Chapter 5

Estimation and hypothesis testing for continuous data

5.1 Objectives

1. To perform point and interval estimation and hypothesis tests for continuous data.
2. To use and interpret point and interval estimates and hypothesis tests relating to differences between the population means of two independent samples.
3. To use R to do all these things.

5.2 Point estimation

Example **Soil pH**

The pH level of the soil in a field was tested by taking samples from 17 scattered locations in the field. The pH readings were as follows:

```
> pH <- c(6, 5.7, 6.2, 6.3, 6.5, 6.4, 6.9, 6.6, 6.8, 6.7, 6.8,
+     7.1, 6.8, 7.1, 7.1, 7.5, 7)
```

What can we infer about the pH level of the soil in this field? We write a model with no explanatory variables

$$pH_i = \mu + e_i, \quad e_i \sim \mathrm{N}(0, \sigma)$$

For this model and many other models, the following procedure can be undertaken:

- Given *any* values of the parameter(s), we can compute *predicted values* (or fitted values) of the response by substituting the explanatory values into the deterministic part of the model equations.
- These predictions give estimated prediction errors (called **residuals**):

  $$\textit{residual} = \textit{estimated error} = \textit{observed response} - \textit{predicted response}$$

  For the soil pH data, the estimate of the parameter μ is the sample mean, 6.68. The residual for the first observation is $6 - 6.68 = -0.68$.
- Assuming that the standard deviation of the errors (σ) is constant, a modification of the usual formula for the sample standard deviation, which includes an adjustment for the number of parameters estimated, gives an acceptable estimate of σ. This quantity is often called the *residual standard deviation*:

$$s_{residual} = \hat{\sigma} = \sqrt{\frac{\text{sum of } (residuals)^2}{\text{no. study units } - \text{ no. parameters estimated}}}$$

The term `no. parameters estimated` does not include σ.

- The denominator [`no. study units − no. parameters estimated`] is called the **residual degrees of freedom**.
- The 'best' values of the estimated parameters are those which give, on average, the smallest estimates of the errors, i.e. the smallest $\hat{\sigma}$. This estimation procedure is called the **Method of Least Squares**.

In most cases the estimates are what we would expect. For population means they are simply the sample means.

5.2.1 Checking the model

We should check that the model adequately describes the data. Both the deterministic function and the random part need to be checked.

For many simple models, the deterministic function can be reasonably checked by looking at graphs of the data.

The random part describes the properties of the errors, which cannot be observed, but we can get estimates of them — the *residuals*:

$$\begin{array}{ccccc} \textit{residual} & = & \textit{observed response} & - & \textit{predicted response} \\ \hat{e}_i & = & y_i & - & \hat{y}_i \end{array}$$

If the model is correct, and the errors are assumed to be normally distributed with constant standard deviation, then the residuals should be consistent with a normal distribution: reasonably symmetric with no outliers, a normal probability plot that is not too different from linear and variation roughly equal along the line (for numerical explanatory variables) or within each group (for categorical explanatory variables), and so on.

The assumption of constant standard deviation (which is more important than that of normality) can be checked with a scatterplot of residuals versus the fitted values, i.e. a plot of $\hat{e}_i$ versus $\hat{y}_i$. If the assumption of constant standard deviation (also called "constant variance" or "homogeneity of variance") is reasonable, then no pattern should be evident in the plot. Departure from the assumption is most commonly revealed through a plot that shows 'fanning', where the magnitude of the errors increases with the fitted values.

