Ch06
Hypothesis testing

This chapter deals with issues very similar to those of the previous chapter on estimation, but examines them in a different way. The estimation of population parameters and the testing of hypotheses about those parameters are similar techniques (indeed they are formally equivalent in a number of respects), but there are important differences in the interpretation of the results arising from each method.

6.1 The concepts of hypothesis testing

- Suppose that you are thinking of taking over a small business franchise. The current owner claims the weekly turnover of each existing franchise is £5000 and at this level you are willing to take on a franchise. You would be more cautious if the turnover is less than this figure. You examine the books of 26 franchises chosen at random and find that the average turnover was £4900 with standard deviation £280. What do you do?
- The ***null hypothesis*** in this case is that average weekly turnover is £5000 (or more; that would be even more to your advantage).
- The ***alternative hypothesis*** is that turnover is strictly less than £5000 per week.
- We may write these more succinctly as follows

$$H_0: \mu = 5000 \text{ and } H_1: \mu < 5000$$

- H_0 is conventionally used to denote the null hypothesis, H_1 the alternative.
- Initially, H_0 is presumed to be true and this presumption will be tested using the sample evidence. Note that the sample evidence is not used as part of the hypothesis.

6.1 The concepts of hypothesis testing

You have to decide whether the owner's claim is correct (H_0) or not (H_1). The two types of error you could make are as follows:

- **Type I error** – reject H_0 when it is in fact true. This would mean missing a good business opportunity.
- **Type II error** – not rejecting H_0 when it is in fact false. You would go ahead and buy the business and then find out that it is not as attractive as claimed. You would have overpaid for the business.

Obviously a good decision rule would give a good chance of making a correct decision and rule out errors as far as possible. Unfortunately it is impossible completely to eliminate the possibility of errors.

As the decision rule is changed to reduce the probability of a Type I error, the probability of making a Type II error inevitably increases. The skill comes in balancing these two types of error.

Again a diagram is useful in illustrating this.

Assuming that the null hypothesis is true, then the sample observations are drawn from a population with mean 5000 and some variance, which we shall assume is accurately measured by the sample variance.

6.1 The concepts of hypothesis testing

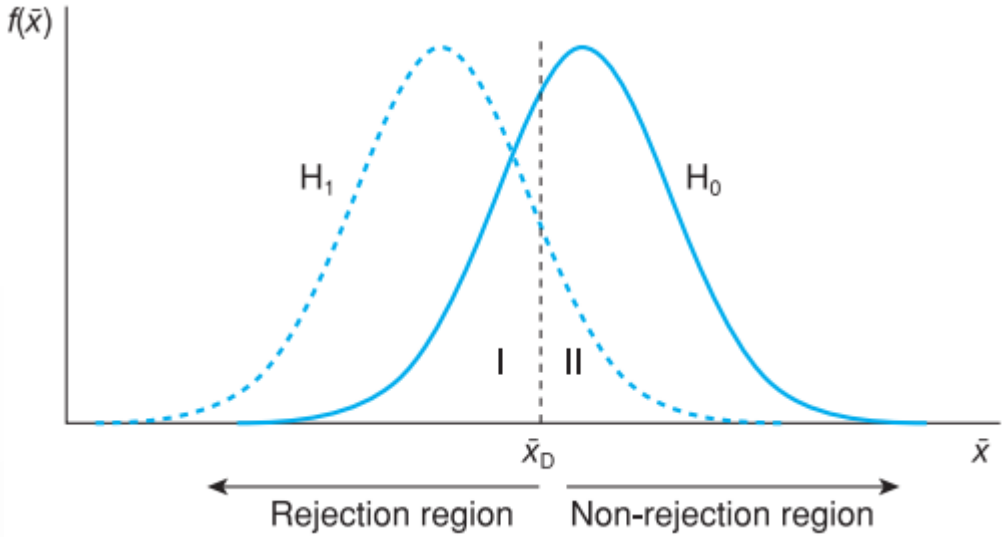
		True Situation	
		H_0 True	H_0 False
Decision	Accept H_0	Correct decision	Type II error
	Reject H_0	Type I error	Correct decision

- The distribution of $\bar{x}$ is then given by

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{or } \bar{x} \sim \left(5000, \frac{(280)^2}{26}\right)$$

- Under the alternative hypothesis the distribution of $\bar{x}$ would be the same except that it would be centered on a value less than 5000.
- These two situations are illustrated in Figure below.
- The distribution of $\bar{x}$ under H_1 is shown by a dashed curve to signify that its exact position is unknown, only that it lies to the left of the distribution under H_0.

