MA181 INTRODUCTION TO STATISTICAL MODELLING
HYPOTHESIS TESTING

Suppose X is a random variable that follows a distribution dependent on a parameter θ and it is desired to test a hypothesis about θ on the basis of a random sample of observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of X.

Definitions The hypothesis under test is called the*null hypothesis* and is denoted by H_0. If the test leads us to reject H_0, then we accept an *alternative hypothesis* H_1. Consequently, H_0 and H_1 must, between them, cover all the possible values of the parameter.

Usually, H_0 is of the form $H_0 : \theta = \theta_0$ while H_1 takes one of the three forms $H_1 : \theta > \theta_0$, $H_1 : \theta < \theta_0$ and $H_1 : \theta \neq \theta_0$.

The problem of hypothesis testing reduces to that of dividing the sample space into two regions, that for which H_0 is rejected, called the *critical region*(C) (or *rejection region*), and that for which H_0 is accepted (or more properly not rejected), called the *acceptance region*.

When an experiment is carried out, there are two possible states of nature, H_0 true and H_0 false, and two possible decisions which might result, reject H_0 and accept H_0. The effects of this set up can be described by the following table.

	Accept H_0	Reject H_0
H_0 true	$\surd$	Type I error
H_0 false	Type II error	$\surd$

Let

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ true}) = P(\mathbf{X} \in C | H_0).$$

Then α is called the *size* of the test or the *significance level*.

Let

$$\beta = P(\text{Type II error}) = P(\text{Accept } H_0 | H_1 \text{ true}).$$

Then

$$1 - \beta = P(\text{Reject } H_0 | H_1) = P(\mathbf{X} \in C | H_1)$$

is called the *power* of the test, or the *power function* when regarded as a function of the parameter.

A good test would have small values of both α and β, i.e. a small size and a large power. It is usually impossible to achieve both of these simultaneously, so the standard procedure is to fix α at some acceptable level and then find the test that maximises $1-\beta$. The commonly used value of α is 0.05 but a value of 0.01 can be used for a more stringent test and a value of 0.001 for a very stringent test.

Normal distribution *One-tailed test*

Suppose $X \sim N(\mu, \sigma^2)$ where, for the moment, σ is assumed known and we wish to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu > \mu_0$. A random sample $X_1, X_2, \ldots, X_n$ is taken and yields observed values $x_1, x_2, \ldots, x_n$. It can be shown that the most powerful test of H_0 against H_1 leads us to reject H_0 is the sample mean $\overline{x}$ is large or, equivalently, if the critical region is defined by

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > c,$$

where z is the standardised form of $\overline{x}$ under H_0. The value of c is found by considering the fact that, under H_0,

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

Hence, if a test of size (significance level) α is desired, c is given by

$$\alpha = P(Z > c | \mu = \mu_0) = 1 - \Phi(c),$$

where $\Phi(.)$ is the cumulative distribution of the standard normal distribution. The commonly used values of α and c are given in the following table to sufficient accuracy for practical purposes:

α	c
0.05	1.645
0.01	2.326
0.001	3.090

We tend to state that, if $z > 1.645$, the test is significant to 5% and there is "some" evidence to reject H_0. If $z > 2.326$, there is "strong" evidence to reject and, if $z > 3.090$, there is "very strong" evidence to reject.

Suppose, however, that instead of the above, we wished to test $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$. Then the best critical region is defined by $z < -c$, where c is given by

$$\alpha = P(Z < -c|\mu = \mu_0) = \Phi(-c).$$

In view of the symmetry of the normal distribution, the values of c usually required are those given in the above table.

Example I.Scribe Inc. manufacture ball point pens and have done so for many years. The amount of ink in a pen is sufficient to draw a line the length of which is normally distributed with standard deviation 0.1 km. The firm tries to keep the ink-filling machine running so that the mean length of line is 4km, but there is a suspicion in the mind of the operator that the mean has fallen somewhat. Consequently, a random sample of eight pens are placed in a birometer and the lengths of line measured (in km) with following results:

$$3.81,\ 3.92,\ 3.94,\ 3.93,\ 3.72,\ 4.05,\ 3.88,\ 3.83.$$

It is desired that the test be carried out with $\alpha = 0.01$.

We are here testing $H_0 : \mu = 4$ against $H_1 : \mu < 4$ so that the critical region of a 1% test is $z < -2.326$. Since $\overline{x} = 31.08/8 = 3.885$, the observed value of z is

$$z = \frac{3.885 - 4}{0.1/\sqrt{8}} = -3.253.$$

The test is significant at 1% so that there is strong evidence to reject H_0. (In fact the test would also be significant at 0.1%.)

