2 **test** (χ is the Greek letter chi – pronounced 'kie', rhyming with 'lie'). This is still a hypothesis test, so its general approach is the same as that of parametric tests. The difference is that it looks at the *frequencies* of observations and sees whether or not these match the expected frequencies.

Suppose you form a hypothesis about the distribution of values for some variable. Then you may expect a distribution with frequencies $E_1, E_2, E_3, \ldots , E_n$, but when you actually check the values you get a series of observations $O_1, O_2, O_3, \ldots , O_n$. The difference between these shows how closely the actual observations match your expectations. Squaring the difference between observed and expected frequencies removes negative values, and then dividing by the expected frequencies gives a distribution with a standard shape. In other words, we define χ^2 as:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \ldots + \frac{(O_n - E_n)^2}{E_n}$$

chi-squared $= \chi^2 = \sum \dfrac{(O - E)^2}{E}$

When observed frequencies are close to the expected frequencies, χ^2 is close to zero; but when there are bigger differences between observed and expected frequencies, χ^2 has a larger value. So for this type of test, we define a **critical value** for χ^2 – when the actual value is above this we say that observations do match our expectations, and reject the hypothesis; when the actual value is below the critical value we cannot reject the hypothesis. There are standard tables of critical values (shown in Appendix F) or you can use standard software, such as Excel's CHIINV function.

Like the *t*-distribution described in Chapter 14, the shape of the χ^2 distribution depends on the degrees of freedom. As we have seen, these measure the number of pieces of information that are free to take any value. Without going into the detailed reasoning, we will simply say that the number of degrees of freedom for a χ^2 distribution is:

degrees of freedom = number of classes − number of estimated
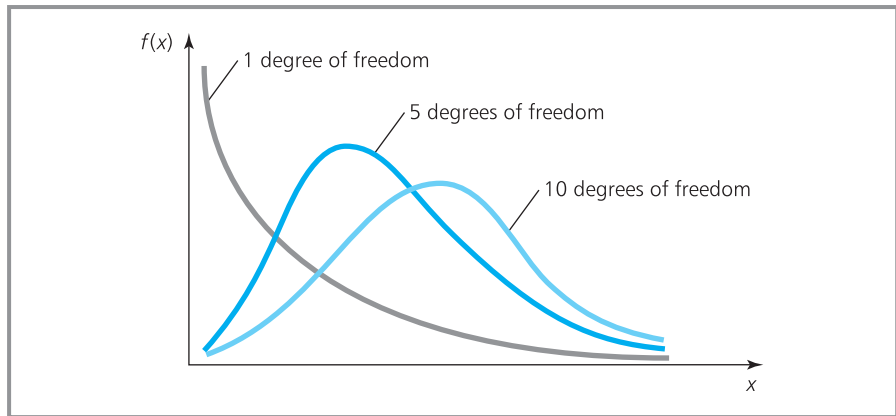variables − 1

Figure 15.7 Chi-squared distribution with varying degrees of freedom

The following examples show how to use this, and Figure 15.7 compares the shapes of a χ^2 distribution with different degrees of freedom.

WORKED EXAMPLE 15.13

Over the past three years, five supermarkets have recorded the following numbers of minor accidents:

Supermarket	1	2	3	4	5
Number of accidents	31	42	29	35	38

Does this suggest that some supermarkets have more accidents than others?

Solution

We use the standard procedure for hypothesis testing.

1 *State the null and alternative hypotheses.* The null hypothesis, H_0, is that each supermarket has the same number of accidents. The alternative hypothesis, H_1, is that each supermarket does not have the same number of accidents.
2 *Specify the level of significance.* We can take this as 5%.
3 *Calculate the critical value of χ^2.* In this problem there are five classes, and no variables have been estimated, so the degrees of freedom are $5 - 0 - 1 = 4$. With a 5% significance level, we look in the '0.05' column of χ^2 tables and with 4 degrees of freedom we find a critical value of

9.49 (or we could get the same value using the CHIINV function in Excel).

4 Find the actual value of χ^2, where:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

There are a total of 175 accidents. If each supermarket expects the same number of accidents, each expects $175 / 5 = 35$. Then we do the following calculations, which show that the actual value of χ^2 is 3.143.