6.1 The concepts of hypothesis testing



- A decision rule amounts to choosing a point or dividing line on the horizontal axis in the Figure.
- If the sample mean lies to the left of this point then H_0 is rejected (the sample mean is too far away from H_0 for it to be credible) in favour of H_1 and you do not buy the firm.
- If $\bar{x}$ lies above this decision point then H_0 is not rejected and you go ahead with the purchase.
- Such a decision point is shown in the Figure, denoted by $\bar{x}_D$. To the left of $\bar{x}_D$ lies the rejection (of H_0) region; to the right lies the non-rejection region.

6.1 The concepts of hypothesis testing

- Based on this point, we can see the probabilities of Type I and Type II errors.
- The area under the H_0 distribution to the left of $\bar{x}_D$, labelled I, shows the probability of rejecting H_0 given that it is in fact true: a Type I error.
- The area under the H_1 distribution to the right of $\bar{x}_D$, labelled II, shows the probability of a Type II error: not rejecting H_0 when it is in fact false (and H_1 is true).
- Shifting the decision line to the right or left alters the balance of these probabilities. Moving the line to the right increases the probability of a Type I error but reduces the probability of a Type II error. Moving the line to the left has the opposite effect.
- The Type I error probability can be calculated for any value of $\bar{x}_D$. Suppose we set $\bar{x}_D$ to a value of 4950. Using the distribution of $\bar{x}$ given in equation above, the area under the distribution to the left of 4950 is obtained using the z score

$$z = \frac{\bar{x}_D - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{4950 - 5000}{\sqrt{\frac{280^2}{26}}} = -0.91$$

6.1 The concepts of hypothesis testing

- From the tables of the standard Normal distribution we find that the probability of a Type I error is 18.1%. Unfortunately, the Type II error probability cannot be established because the exact position of the distribution under H_1 is unknown.
- Therefore we cannot decide on the appropriate position of $\bar{x}_D$ by some balance of the two error probabilities.
- The convention therefore is to set the position of $\bar{x}_D$ by using a Type I error probability of 5%, known as the significance level1 of the test.
- In other words, we are prepared to accept a 5% probability of rejecting H_0 when it is, in fact, true. This allows us to establish the position of $\bar{x}_D$.
- From the table, we find that $z = -1.64$ cuts off the bottom 5% of the distribution, so the decision line should be 1.64 standard errors below 5000.
- The value −1.64 is known as the critical value of the test. We therefore obtain

$$\bar{x}_D = 5000 - 1.64\sqrt{\frac{280^2}{26}} = 4910$$

6.1 The concepts of hypothesis testing

- Since the sample mean of 4900 lies below 4910 we reject H0 at the 5% significance level or equivalently we reject with 95% confidence.
- The significance level is generally denoted by the symbol α and the complement of this, given by $1 - \alpha$, is known as the confidence level (as used in the confidence interval).
- An equivalent procedure would be to calculate the *z-score* associated with the sample mean, known as the **test statistic**, and then compare this to the critical value of the test. This allows the hypothesis testing procedure to be broken down into five neat steps.
- (1) Write down the null and alternative hypotheses:

$$H_0: \mu = 5000 \; ; \qquad H_1: \mu < 5000$$

- (2) Choose the significance level of the test, conventionally α = 0.05 or 5%.
- (3) Look up the critical value of the test from statistical tables, based on the chosen significance level. z* = 1.64 is the critical value in this case.
- (4) Calculate the test statistic.
- (5) Decision rule. Compare the test statistic with the critical value: if $z < -z*$ reject H_0 in favour of H_1.

6.1 The concepts of hypothesis testing

- **Example:** A sample of 100 workers found the average overtime hours worked in the previous week was 7.8, with standard deviation 4.1 hours. Test the hypothesis that the average for all workers is 5 hours or less.

