10 Hypothesis Testing

Very often and particularly in the natural, health and social sciences, data are collected in order to test a certain hypothesis. In most cases we express such a hypothesis in terms of some feature of the population(s) or probability distribution(s) being studied. Typical examples of hypotheses are:

- the mean change in blood pressure due to a new drug is zero;
- the probability of males buying a product recommended to them is the same as the corresponding probability for females;
- the mean change in blood pressure due to a new drug is not zero;
- the probability that an electronic component will fail before 100 hours of use is greater than 0.1.

Hypothesis testing is a way of deciding whether or not a set of data are consistent with a hypothesis about the underlying population.

10.1 Introduction

In hypothesis testing, we usually compare two hypotheses. The first, called the *null hypothesis* H_0, generally reflects the status-quo (e.g. the mean change in blood pressure due to a new drug is zero). The second, called the *alternative hypothesis* H_1, is the conclusion to be reached if we find evidence to reject H_0 (e.g. the mean change in blood pressure due to a new drug is not zero). The objective is then to determine whether the null hypothesis is plausible in light of a sample of data.

Let us consider again the example from Section 9.1. Suppose previous market research suggested that the proportion of internet users with more than one email account was $\theta = 0.6$. We want to collect some more up-to-date data in order to assess whether the value $\theta = 0.6$ is still plausible. In this case our null and alternative hypotheses might be $H_0 : \theta = 0.6$ versus $H_1 : \theta \neq 0.6$. Alternatively, if it was only really reasonable to suspect that θ was *at least* 0.6 (i.e. the proportion has not decreased since the original market research was done) then we might use the alternative $H_1 : \theta > 0.6$ instead.

For a given pair of hypotheses, we need a way of deciding whether a sample contains enough evidence to reject H_0 or not. The decision is made by considering the distribution of some statistic (which we will calculate from our sample) whose probability distribution is known under the assumption that the null hypothesis is correct. Such a statistic is called a *test statistic*. Given our sample of data, we can compute the observed value of this test statistic. We know its distribution if the null hypothesis is correct and this allows us to question whether the value we have observed would be likely in light of this theoretical distribution. If the observed value would be very unlikely (e.g. it's very far into a tail of the distribution) then we reject H_0 in favour of H_1. Otherwise we do not.

p-value	Interpretation
$p \geq 0.1$	No evidence against H_0: do not reject H_0.
$0.05 \leq p < 0.1$	Slight evidence against H_0, but not enough to reject it.
$0.01 \leq p < 0.05$	Moderate evidence against H_0: reject it and go with H_1.
$0.001 \leq p < 0.01$	Strong evidence against H_0: reject it and go with H_1.
$p < 0.001$	Very strong evidence against H_0: reject it and go with H_1.

Table 1: Rule-of-thumb for the interpretation of p-values.

10.2 Hypothesis tests

Philosophically there are two ways of deciding whether our observed test statistic is sufficiently unlikely under H_0 to cause us to reject it: *hypothesis testing* (due primarily to Karl Pearson, William Sealy Gosset and Ronald Fisher) and *significance testing* (due primarily to Jerzy Neyman and Egon Pearson, son of Karl). In practice, some hybrid of the two approaches is generally used. In this course we will not dwell on the differences and instead focus on a common hybrid approach which is especially compatible with the statistical software packages you are likely to come across. In this case we weigh up the evidence against H_0 and in favour of H_1 by computing a *p-value*. This is *the probability of observing a test statistic at least as extreme as our observed value under the null hypothesis*. A small p-value suggests that the observed value for the test statistic would be unlikely if H_0 was true, and if it is "small enough", we reject H_0 in favour of H_1. To help in deciding what is "small enough", Table 1 gives a rule-of-thumb for the interpretation of p-values.