Supermarket	O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
1	31	35	−4	16	0.457
2	42	35	7	49	1.400
3	29	35	−6	36	1.029
4	35	35	0	0	0.000
5	38	35	3	9	0.257
Total	175	175			3.143

5 *Decide whether or not to reject the null hypothesis.* The actual value (3.143) is less than the critical value (9.4877), so we cannot reject the null hypothesis.
6 *State the conclusion.* The evidence supports the view that each supermarket has the same number of accidents, with any variation due to chance.

Worked example 15.13 effectively tested whether or not there was a uniform distribution of accidents across supermarkets. In other words, we hypothesised that the accident rate was uniformly distributed and checked if the data fitted this. We can use the same approach for other distributions, to see whether or not a set of data follows a specific distribution.

WORKED EXAMPLE 15.14

MailFast asked 10 people in each of 100 postcode areas if they would be prepared to pay more for a faster delivery service. They collected the following results. Do these follow a binomial distribution?

Number in sample willing to pay higher charges	0	1	2	3	4
Number of postcodes	26	42	18	9	5

Solution

If we define success as being willing to pay more and failure as being unwilling to pay, we have a binomial process. Then the number of successes in 10 trials should follow a binomial distribution, and we can test this using the standard procedure.

1 *State the null and alternative hypotheses.* The null hypothesis, H_0, is that the distribution of successes is binomial. The alternative hypothesis, H_1, is that the distribution is not binomial.
2 *Specify the level of significance.* We take this as 5%.
3 *Calculate the critical value of χ^2.* There are 5 classes – but we also need to find one parameter for the binomial distribution, which is the probability of success. So the number of degrees of freedom is:

   number of classes – number of estimated parameters – 1 = 5 – 1 – 1 = 3

   Looking up the critical value for a significance level of 5% and 3 degrees of freedom gives a critical value of 7.81.

4 *Find the actual value of χ^2.* We have 10 results from each of 100 postcodes, giving 1,000 opinions, and of these the total number willing to pay more is:

   $(26 \times 0) + (42 \times 1) + (18 \times 2) + (9 \times 3) + (5 \times 4)$
   $= 125$

So the probability that someone is willing to pay more is 125 / 1,000 = 0.125. Now we can find the probabilities of 0, 1, 2, 3, 4 etc. successes out of 10 when the probability of success is 0.125, using the binomial calculations described in Chapter 13 (or the tables in Appendix B or statistical software). The probability of no successes is 0.263, the probability of one success is 0.376, the probability of two successes is 0.242 and so on. Multiplying these probabilities by the number of postcodes, 100, gives the expected number of postcodes with each number willing to pay higher charges. For convenience, we have taken the last class as '4 or more'.

Number willing to pay more	0	1	2	3	4 or more
Probability	0.263	0.376	0.242	0.092	0.027
Expected number of postcodes	26.3	37.6	24.2	9.2	2.7

Now we have both the expected and observed distributions of postcodes, and can calculate the value of χ^2.

Frequency	O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
0	26	26.3	−0.3	0.09	0.003
1	42	37.6	4.4	19.36	0.515
2	18	24.2	−6.2	38.44	1.588
3	9	9.2	−0.2	0.04	0.004
4 or more	5	2.7	2.3	5.29	1.959
Total	100	100			4.069

5 *Decide whether or not to reject the null hypothesis.* The actual value of χ^2 (4.069) is less than the critical value (7.81), so we cannot reject the null hypothesis.
6 *State the conclusion.* The evidence supports the view that the observations follow a binomial distribution.

WORKED EXAMPLE 15.15

Performance Cables record the following numbers of faults per kilometre in a cable. Does this data follow a Poisson distribution?

Number of faults	0	1	2	3	4	5	6
Number of kilometres	37	51	23	7	4	2	1

Solution

1 *State the null and alternative hypotheses*. The null hypothesis, H_0, is that the distribution is Poisson. The alternative hypothesis, H_1, is that the distribution is not Poisson.
2 *Specify the level of significance*. We take this as 5%.
3 *Calculate the critical value of χ^2*. A small problem here is that the χ^2 distribution does not work well with expected frequencies of less than 5. We should really combine small adjacent classes so that the expected number of observations becomes greater than 5. Because of this adjustment, we will return to the critical value of χ^2 a little later.
4 *Find the actual value of χ^2*. We have 125 kilometres' worth of data, and the total number of defects is:

$$(37 \times 0) + (51 \times 1) + (23 \times 2) + (7 \times 3) + (4 \times 4) + (2 \times 5) + (1 \times 6) = 150$$

The mean number of defects is 150 / 125 = 1.2 per kilometre. Using this as the mean of a

Poisson distribution, we can find the expected number of defects per kilometre in the usual ways. Then multiplying these probabilities by the number of observations, 125, gives the expected frequency distribution of defects. Figure 15.8 shows a spreadsheet with these results in columns A to C.