- **One-tail and two-tail tests**

- In the above example the rejection region for the test consisted of one tail of the distribution of $\bar{x}$, since the buyer was only concerned about turnover being less than claimed. For this reason it is known as a one-tail test.

- Suppose now that an accountant is engaged to sell the franchise and wants to check the claim about turnover before advertising the business for sale.

- In this case she would be concerned about turnover being either below or above 5000.

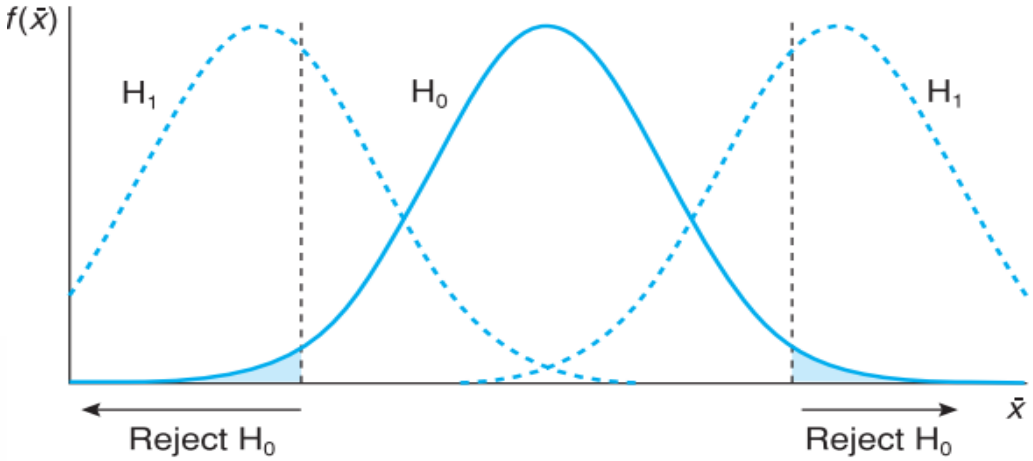
- This would now become a two-tail test with the null and alternative hypotheses being

$$H_0: \mu = 5000 \text{ and } H_1: \mu \neq 5000$$

- Now there are two rejection regions for the test. Either a very low sample mean or a very high one will serve to reject the null hypothesis.

- The situation is presented graphically in the next Figure.

6.1 The concepts of hypothesis testing



- The distribution of $\bar{x}$ under H_0 is the same as before, but under the alternative hypothesis the distribution could be shifted either to the left or to the right, as depicted.
- If the significance level is still chosen to be 5%, then the complete rejection region consist of the two extremes of the distribution under H_0, containing 2.5% in each tail (hence 5% in total).
- This gives a Type I error probability of 5% as before.
- The critical value of the test therefore becomes z* = 1.96, the value which cuts off 2.5% in each tail of the standard Normal distribution. Only if the test statistic falls into one of the rejection regions beyond 1.96 standard errors from the mean is H_0 rejected.

6.1 The concepts of hypothesis testing

Exercise1:

(a) Two political parties are debating crime figures. One party says that crime has increased compared to the previous year. The other party says it has not. Write down the null and alternative hypotheses.

(b) Explain the two types of error that could be made in this example and the possible costs of each type of error.

Exercise2:

(a) We test the hypothesis $H_0: \mu = 100$ against $H_1: \mu > 100$ by rejecting H_0 if our sample mean is greater than 108. If in fact $\bar{x} \sim N(100, 900/25)$, what is the probability of making a Type I error?

(b) If we wanted a 5% Type I error probability, what decision rule should we adopt?

(c) If we knew that μ could only take on the values 100 (under H_0) or 112 (under H_1) what would be the Type II error probability using the decision rule in part (a)?

Exercise3:

Test the hypothesis $H_0: \mu = 500$ versus $H_1: \mu \neq 500$ using the evidence $\bar{x} = 530$, s = 90 from a sample of size n = 30.