10.2.1 Example: Testing a binomial proportion

In the internet users example, suppose that we have specified our null and alternative hypotheses as

$$H_0 : \theta = 0.6 \quad \text{vs.} \quad H_1 : \theta > 0.6.$$

Suppose further that we have taken a random sample of 50 users and noted that 38 of them have more than one email account. This gives us an observation $X = 38$ where we assume $X \sim \text{Bin}(50, \theta)$. Under the null hypothesis, $\theta = 0.6$ and so we can compute the p-value as

$$p = \Pr(X \geq 38 | \theta = 0.6) = 1 - \Pr(X \leq 37 | \theta = 0.6) = 0.0132,$$

where we focus on the region $X \geq 38$ because our alternative hypothesis here implies that "extreme" corresponds to large values of X. If we're satisfied that the binomial model for our data is reasonable and that θ cannot be less than 0.6, there are two explanations for our small p-value, $p = 0.0132$. Either the null hypothesis is true and we have observed a very unusual sample by chance. Or the null hypothesis is not true and it is because $\theta > 0.6$ that we have observed a large number of internet users with more than one account. The latter explanation seems more plausible and so, as suggested in Table 1, we conclude that we have moderate evidence against H_0 and reject it in favour of H_1. It appears that the proportion of internet users with more than one account is greater than 0.6.

10.2.2 Summary of steps

The testing procedure employed in the previous example is common to all hypothesis tests. The steps are as follows:

1. State the null and alternative hypotheses, H_0 and H_1;
2. Calculate the observed test statistic;
3. Compute the p-value of the test by comparing the observed value of the test statistic to its theoretical distribution under H_0;
4. Reach a conclusion about whether or not to reject H_0.

For most types of statistical test, after choosing your hypotheses (step 1), computer software is available to perform steps 2 and 3 for you. You can then use the p-value produced by the software to reach your conclusions (step 4).

10.2.3 Terminology

Before moving on to a very common type of hypothesis test, we will review some terminology which you may come across if you do any reading around the subject. In making a decision about whether or not to reject H_0 in favour of H_1, one of four things might happen. This is summarised in the following table:

		Actual situation	
		H_0 true	H_1 true
Decision	Reject H_0	Type I error	Correct decision
	Do not reject H_0	Correct decision	Type II error

So conceptually, there are two kinds of error we can make: a *type I error* in which we reject H_0 when it was true, and a *type II error* in which we fail to reject H_0 when H_1 was true. Much theory, for example, *sample size determination* is based around the idea of controlling the probabilities of making these two kinds of errors.

10.3 One sample t-test

The one sample t-test is one of the most commonly used hypothesis tests. The test assumes that we have a random sample from a normal distribution with unknown mean and variance. It is designed to help us assess the plausibility of a hypothesised value for the population mean, based on information in the sample.

Consider a single normal population with unknown mean μ and unknown variance σ^2. We draw a random sample $X_1, X_2, \ldots, X_n$ where each $X_i \sim N(\mu, \sigma^2)$ and estimate μ by the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We now want to test how convincing a proposal (say μ_0) is for the population mean based on the information in our sample. Our null hypothesis is therefore $H_0 : \mu = \mu_0$. For H_1 we might choose a *two-sided* alternative $H_1 : \mu \neq \mu_0$ or a *one-sided* alternative, either $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$.

Under H_0, $X_1, X_2, \ldots, X_n$ are IID $N(\mu_0, \sigma^2)$ and so we know from Section 9.3 that

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \sim t_{n-1}.$$

T is a standardised difference between the sample mean and our hypothesised mean μ_0 and so we use it as our test statistic for weighing up the evidence against H_0. Based on observations $x_1, x_2, \ldots, x_n$, we can compute the observed value of our test statistic

$$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}},$$

where $\bar{x}$ and s^2 are simply the observed sample mean and variance. The next step is to calculate the associated p-value. The way the p-value is calculated depends on whether the alternative hypothesis is one-sided or two-sided. If we have a two-sided alternative hypothesis, either small or large values of T could be expected under H_1 and so the p-value is given by