Now we should really combine adjacent classes so that each class has more than 5 observations. Adding the last four classes gives the revised table of expected values in column D, with subsequent calculations in columns F and G. Now you can see why we delayed the calculation of the critical value; there are now only four classes, so the number of degrees of freedom is:

number of classes – number of estimated parameters – 1 = 4 – 1 – 1 = 2

Here the parameter estimated is the mean of the distribution. Looking up the critical value for a significance level of 5% and 2 degrees of freedom gives a value of 5.99.

5 *Decide whether or not to reject the null hypothesis*. The actual value (1.459) is less than the critical value (5.99), so we cannot reject the null hypothesis.
6 *State the conclusion*. The evidence supports the view that the observations follow a Poisson distribution.

	A	B	C	D	E	F	G
1	Chi-squared test						
2							
3	Number of defects	Probability	Expected frequency	Revised frequency (E)	Actual frequency (O)	(O – E)	(O – E)²/E
4	0	0.301	37.649	37.649	37	−0.649	0.011
5	1	0.361	45.179	45.179	51	5.821	0.750
6	2	0.217	27.107	27.107	23	−4.107	0.622
7	3	0.087	10.843	15.064	14	−1.064	0.075
8	4	0.026	3.253				
9	5	0.006	0.781				
10	>=6	0.002	0.188				
11	Totals	1.000	125.000	125.000	125.000		1.459

Figure 15.8 Spreadsheet calculations for the chi-squared test in Worked example 15.15

Review questions

15.13 What is the main difference between a parametric test and a non-parametric test?

15.14 When would you use a non-parametric test?

15.15 'When you cannot use a parametric test, you can always use a non-parametric test instead.' Do you think this is true?

15.16 Why does a χ^2 test have only a critical value rather than an acceptance range?

15.17 What is χ (the square root of χ^2) used for?

IDEAS IN PRACTICE   Humbolt Farm Products

Humbolt Farm Products routinely analyse the weights of materials received from suppliers. One type of delivery is known to have a mean weight of 45 g and a standard deviation of 15 g, but a sample of 500 units had this distribution:

Weight (in grams)	Number of observations	Weight (in grams)	Number of observations
less than 10	9	40 to 49.99	115
10 to 19.99	31	50 to 59.99	94
20 to 29.99	65	60 to 69.99	49
30 to 39.99	97	70 to 79.99	24
		80 to 89.99	16

Many types of statistical analyses are valid only if the data are Normally distributed, so Humbolt test this using χ^2 tests. For this example, they use the standard procedure as follows.

1 The null hypothesis, H_0, is that the distribution is Normal; the alternative hypothesis, H_1, is that the distribution is not Normal.
2 The significance level is 5%.
3 The number of degrees of freedom is $9 - 0 - 1 = 8$, giving a critical value of $\chi^2 = 15.5$.

4 The probability that an observation is in the range, say, 10 to 19.99 is:

$$P(\text{between 10 and 19.99}) = P(\text{less than 20}) - P(\text{less than 10})$$

Now 20 is $(20 - 45) / 15 = -1.67$ standard deviations from the mean, which corresponds to a probability of 0.048, and 10 is $(10 - 45) / 15 = -2.33$ standard deviations from the mean, which corresponds to a probability of 0.010, so:

$$P(\text{between 10 and 19.99}) = 0.048 - 0.010 = 0.038$$

The expected number of observations in this range is $0.038 \times 500 = 19$. Figure 15.9 shows a spreadsheet with these calculations, leading to a calculated value of χ^2 as 43.56.
5 The actual value of χ^2 (43.56) is greater than the critical value (15.5), so they reject the hypothesis that the sample is Normally distributed.
6 The evidence does not support the view that observations follow a Normal distribution, so Humbolt have to be careful with the analyses they do on this sample.