We can calculate the power of the test for any alternative value of μ. If, for instance, $\mu = 3.9$, then

$$1 - \beta \ = \ P\left(\frac{\overline{X} - 4}{0.1/\sqrt{8}} < -2.326 \middle| \mu = 3.9 \right)$$

$$
\begin{aligned}
&= P\left(\frac{\overline{X}-3.9}{0.1/\sqrt{8}} < -2.326 + \frac{0.1}{0.1/\sqrt{8}}\middle|\, \mu = 3.9\right) \\
&= P(Z < 0.502) = 0.692
\end{aligned}
$$

Two-tailed test

Assume again that $X \sim N(\mu\sigma^2)$ with σ assumed known but that we now wish to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$. The natural critical region to use is that consisting of both large and small values of $\overline{x}$ or, equivalently, of $z = \sqrt{n}(\overline{x} - \mu_0)/\sigma$. In view of the symmetry of the distribution of Z, the critical region takes the form $|z| > c$ where, for a test of size α, c is defined by

$$\alpha = P(|Z| > c|\mu = \mu_0) = P(Z > c|\mu = \mu_0) + P(Z < -c|\mu = \mu_0) = 2[1-\Phi(c)]$$

or $1 - \Phi(c) = \frac{\alpha}{2}$.

Such a test is called a *two-tailed test*, as opposed to the one-tailed tests described above, since the critical region is to be found in the two tails of the distribution of Z. The commonly used values of α and c are given in the following table:

α	c
0.05	1.960
0.01	2.576
0.001	3.291

Example Milk bottles are vacuum-formed from molten gobs of glass that are weighed as they fall into moulds. The weight is known to be normally distributed with standard deviation 2.5 g and it is important that the bottles mean weight is maintained st 2252; too low a mean results in too many fragile bottles, too high a mean in an excess consumption of glass as well as too many bottles having a low internal volume.

The hypotheses under test here are $H_0 : \mu = 255$ against $H_1 : \mu \neq 255$.

As the manufacturer does not want to stop production and check his equipment unless there is strong evidence that the mean has changed, he permits a probability of only 0.01 of making a Type I error.

The value of c is given by $1 - \Phi(c) = \frac{\alpha}{2} = 0.005$ so that $c = 2.576$ and the critical region is given by $|z| > 2/576$.

A random sample of eight bottles has a mean weight of 253.25g. So

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{253.25 - 255}{2.5/\sqrt{8}} = -1.980$$

and $|z|$ does not fall in the critical region. There is therefore no evidence to reject H_0.

(Note that the test would just be significant with $\alpha = 0.05$.)

Suppose the population mean weight of the bottles changes to 275g. Then the power of the test is

$$\begin{aligned}
1 - \beta &= P\left(\frac{\overline{X} - 255}{2.5/\sqrt{8}} > 2.576 \middle| \mu = 257\right) + P\left(\frac{\overline{X} - 255}{2.5/\sqrt{8}} < -2.576 \middle| \mu = 257\right) \\
&= P\left(Z > 2.576 - \frac{2}{2.5/\sqrt{8}}\right) + P\left(Z < -2.576 - \frac{2}{2.5/\sqrt{8}}\right) \\
&= 1 - \Phi(0.313) + \Phi(-4.839) \\
&= 0.377 + 0 = 0.377
\end{aligned}$$

Variance unknown

In practice, cases where the variance of X is known are uncommon. Usually σ^2 is not known. As a result, although

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

remains true, it cannot be used to test hypotheses about μ as it depends on σ. The natural way around this problem is to estimate σ^2 by

$$S^2 = \frac{\sum(X_i - \overline{X})^2}{n - 1}.$$

There are several reasons for setting the divisor equal to $n - 1$ (rather than, say, n), one of which is to ensure that $E(S^2) = \sigma^2$.

the use of S leads to a modified statistic defined by

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}.$$

Under $H_0, T \sim t_{n-1}$, i.e. T follows the (Student) t distribution with $n-1$ degrees of freedom. One-tailed and two-tailed tests can be carried out as before except that the critical region is given by $t > c, t < -c$ or $|t| > c$ as appropriate, where the critical point c must be read from the table of the t distribution rather than the standardised normal.

Example A particular task in a manufacturing industry has long been scheduled to take 15 minutes on average. The management wants to introduce a new way of doing the job that should prove more economical in the tools required and cleaner for the workers involved. They want, however, to keep the mean time unchanged so that the task still fits perfectly into the production line schedule. Consequently, they measure the times for 12 completions of the task under the new conditions and obtain the following results, in minutes.

13.6, 12.3, 16.3, 15.1, 13.8, 15.2, 14.5, 14.0, 13.3, 15.2, 16.1, 14.1.

We will assume that the time to complete the task is normally distributed and we want to test the null hypothesis $H_0 : \mu = 15$ against the alternative $H_1 : \mu \neq 15$. Since $n = 12$, the critical region of a test size $\alpha = 0.05$ is given by

$$|t| = \frac{|\overline{x} - 15|}{s/\sqrt{12}} > 2/201.$$

Now $\sum x_i = 173.5$ and $\sum x_i^2 = 2523.63$. Hence$\overline{x} = \frac{173.5}{12} = 14.458\dot{3}$ and $s^2 = \frac{(2523.63 - 173.5^2/12)}{11} = \frac{15.1092}{11} = 1.3736$. The observed value of $|T|$ is therefore

$$|t| = \frac{|13.458\dot{3} - 15|}{\sqrt{1.3736/12}} = 1.601.$$

This does not fall in the critical region so there is no evidence to reject H_0.