$$p = \Pr(T \geq |t| \text{ or } T \leq -|t| \;\mid\; \mu = \mu_0) = 2 \times \Pr(T \leq -|t| \;\mid\; \mu = \mu_0)\,,$$

where $T \mid \mu = \mu_0 \sim t_{n-1}$ and the second equality follows from the symmetry of the Student's t distribution about 0. Here $|.|$ denotes the absolute value, that is, $|t| = t$ if $t \geq 0$ or $|t| = -t$ is $t < 0$ (e.g. $|-3| = 3$ and $|5| = 5$). If we have chosen the one-sided alternative $H_1 : \mu < \mu_0$, then small values of T would be expected under H_1 and so the p-value is given by

$$p = \Pr(T \leq t \mid \mu = \mu_0)\,,$$

where $T \mid \mu = \mu_0 \sim t_{n-1}$. Conversely if $H_1 : \mu > \mu_0$, the p-value is given by

$$p = \Pr(T \geq t \mid \mu = \mu_0)\,.$$

Probabilities like these can be calculated in R using the `pt` function.

10.3.1 Example

An archaeologist has recorded the following measurements (in cm) of the total length of the skeletons of 10 adults from the extinct species *Homo heidelbergensis* which she found preserved in a recently excavated peat bog:

178.5 169.2 141.2 167.5 159.8 167.7 155.3 164.8 150.5 170.7

Assuming that these are independent observations of $X \sim N(\mu, \sigma^2)$, (a) test $H_0 : \mu = 156$ versus $H_1 : \mu \neq 156$; and (b) test $H_0 : \mu = 156$ versus $H_1 : \mu > 156$.

Solution. We have $\mu_0 = 156$ and calculate $n = 10$, $\bar{x} = 162.52$ and $s^2 = 120.25$. Our observed test statistic is therefore

$$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}} = \frac{162.52 - 156}{\sqrt{120.25/10}} = 1.8802$$

(a) The p-value for a two-sided test is

$$p = 2 \times \Pr(T \leq -|t| \mid \mu = \mu_0) = 2 \times \Pr(T \leq -1.8802 \mid \mu = 156)\,,$$

where $T \sim t_9$, which we can evaluate in R using

```
2*pt(-1.8802,9) # Gives 0.0928.
```

Therefore $p = 0.0928$ and so we conclude that there is little evidence against H_0 and so we cannot reject it. A mean skeleton length of 156 is broadly consistent with these data.

(b)

Two-sample t-test

Another common situation encountered in statistical inference is the desire to compare multiple groups to investigate if their population means differ. If we have two groups to compare, the most common method for carrying this out is the two-sample t-test. Suppose we have a sample A, which comes from a population with mean μ_A and variance σ_A^2; and a sample

B which comes from a population with mean μ_B and variance σ_B^2. A two-tailed two sample t-test would test the hypotheses

$$\mu_A = \mu_B,$$
$$\mu_A \neq \mu_B.$$

One-tailed alternatives are also possible and work the same as for the one-sample case.

10.4 Equal Variance Assumption

The exact form of test statistic we use actually depends on the two population variances σ_A^2 and σ_B^2, and whether we can assume they are equal, i.e. $\sigma_A^2 = \sigma_B^2$. In order to determine whether we can make this assumption, we carry out another hypothesis test, testing the hypotheses:

$$H_0 : \sigma_A^2 = \sigma_B^2,$$
$$H_1 : \sigma_A^2 \neq \sigma_B^2.$$

There are several hypothesis tests we can use to investigate the equal variance assumption. A commonly used test which comes in built to base R is Bartlett's test, which we carry out using the `bartlett.test` command. If our test gives a p-value less than 0.05, we reject the null hypothesis, and therefore cannot assume the population variances are equal.

10.4.1 Equal Variances

If we can assume this population variances are approximately equal (i.e. if the Bartlett test gives $p > 0.05$), we can perform the standard two sample t-test, where the test statistic is

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

where s_p is the pooled standard deviation, calculated by

$$s_p = \sqrt{\frac{(n_A - 1)\, s_A^2 + (n_B - 1)\, s_B^2}{n_A + n_B - 2}}$$

where $\bar{x}_A$ and $\bar{x}_B$ are the sample means for groups A and B, s_A^2 and s_B^2 are the sample standard deviations, and n_A and n_B are the sample sizes.