Ideas in practice continued

	A	B	C	D	E	F
1	Chi-squared test – Humbolt Farm Products					
2						
3	Weight from	to	Probability	Expected frequency	Observed frequency	$(O - E)^2/E$
4	0	10.00	0.010	4.9	9	3.41
5	10	20.00	0.038	19.0	31	7.61
6	20	30.00	0.111	55.4	65	1.65
7	30	40.00	0.211	105.4	97	0.67
8	40	50.00	0.261	130.5	115	1.85
9	50	60.00	0.211	105.4	94	1.23
10	60	70.00	0.111	55.4	49	0.75
11	70	80.00	0.038	19.0	24	1.32
12	80	90.00	0.010	4.9	16	25.07
13			1.000	500.0	500	**43.56**

Figure 15.9 Calculations for Humbolt Farm Products (Worked example 15.16)

Tests of association

Suppose that you have a set of data that can be classified in two ways, such as a mortgage that can be classified according to both the amount borrowed and the type of lender. Then you can use a frequency table – called a **contingency table** – to show the number of observations that fall into each category. And a χ^2 test checks for associations between the two classifications. For example, if you build the following contingency table to show the number of mortgages with particular sizes and sources, you might ask whether or not there is any association between the size of a loan and the source.

		Size of loan			Total
		Less than £80,000	£80,000 to £150,000	More than £150,000	
Source of mortgage	Building society	30	55	40	**125**
	Bank	23	29	3	**55**
	Elsewhere	12	6	2	**20**
Total		**65**	**90**	**45**	**200**

WORKED EXAMPLE 15.16

Is there any association between the size of loan and its source in the contingency table above?

Solution

This is a hypothesis test, so we can use the standard procedure.

1 *State the null and alternative hypotheses.* When testing for association you normally use a null hypothesis that there is no association. Then the null hypothesis, H_0, is that there is no association between the size of mortgage and its source – in other words, the two are independent. The alternative hypothesis, H_1, is that there is an association.

2 *Specify the level of significance.* We take this as 5%.

3 *Calculate the critical value of χ^2.* For a contingency table the number of degrees of freedom is:

degrees of freedom = (number of rows – 1) × (number of columns – 1)

Here there are three rows and three columns (ignoring the totals), so there are

$(3 - 1) \times (3 - 1) = 4$ degrees of freedom

Looking up χ^2 tables for a 5% significance level and 4 degrees of freedom gives a critical value of 9.49.

4 *Find the actual value of χ^2.* For this we have to calculate the expected number of replies in each cell of the matrix. Starting at the top left-hand cell, we have the number of people who have a loan of less than £80,000 from a building society. A total of 125 loans come from a building society, so the probability that a particular loan comes from a building society is 125 / 200 = 0.625. A total of 65 loans are less than £80,000, so the probability that a particular loan is less than £80,000 is 65 / 200 = 0.325. Then the probability that a loan comes from a building society and is less than £80,000 is 0.625 × 0.325 = 0.203. Since there are 200 loans, the expected number of this type is 0.203 × 200 = 40.625. Repeating this calculation for every cell in the matrix gives these results:

		Size of loan			Total
		Less than £80,000	£80,000 to £150,000	More than £150,000	
Source of mortgage	Building society	40.625	56.250	28.125	**125**
	Bank	17.875	24.750	12.375	**55**
	Elsewhere	6.500	9.000	4.500	**20**
Total		**65**	**90**	**45**	**200**

Now we have a set of 9 observed frequencies, and a corresponding set of 9 expected frequencies. When we calculated the expected frequencies we assumed that there is no connection between the loan size and its source. Any significant differences between expected and observed values are caused by an association between the loan size and its source. The closer the association, the larger is the difference, and the larger the calculated value of χ^2. So now we have to find the actual value of χ^2, which you can see from the following table is 24.165.