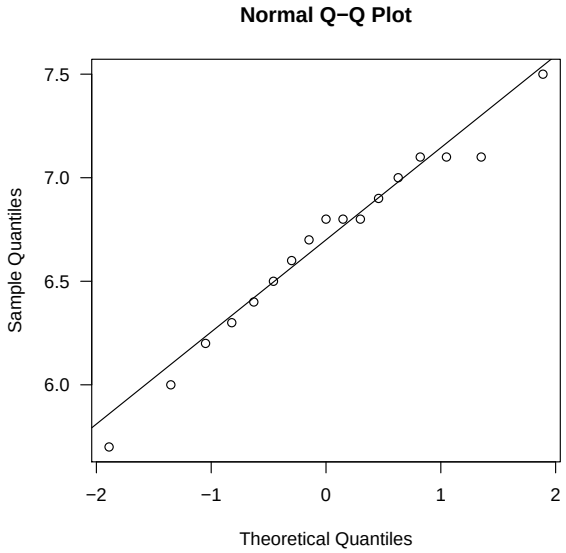
If there are numerical explanatory variables, graphs of the residuals versus these variables are also good. Any pattern indicates a departure from the model assumptions.

The normality of the errors can be checked with a normal probability plot of the residuals (also referred to as a normal Q–Q (Quantile–Quantile) plot.) In the QQ–plot of the pH residuals shown on the next page, all of the points lie quite close to a straight line, so the assumption of normality seems reasonable. In R:

```
> qqnorm(pH)
> qqline(pH)
```

The errors are also assumed to be **independent**, meaning that they are not related to each other but are like random draws from a normal distribution. Assessing this assumption usually requires knowledge of how the data were collected. Common causes of non-independence are data collected in time order, or several items taken from the same study unit, such as plants taken from the same pot. In these cases, some of the error is due to a cause that could be explained.

Lack of independence is a serious problem and one that can completely invalidate a statistical analysis if ignored. It is sometimes hard to deal with and is beyond the scope of this course.



5.3 Interval estimation of the mean

Continuing with the soil pH data:

```
> summary(pH)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.700   6.400   6.800   6.676   7.000   7.500

> sd(pH)

[1] 0.4548755
```

Model: Let the random variable X_i denote the pH of the ith soil sample, $i = 1, \ldots, 17$. Under the model we specified at the start of section 5.2, X_i has a normal probability distribution with mean μ and standard deviation σ, i.e. $X_i \sim \mathrm{N}(\mu, \sigma)$. The usual estimator of the sample mean $\bar{X}_n$ ($\bar{x}_n = 6.676$ is an estimate) is also a random variable, and its distribution is

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{17}}\right).$$

The standard error of this estimator is $sd(\bar{X}_n) = \dfrac{\sigma}{\sqrt{n}}$, which can be *estimated* by $\dfrac{s}{\sqrt{n}}$, the **standard error** of the (sample) mean;

A confidence interval for μ is *of the familiar form* $\bar{x} \pm k \times \dfrac{s}{\sqrt{n}}$, but we can't use the normal distribution quantiles because we don't know σ; we have to 'pay a price' for estimating σ, and this results in a slightly larger value for k. This requires us to use another table of values, but otherwise nothing changes.

From statistical theory, it is possible to work out the distributions involved – they are not quite normal, but involve another bell-shaped distribution called the *t distribution*.

If $X_1, \ldots, X_n$ is a random sample on $X \sim N(\mu, \sigma)$, then it turns out that

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

where S is an estimator of σ.

From which it follows that an observed 95% confidence interval for μ is

$$\bar{x} \pm t_{n-1}^{0.975} \times \frac{s}{\sqrt{n}}$$

where $t_{n-1}^{0.975}$ is the 0.975 quantile (97.5 percentile) of the t-distribution on $n-1$ df. Here are some properties of the t_ν distribution:

1. t_ν denotes the 't distribution with ν degrees of freedom'.
2. It is symmetrical about zero, bell-shaped, but more spread out than the normal distribution.
3. $t_\nu \xrightarrow{d} N(0,1)$ as $\nu \to \infty$. ($\xrightarrow{d}$ means "converges in distribution")

This distribution was derived by Gosset in 1905, *The Probable Error of a Mean*. Gosset worked for Guinness Breweries in Dublin, but wasn't allowed to publish under his own name, so he published under the name 'Student'. The t distribution is sometimes called 'Student's t distribution'.