6.2 Significance, effect size and power

- Researchers usually look for ‘significant’ results. Academic papers report that ‘the results are significant’ or that ‘the coefficient is significantly different from zero at the 5% significance level’.
- It is vital to realise that the word ‘significant’ is used here in the statistical sense and not in its everyday sense of being important.
- Something can be statistically significant yet still unimportant. Suppose that we have some more data about the business examined earlier.
- Data for 100 franchises have been uncovered, revealing an average weekly turnover of £4975 with standard deviation £143.
- Can we reject the hypothesis that the average weekly turnover is £5000?
- The test statistic is

$$z = \frac{4975 - 5000}{\sqrt{\frac{143^2}{100}}} = -1.75$$

6.2 Significance, effect size and power

- Since this is less than $-z* = -1.64$ the null is rejected with 95% confidence.
- True average weekly turnover is less than £5000. However, the difference is only £25 per week, which is 0.5% of £5000. Common sense would suggest that the difference may be unimportant, even if it is significant in the statistical sense.
- One should not interpret statistical results in terms of significance alone, therefore; one should also look at the size of the difference (sometimes known as the effect size) and ask whether it is important or not.
- This problem with hypothesis testing paradoxically grows worse as the sample size increases. For example, if 250 observations reveal average sales of 4985 with standard deviation 143, the null would (just) be rejected at 5% significance.
- In fact, given a large enough sample size we can virtually guarantee to reject the null hypothesis even before we have gathered the data. This can be seen from the equation for the z score test statistic: as n grows larger, the test statistic also inevitably increases.
- A good way to remember this point is to appreciate that it is the evidence which is significant, not the size of the effect. Strictly, it is better to say '. . . There is significant evidence of difference between . . .' than '. . . there is a significant difference between . . .'.

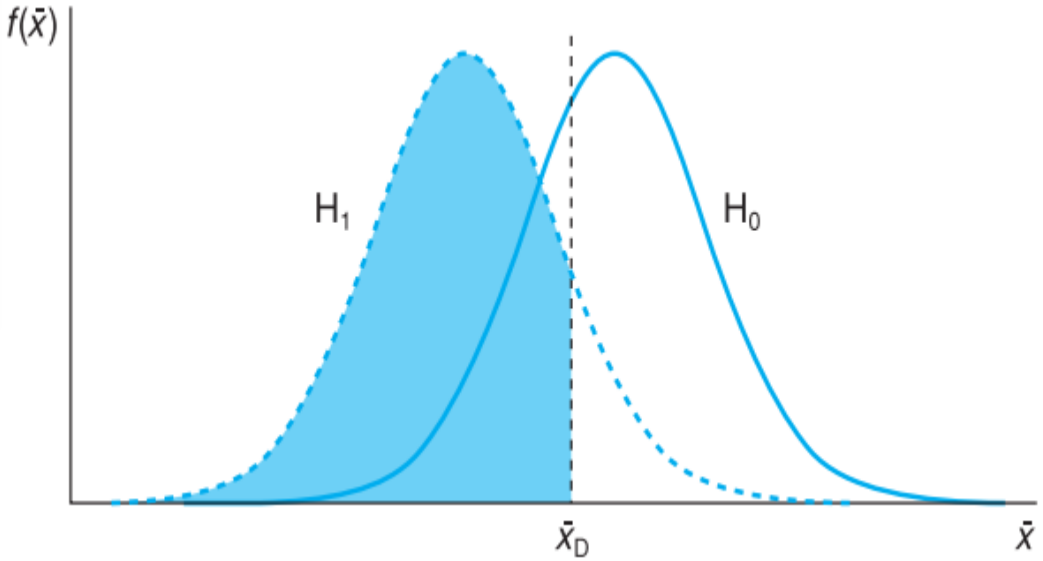
6.2 Significance, effect size and power

- A related way of considering the effect of increasing sample size is via the concept of the power of a test. This is defined as

$$Power\ of\ a\ test\ =\ 1\ -\ \Pr(Type\ II\ error)\ =\ 1\ -\ \beta$$

- where β is the symbol conventionally used to indicate the probability of a Type II error. As a Type II error is defined as not rejecting H_0 when false (equivalent to rejecting H_1 when true), power is the probability of rejecting H_0 when false (if H_0 is false, it must be either accepted or rejected; hence these probabilities sum to one).