O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
30	40.625	– 10.625	112.891	2.779
55	56.250	−1.25	1.563	0.028
40	28.125	11.875	141.016	5.014
23	17.875	5.125	26.266	1.469
29	24.750	4.25	18.063	0.730
3	12.375	−9.375	87.891	7.102
12	6.500	5.500	30.250	4.654
6	9.000	−3.000	9.000	1.000
2	4.500	−2.500	6.250	1.389
200	200			24.165

5 *Decide whether or not to reject the null hypothesis.* The actual value (24.165) is greater than the critical value (9.49), so we reject the null hypothesis and accept the alternative hypothesis.

6 *State the conclusion.* The evidence supports the view that there is an association, and the size of a mortgage is related to its source.

Statistical packages can be quite difficult to use, confusing and give results that are difficult to

Worked example 15.16 continued

interpret. Figure 15.10 shows the printout from a simple package for this problem. This follows the steps that we have used, but notice that it gives a warning that one cell has fewer than 5 expected observations. We should be careful about this because it means that a chi-squared test may not work well, and we should really combine this cell with others to raise the number of expected observations. Alternatively, we could simply collect more observations.

```
001 > read data columns c1–c3
002 > 30 55 40
003 > 23 29 3
004 > 12 6 2
005 > column titles <80, 80–150, >150
006 > row titles building society, bank, elsewhere
007 > end data
008 > chisquare calculate c1–c3
```

Expected counts are printed below observed counts

	<80	80–150	>150	Total
building society	30	55	40	125
	40.62	56.25	28.12	
bank	23	29	3	55
	7.88	24.75	12.37	
elsewhere	12	6	2	20
	6.50	9.00	4.50	
Total	65	90	45	200

```
ChiSq = 2.779 + 0.028 + 5.014 +
        1.469 + 0.730 + 7.102 +
        4.654 + 1.000 + 1.389 = 24.165
df = 4

* WARNING *
1 cell with expected counts less than 5.0
   Merge rows or columns?
09 > no

010 > significance = 0.05
011 > chisquare test c1–c3

Critical value of ChiSq = 9.488
Calculated value of ChiSq = 24.165

Conclusion = reject null hypothesis
```

Figure 15.10 Printout from statistics package for chi-squared test

15.18 What is a test of association?

15.19 Why would you use a statistical package for χ^2 tests?

Reports of findings

Newspapers find it difficult to report statistical information – they realise that most people will not fully understand it. Rather than confusing or boring their readers, they give heavily summarised results. You can see this effect in most reports, including the results from hypotheses tests. For example, one study found evidence that women who smoke during pregnancy have children who subsequently perform less well in school exams. The newspapers reported the headline results, 'Children born to women who smoke during pregnancy score an average of 9.2% lower marks in exams'. But it omitted the associated analysis, assumptions, significance levels and so on.

One survey of weight loss found that 47 people who only followed diets for six months lost 5.9 kg, while 43 people who only exercised lost 4.0 kg. The difference seems relatively large, but several magazines interpreted the results to mean that both methods are equally successful. However, the full study included a hypothesis test, using the standard deviations of weight loss (4.0 kg for diets and 3.9 kg for exercise). With a null hypothesis that the difference in mean weight loss between the two samples is zero, and an alternative hypothesis that the difference in mean weight loss is not zero, this result would occur in only 3% of tests. A fairer conclusion is that there is a statistically significant difference between the average weight loss for the two methods.

Abbreviated reports commonly do not give the full story about, for example, the effectiveness of alternative medicines, the effects of drinking alcohol, public opinions about government policies, customer reaction to new products, success rates of treatments for illnesses and so on.

CHAPTER REVIEW

This chapter described the approach of hypothesis testing, which reveals whether or not a statement about a population is supported by the evidence in a sample.

- Hypothesis testing starts with a null hypothesis, which is a precise statement about a population. Then it tests a sample from the population to see if there is evidence to support the null hypothesis. If the evidence does not support the null hypothesis it is rejected, otherwise it cannot be rejected.

- Samples always involve uncertainty and in hypothesis testing there are two types of error: Type I errors reject a null hypothesis that is true; and Type II errors do not reject a null hypothesis that is false.

- A significance level is the minimum acceptable probability that a value is a random sample from the hypothesised population. It is equivalent to the probability of making a Type I error.

- A common use of hypothesis testing checks whether or not the mean of a population has a specified value. A two-sided test checks the range within which a population mean is likely to lie; a one-sided test checks whether the population mean is likely to be above or below a specified value.