In reality we will almost never calculate this by hand, and instead allow R to do all the work for us, using the `t.test` command, and indicate we wish to use this form of the test by specifying `var.equal = TRUE`.

10.4.2 Unequal Variances

If we cannot assume the population variances are equal, we perform a slightly different version of the test, known as the Welch test. In this case the test statistic is

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

with the same notation as in the equal variance example. Again we can carry out this test in R using the `t.test` command, this time specifying `var.equal = FALSE`. By default R will assume we have unequal variances, so a Welch test will be performed unless we specify `var.equal = TRUE`.

Example

Suppose we have the marks of a group of students taking their first data science module, some from a statistics background and some from a computing background. We wish to know if there is a significant difference in the average score between the two groups of students.

We have the grades of 11 students from a statistics background:

$$\mathbf{x}_S = (64, 52, 74, 75, 76, 54, 77, 84, 57, 75, 56)$$

and 7 students from a computing background:

$$\mathbf{COMP} = (80, 63, 86, 63, 76, 58, 78)$$

We can then test whether there is a significant different between the mean score of the stats students μ_S and computing students μ_C.

$$\mu_S = \mu_C,$$
$$\mu_S \neq \mu_C$$

We then need to test whether the equal variance assumption is valid for this data. In order to carry out the Bartlett test in R, we first collect our data into a data frame with 2 columns corresponding to the score and the class it was from.

```
> head(marks)
  Mark     Group
1   80 computing
2   64     stats
3   63 computing
4   52     stats
5   86 computing
6   63 computing
```

Then we carry out the Bartlett test

```
> bartlett.test(Mark ~ Group, data = marks)

Bartlett test of homogeneity of variances

data:  Mark by Group
Bartlett's K-squared = 0.02754, df = 1, p-value =
0.8682
```

We see we have a p-value of 0.8682, which is greater than 0.05, and so the assumption of equal variances is valid for our t-test. We specify this in R by adding `var.equal = TRUE` to our `t.test` command.

```
> t.test(Mark ~ Group, data = marks, var.equal = T)

Two Sample t-test

data:  Mark by Group
t = 0.82036, df = 16, p-value = 0.4241
alternative hypothesis: true difference in means between group computing and group stats
95 percent confidence interval:
 -6.912523 15.639796
sample estimates:
mean in group computing     mean in group stats
               72.00000                67.63636
```

We see that we obtain a p-value of 0.4241, which is bigger than 0.05, so we cannot reject the null and conclude there is no evidence of a significant difference between the population mean marks in the stats and computing groups.

11 Paired t-test

A particularly special case of the two sample t-test is where we have paired data. Data are paired when the two samples being investigated are drawn from the same group of individuals, i.e. if we have samples $A = (x_{A1}, x_{A2}, \ldots, x_{An})$ and $B = (x_{B1}, x_{B2}, \ldots, x_{Bn})$ are formed such that observations x_{A1} and x_{B1} come from the same individual 1, observations x_{A2} and x_{B2} come from the same individual 2 etc, then samples A and B are paired. Examples of paired data might be measuring the sea temperature at particular locations 20 years apart; comparing the sales at a series of shops before and after deploying a new marketing campaign; or comparing the blood pressure of patients before and after they've taken a new treatment.