An observed 95% confidence interval (CI) for μ is

$$\bar{x} \pm t_{n-1}^{0.975} \frac{s}{\sqrt{n}} = 6.676 \pm 2.120 \times \frac{0.455}{\sqrt{17}} = 6.676 \pm 2.120 \times 0.110 = (6.443,\ 6.910).$$

The value 2.120 is larger than the 0.975 quantile of the standard normal distribution, 1.96. Using R:

```
> n <- length(pH)
> mean(pH) + sd(pH) / sqrt(n) * qt(c(0.025, 0.975), df = n-1)

[1] 6.442595 6.910346
```

Such confidence intervals rest on certain assumptions about the data, the main ones being:

- The observations are a random sample from the population. This cannot always be assessed from the data; one must find out how the data were collected.
- If the sample size is small then the confidence interval is valid only if the data come from a normal distribution. This assumption can be checked using a Q-Q plot. For larger sample sizes the assumption is not so critical due to the Central Limit Theorem.
- It is inappropriate to use methods based on samples when the data refer to the whole population. If we have taken the whole population as a sample (a census), sample uncertainty no longer remains. For example, it doesn't make sense to construct a confidence interval for physical properties of the planets in our solar system, or for the mean income of employees of a company if every employee's income is known.

5.4 Hypothesis testing

Suppose that the soil in the field had a desired pH of 7 (i.e. neutral). Should we add some alkaline chemicals to change the pH of the soil? In other words, is the (mean) pH of the soil different from 7?

To answer this question, we first ask another: What is the probability of observing a sample mean at least as small as 6.676, if the population mean is 7? This is a one-sided P-value.

This probability is about 0.005 (we calculate this shortly), which is small. So either

1. the true mean is 7 and we have observed something that is very unusual, or
2. our assumption about the mean pH ($\mu = 7$) is wrong.

Here we would most likely conclude that $\mu < 7$ and add some chemicals.

5.4.1 Terminology and notation

(There is a degree of repetition with lab 4 for some of this section; but the reinforcement won't hurt!)

The null hypothesis, denoted by H_0, is usually a statement of 'no change' or 'no difference' (between a drug and a placebo, for example). It is expressed in terms of the parameters of the model, e.g. $H_0 : \mu = 7$.

The alternative hypothesis, denoted by H_1, often specifies what we hope, or expect, to be true. For the soil pH data, a one-sided alternative hypotheses would be $H_1 : \mu < 7$, which would require us to believe that the soil couldn't possibly be alkaline.

A two-sided alternative would be $H_1 : \mu \neq 7$, which would mean that we do not wish to specify

the direction in which the mean pH may differ from 7. If in doubt, use a two-sided H_1, as it is more conservative and, in most situations, more appropriate.

Statistical tests are performed using a **test statistic** which measures compatibility between the null hypothesis and the data. The alternative hypothesis determines which values of the test statistic count against H_0.

The result of a test is based on its ***P*-value** — the probability of observing a value of the test statistic as or more extreme than the one actually observed, assuming that H_0 is true.

For the soil pH example, the test statistic is the sample mean. The observed sample mean, $\bar{x}_n = 6.676$, has a one-sided P-value of 0.005. However, the two-sided alternative $H_1 : \mu \neq 7$ implies that values ≤ 6.676 or ≥ 7.324 are as or more extreme, and hence $P = 2 \times 0.005 = 0.010$.