- The standard analysis can be extended to deal with small samples using a t-distribution, proportions of a population sharing some feature, differences between means and paired observations.

- When there is no parameter to test, typically with nominal data, we have to use a distribution-free, or non-parametric, test. The most common approach uses chi-squared tests, which check whether or not data follows a specified distribution. Chi-squared distribution can also be used to test the association between two parameters in a contingency table.

Willingham Consumer Protection Department

Willingham Consumer Protection Department (WCPD) is responsible for administering all weights and measures laws in its area of North Carolina. A part of its service makes sure that packages of food and drink contain the quantities stated. One week, WCPD decided to test containers of milk. Most of these tests were done at dairies, where procedures and historical data were also examined, with other random samples taken from local shops and milk delivery services.

On two consecutive days WCPD bought 50 containers with a nominal content of 4 pints or 2.27 litres. The actual contents of these, in litres, are as follows.

- Day 1: 2.274 2.275 2.276 2.270 2.269 2.271
          2.265 2.275 2.263 2.278 2.260 2.278
          2.280 2.275 2.261 2.280 2.279 2.270
          2.275 2.263 2.275 2.281 2.266 2.277
          2.271 2.273 2.283 2.260 2.259 2.276
          2.286 2.275 2.271 2.273 2.291 2.271
          2.269 2.265 2.258 2.283 2.274 2.278
          2.276 2.281 2.269 2.259 2.291 2.289
          2.276 2.283
- Day 2: 2.270 2.276 2.258 2.259 2.281 2.265
          2.278 2.270 2.294 2.255 2.271 2.284
          2.276 2.293 2.261 2.270 2.271 2.276
          2.269 2.268 2.272 2.272 2.273 2.280

2.281 2.276 2.263 2.260 2.295 2.257
2.248 2.276 2.284 2.276 2.270 2.271

When they were collecting these figures, WCPD inspectors were convinced that there were no problems with the main dairies, but some small operations were not so reliable. This was because large dairies invariably used modern, well-designed equipment, and they employed special quality assurance staff. Smaller operators tended to use older, less reliable equipment and could not afford to run a quality assurance department. Two companies, in particular, were identified as needing further checks. WCPD took random samples of 15 containers from each of these dairies, with these results:

- Company 1: 2.261 2.273 2.250 2.268 2.268
              2.262 2.272 2.269 2.268 2.257
              2.260 2.270 2.254 2.249 2.267
- Company 2: 2.291 2.265 2.283 2.275 2.248
              2.286 2.268 2.271 2.284 2.256
              2.284 2.255 2.283 2.275 2.276

Question

- **What could the milk inspectors report about their findings? What follow-up action could they recommend? Are there any improvements they could make to their data collection and analysis?**

PROBLEMS

15.1 The mean wage of people living in Alto Canillas is said to be £400 a week with a standard deviation of £100. A random sample of 36 people was examined. What is the acceptance range for a 5% significance level? What is the acceptance range for a 1% significance level?

15.2 The weight of packets of biscuits is claimed to be 500 g. A random sample of 50 packets has a mean weight of 495 g and a standard

deviation of 10 g. Use a significance level of 5% to see whether or not the data from the sample supports the original claim.

15.3 Hamil Coaches Ltd say that their long-distance coaches take 5 hours for a particular journey. Last week a consumer group tested these figures by timing a sample of 30 journeys. These had a mean time of 5 hours 10 minutes with a standard deviation of 20 minutes. What report can the consumer group make?

15.4 A food processor specifies the mean weight of a product as 200 g. A random sample of 20 has a mean of 195 g and a standard deviation of 15 g. Does this evidence suggest that the mean weight is too low?

15.5 An emergency breakdown service suggests that 50% of all drivers are registered with their service. A random sample of 100 people had 45 who were registered. Does this sample support the original claim?

15.6 Quality Managers at CentralGen say that 12% of the letters they post contain errors. A sample of 200 letters was checked and 31 of them contained errors. What do these results suggest?

15.7 Health service managers say that doctors should not spend more than 2 hours a day doing paperwork. A sample of 40 doctors spends an average of 2 hours 25 minutes a day doing paperwork, with a standard deviation of 55 minutes. What does this show?

15.8 A mobile phone has an advertised life of 30,000 hours. A sample of 50 phones had a life of 28,500 hours with a standard deviation of 1,000 hours. What can you say about the advertisements?