The power of a test is therefore given by the area under the H_1 distribution, to the left of the decision line, as illustrated (shaded) in the next Figure (for a one-tail test).



6.2 Significance, effect size and power

It is generally desirable to maximise the power of a test, as long as the probability of a Type I error is not raised in the process. There are essentially three ways of doing this

● Avoid situations where the null and alternative hypotheses are very similar, i.e. the hypothesised means are not far apart (a small effect size).

● Use a large sample size. This reduces the sampling variance of X (under both H0 and H1) so the two distributions become more distinct.

● Use good sampling methods which have small sampling variances. This has a similar effect to increasing the sample size.

Exercise 1 If a researcher believes the cost of making a Type I error is much greater than the cost of a Type II error, should they choose a 5% or 1% significance level? Explain why.

Exercise 2

(a) A researcher uses Excel to analyse data and test a hypothesis. The program reports a test statistic of z = 1.77 (P value = 0.077). Would you reject the null hypothesis if carrying out (i) a one-tailed test (ii) a two-tailed test? Use the 5% significance level.

(b) Repeat part (a) using a 1% significance level.

6.3 Further hypothesis tests

- *Testing the difference of two means*
- Suppose a car company wishes to compare the performance of its two factories producing an identical model of car.
- The factories are equipped with the same machinery but their outputs might differ due to managerial ability, labour relations, etc. Senior management wishes to know if there is any difference between the two factories.
- Output is monitored for 30 days, chosen at random, with the following results:

	Factory 1	Factory 2
Average daily output	420	408
Standard deviation of daily output	25	20

- Does this produce sufficient evidence of a real difference between the factories, or does the difference between the samples simply reflect random differences such as minor breakdowns of machinery?

6.3 Further hypothesis tests

- The information at our disposal may be summarised as

$$\begin{array}{ll} \bar{x}_1 = 420 & \bar{x}_2 = 408 \\ s_1 = 25\ ; & s_2 = 20 \\ n_1 = 30 & n_2 = 30 \end{array}$$

- The hypothesis test to be conducted concerns the difference between the factories' outputs, so the appropriate random variable to examine is $\bar{x}_1 - \bar{x}_2$. From Chapter 4, this has the following distribution, in large samples

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

- The hypothesis test is therefore as follows.

- $H_0\colon\ \mu_1 - \mu_2 = 0\ ;\ H_1\colon \mu_1 - \mu_2 \neq 0$

- The null hypothesis posits no real difference between the factories. This is a two-tail test since there is no a priori reason to believe one factory is better than the other, apart from the sample evidence.

6.3 Further hypothesis tests

- Significance level: $\alpha = 1\%$. This is chosen since the management does not want to interfere unless it is really confident of some difference between the factories. In order to favour the null hypothesis, a lower significance level than the conventional 5% is set.
- The critical value of the test is $z^* = 2.57$. This cuts off 0.5% in each tail of the standard Normal distribution.
- The test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}}$$

- The hypothesised value of the difference (zero in this case) is subtracted from the sample difference and this is divided by the standard error of the random variable.
- **Decision rule:** $z < z^*$ so the test statistic falls into the non-rejection region.
- There does not appear to be a significant difference between the two factories.

6.3 Further hypothesis tests

- **Exercise1:** A survey of 50 teenage girls found that on average they spent 3.6 hours per week chatting with friends over the internet. The standard deviation was 1.2 hours. A similar survey of 90 teenage boys found an average of 3.9 hours, with standard deviation 2.1 hours. Test if there is any difference between boys' and girls' behaviour.
- **Exercise 2:** One gambler on horse racing won on 23 of his 75 bets. Another won on 34 out of 95. Is the second person a better judge of horses, or just luckier?

6.4 Hypothesis tests with small samples

- *Testing the sample mean*
- As with estimation, slightly different methods have to be employed when the sample size is small ($n < 25$) and the population variance is unknown. When both of these conditions are satisfied the t distribution must be used rather than the Normal, so a t test is conducted rather than a z test. This means consulting tables of the t distribution to obtain the critical value of a test, but otherwise the methods are similar.