In general, a small P-value is evidence against H_0. It highlights that the result has only a small probability of being explained by chance, and therefore is likely to indicate a real effect. But we still have to decide how small it should be. The following conventions have arisen in many areas of research (though there are some variations on these):

$P \leq 0.05$	evidence, significant (*)
$P \leq 0.01$	strong evidence, highly significant (**)
$P \leq 0.001$	very strong evidence, extremely significant (***)

5.4.2 Type I and type II errors

Hypothesis testing can also be thought of as a decision making process in which we do or don't 'Reject H_0'. There are then two kinds of errors we can make:

	Accept H_0	Reject H_0
H_0 true	Correct	Type I error
H_0 false	Type II error	Correct

A *Type I* error is rejecting H_0 when in fact H_0 is true. A *Type II* error is not rejecting H_0 when in fact H_0 is false (i.e. when H_1 is true).

Example

Declaring $\mu \neq 7$ when in fact $\mu = 7$ is a Type ___ error.
Not detecting that $\mu \neq 7$ when in fact $\mu \neq 7$, is a Type ___ error.

It is usually thought to be a worse error to reject H_0 when it is true, i.e. to go against the current norm or standard. It is therefore conventional to require the P-value to be small before rejecting H_0, and so reduce the probability of a type I error. However, this increases the probability of making a Type II error. It is always a trade-off.

There is a useful analogy between hypothesis testing and criminal trials, where it is usually assumed that it is worse to convict an innocent person than to not convict a guilty person.

H_0: accused is innocent.

We want proof 'beyond reasonable doubt' of guilt (if we want absolutely certain proof, we'll never convict.)

H_1: accused is guilty.

P-value = $P(evidence|H_0)$.

If the P-value is small, then we convict (reject H_0). The probability of obtaining the evidence that has been obtained will be small if the accused is innocent, therefore we conclude that (s)he is probably guilty — *beyond reasonable doubt*. Note that a large P-value does not mean innocence (H_0 true), it means not convicted (H_0 not rejected).

Notes

1. A result with $P \leq 0.05$ is, by convention, called 'statistically significant'. For some applications, though, we may make other requirements. For instance, if we are doing a pilot study, we may be willing to include any variables significant at the $P = 0.10$ level in our next study.

2. If we decide that we will reject H_0 if the P-value is ≤ 0.05, then $P(\text{Type I error}) = 0.05$. The largest value of P for which we will reject H_0 is referred to as the **level of significance** of the test.

3. When H_0 is *not* true, it is possible to make a Type II error – to accept H_0 when H_1 is true. The probability that we correctly reject H_0 when H_1 is true $(1 - \text{P(Type II error)})$ is called the **power** of the test. It is obviously desirable for the power to be large.

4. Statistical significance is not the same thing as practical significance. If we took enough soil samples we could detect a mean of 6.999 as being statistically significantly different from 7. But such a departure from $\mu = 7$ is likely to be of no practical importance. The P-value tells how likely it is that the data could have been obtained if H_0 is true, *not* how large or important any departure from H_0 is.

5. A large P-value, giving a non-significant result, does *not* mean that H_0 is true. If we took only two or three soil samples, we might not detect departure from $\mu = 7$ even if μ were actually as small as 6 (i.e., we may get a large P-value). A large P-value only means that we don't have enough evidence to reject H_0. **Tests based on a small number of samples (observations) tend to have little power against reasonable alternatives.**

5.4.3 Testing hypotheses about μ

To test the hypothesis $\mu = \mu_0$ we need to calculate a P-value, which can be obtained using the result that *if H_0 is true, then* $\dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$.

Some percentiles of the t distribution can be obtained from tables, but more precise P-values can usually be obtained from a statistics package, such as R.

Example Soil pH (continued).

Carry out a test of $H_0: \mu = 7$ versus $H_1: \mu \neq 7$.