15.9 Dorphmund Industries have two similar factories. There is some disagreement, because people working in each factory think those in the other factory are getting higher wages. A sample of wages was taken from each factory with the following results:

- Sample 1: size = 45, mean = $250, standard deviation = $45
- Sample 2: size = 35, mean = $230, standard deviation = $40

What can you say about the wages?

15.10 A car manufacturer says that its cars cost €500 a year less to maintain than those of its competitors. To test this, a consumer group found the cost of maintaining 10 cars for a year, and the mean saving was €79 with a standard deviation of €20. What does this say about the manufacturer's claim?

15.11 Five factories reported the following numbers of minor accidents in a year:

Factory	1	2	3	4	5
Number of accidents	23	45	18	34	28

Does this suggest that some factories have more accidents than others?

15.12 The following figures show the number of defective components supplied each day by a factory. Does this data follow a binomial distribution?

Number of defects	0	1	2	3	4	5
Number of days	8	22	33	29	15	3

15.13 The number of road accident victims reporting to a hospital emergency ward is shown in the following table. Do these figures follow a Poisson distribution?

Number of accidents	0	1	2	3	4	5	6
Number of days		17	43	52	37	20	8	4

15.14 Do the following figures follow a Normal distribution?

Weight (in grams)	Number of observations	Weight (in grams)	Number of observations
less than 5	5	65 to 79.99	97
5 to 19.99	43	80 to 94.99	43
20 to 34.99	74	95 to 109.99	21
35 to 49.99	103	110 and more	8
50 to 64.99	121		

15.15 Figure 15.11 shows a spreadsheet doing the calculations for a *t*-test on the mean of two samples. Explain the results and check the calculations. How could you improve the format?

	A	B	C	D	E	F	G
1	Two samples – *t*-test						
2							
3	Data			*t*-Test: Two-Sample assuming equal variances			
4	Variable 1	Variable 2					
5	10	8			*Variable 1*	*Variable 2*	
6	16	10		Mean	10.6	7.786	
7	13	9		Variance	12.489	4.335	
8	6	6		Observations	10	14	
9	8	4		Pooled variance	7.671		
10	9	10		Hypothesised mean difference	0		
11	16	9		df	22		
12	9	9		*t* Stat	2.454		
13	12	7		P(*T*<=*t*) one-tail	0.011		
14	7	4		*t* Critical one-tail	1.717		
15		10		P(*T*<=*t*) two-tail	0.023		
16		9		*t* Critical two-tail	2.074		
17		6					
18		8					

Figure 15.11 Calculations for Problem 15.15

RESEARCH PROJECTS

15.1 This chapter mentioned several examples of automatic systems that implicitly include hypothesis tests – including airport security systems, e-mail spam filters, Internet search results, automatic recorders of earthquakes, burglar alarms and roadside speed cameras. What other examples can you find? How do such systems actually incorporate hypothesis testing?

15.2 Hypothesis testing comes in many different forms, and it always seems to involve judgement. This makes it difficult to design a package that automatically takes data and does an appropriate hypothesis test. Do a small survey to see what facilities statistical packages have for hypothesis testing. How do they get around the practical problems?

15.3 Supermarkets and other retailers often claim that they offer the lowest prices in their area. How can you check their claims? Collect some data from competing stores and analyse the results. What conclusions can you reach?

15.4 Often a hypothesis may seem 'obvious', but on closer examination there is no evidence to support it. Find some real examples of this effect. What are the consequences?

Sources of information

Reference

1 Hooke R., *How to Tell the Liars from the Statisticians*, Marcel Dekker, New York, 1983.

Further reading

There are very few books specifically about hypothesis testing – and the odd ones are very

technical. Three possible ones are given, but it may be better to look in general statistics books, like those listed in Chapter 12.

Lehmann E. and Romano J.P., *Testing Statistical Hypotheses* (3rd edition), Springer, New York, 2010.

Wellek S., *Testing Statistical Hypotheses of Equivalence*, Chapman and Hall / CRC, Boca Raton, FL, 2003.

Wilcox R.R., *Introduction to Robust Estimation and Hypothesis Testing* (2nd edition), Elsevier, Amsterdam, 2005.