In a hypothesis test with paired data, we typically want to know if there is a significant difference between the two groups, however we cannot use a standard two-sample t-test since one of the test assumptions (that the two samples A and B are independent) is violated. Instead,

rather than carry out a hypothesis test on the two samples, we calculate the differences for each individual between the two samples, $d_1 = x_{A1} - x_{B1}, d_2 = x_{A2} - x_{B2}, \ldots, d_n = x_{An} - x_{Bn}$. We can then carry out a one-sample t-test on the differences:

$$H_0 : \mathrm{D} = \mu_0,$$
$$H_1 : \mathrm{D} \neq \mu_0$$

$$t = \frac{\bar{d} - \mu_0}{\frac{\sigma_d}{\sqrt{n}}}$$

where $\bar{d}$ is the sample mean of the differences $d_1, d_2, \ldots, d_n$ and s_d is the standard deviation of the differences. If we wish to simply test if there is a significant difference between samples A and B, we would take $\mu_0 = 0$ (i.e. the null hypothesis is that there is no significant difference between the samples)

Example

Revisiting our previous example, suppose we have the marks for our stats and computing students on their next module. We wish to see if the group as a whole perform significantly differently in their second module compared to their first.

```
> head(mod.comp)
  mark     group module
1   89     stats second
2   88 computing second
3   86 computing  first
4   84     stats  first
5   81     stats second
6   80 computing  first
```

We can therefore compare the mean performance in module 1, μ_1 against mean performance in module 2, μ_2.

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

Since we are comparing the scores of the same 18 students across the two modules, this is paired data, and so we should carry out a paired t-test. We can do this in R using the `paired = TRUE` argument.

```
> t.test(mark ~ module, paired = TRUE, data = mod.comp)

Paired t-test
```

```
data:  mark by module
t = -5.46, df = 17, p-value = 4.235e-05
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -2.541757 -1.124909
sample estimates:
mean difference
      -1.833333
```

We see our p-value is given as `4.235e-05`, which means $4.235 \times 10^{-5} = 0.00004235$. This is definitely much less than 0.05, and so we can say we have strong evidence against the null hypothesis and hence strong evidence to suggest there is a difference in mean mark between the two modules.

Issues with t-tests

While t-tests are the most widely used form of hypothesis test, there are some issues which we should be mindful of when working with them. The first is that the statistical theory underpinning them assumes that the sample means are Normally distributed. Recall from Chapter 6, that this can be assumed either when we have a large sample size (typically $n > 30$) so that the central limit theorem applies, or that the sample itself appears Normally distributed. When this is not the case (i.e. we have a small sample which does not appear to be Normally distributed) then t-tests will not be suitable hypothesis tests to carry out, and we should choose a non-parametric test instead, such as the Wilcoxon signed rank test (for one sample t-tests) or Mann-Whitney U test (for two sample t-tests). In the case of the two-sample t-test, we require this assumption to be valid for both samples, although in the paired t-test case it only needs to be true for the differences.

Another issue which affects hypothesis tests more generally, is the risk of generating false positives. Since we typically reject the null hypothesis when $p < 0.05$, we therefore have a 5% chance that we will reject the null hypothesis when we shouldn't. While this is generally deemed acceptable for single tests, the problem becomes magnified when we repeatedly carry out tests.

If we repeatedly carry out k hypothesis tests where the null hypothesis shouldn't be rejected, our probability of at least 1 false positive is

$$\begin{aligned}\Pr(\text{At Least 1 False Positive}) &= 1 - \Pr(\text{No False Positives}),\\ &= 1 - (0.95 \times 0.95 \times \cdots \times 0.95) \qquad = 1 - 0.95^k\end{aligned}$$

Hence we can see how the probability of a false positive increases with the number of tests we carry out. Carrying out $k = 5$ tests gives a probability of false positive probability of around 23%, $k = 10$ tests give a false positive probability of around 40%, and $k = 14$ tests make it more likely than not that we will reject the null hypothesis even when we shouldn't, with a probability of 51%. The message here therefore is that we should be wary of carrying

out lots of tests. Techniques such as analysis of variance (ANOVA) allow us to compare multiple groups simultaneously without the need for repeated testing. If repeated testing is unavoidable, then it is important to ensure the results of all tests are made known, and not just the ones producing a significant result (known as publication bias). This allows for the results to be put into context of the number of tests carried out, and avoids potentially misleading conclusions being reached.