The observed value of the test statistic is $\dfrac{\bar{x}_n - \mu_0}{s/\sqrt{n}} = \dfrac{6.676 - 7}{0.455/\sqrt{17}} = -2.936$. The P-value is found using a one-sample t-test, which is performed in R by:

```
> t.test(pH, mu = 7)

        One Sample t-test

data:  pH
t = -2.9326, df = 16, p-value = 0.009758
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 6.442595 6.910346
sample estimates:
mean of x
  6.67647
```

To carry out a one-sided test of $H_0: \mu = 7$ vs $H_1: \mu < 7$ we do the following:

```
> t.test(pH, mu = 7, alternative = "less")

        One Sample t-test

data:  pH
t = -2.9326, df = 16, p-value = 0.004879
alternative hypothesis: true mean is less than 7
95 percent confidence interval:
     -Inf 6.869083
sample estimates:
mean of x
  6.67647
```

5.4.4 Hypothesis tests and confidence intervals

There is a very close relation between hypothesis testing and confidence intervals. When testing a hypothesis about a single parameter (with a two-sided alternative), the test will reject H_0 at the level α (α is usually 0.05) whenever the null value being tested is outside a $100(1-\alpha)\%$ confidence interval. A value outside the interval is not consistent with the data, and the test backs this up.

Example An observed 95% CI for μ in the soil pH example is $6.676 \pm t_{16}^{0.975} \times 0.455/\sqrt{17}$, or (6.443, 6.910).

The null value $\mu = 7$ is not in this interval so the test will reject H_0 (at the 5% level), and accept H_1: $\mu \neq 7$. Similarly, a test of $H_0 : \mu = 6.5$ would not be significant (at the 5% level), because 6.5 is in the 95% confidence interval.

5.5 Inference for two populations

5.5.1 Independent samples

Example Weight gain of chicks

A small experiment compared feeds 1 and 2, for which the weight gains were:

feed	weight gain
1	42 68 85
2	42 97 81 95 61 103

```
> chicks <- data.frame(feed = rep(c(1,2), times=c(3,6)),
+                      weight_gain = c(
+                        42, 68, 85,
+                        42, 97, 81, 95, 61, 103))
```

To test equality of the means for the two feeds or to find a confidence interval for the difference between the two means we can assume either equal variances for the two populations ($\sigma_1 = \sigma_2$) or unequal variances ($\sigma_1 \neq \sigma_2$).

Assuming equal variances: $\sigma_1^2 = \sigma_2^2 \ (= \sigma^2)$

Let $X_1, \ldots, X_{n_1}$ denote the weight gain of chicks on feed 1 and $Y_1, \ldots, Y_{n_2}$ the weight gain of chicks on feed 2. If we assume that $X_i \sim \mathrm{N}(\mu_1, \sigma)$, for $i = 1, \ldots, n_1$, and $Y_j \sim \mathrm{N}(\mu_2, \sigma)$, for $j = 1, \ldots, n_2$, it then follows that

$$\bar{X}_{n_1} - \bar{Y}_{n_2} \sim \mathrm{N}\left(\mu_1 - \mu_2,\ \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}\right)$$

where $\bar{X}_{n_1}$ and $\bar{Y}_{n_2}$ are the means of the random samples. This simplifies to

$$\bar{X}_{n_1} - \bar{Y}_{n_2} \sim \mathrm{N}\left(\mu_1 - \mu_2,\ \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

Given data from the two groups we estimate $\mu_1 - \mu_2$ by $\bar{x}_{n_1} - \bar{y}_{n_2}$. We also get *two* estimates of σ (s_1 and s_2) and the question is how should we combine these estimates into a single estimate. It turns out to be desirable statistically to use

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

which is an estimate of σ on $(n_1 + n_2 - 2)$ df, the so-called **pooled estimate of σ**, where s_1^2 and s_2^2 are weighted by weights related to the size of the sample they came from.

An observed 95% CI for $\mu_1 - \mu_2$ is then given by

$$\bar{x}_{n_1} - \bar{y}_{n_2} \pm t_{(n_1+n_2-2)}^{0.975} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the test statistic

$$t = \frac{\bar{x}_{n_1} - \bar{y}_{n_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is used for testing $H_0 : \mu_1 = \mu_2$, which has a t distribution with $(n_1 + n_2 - 2)$ df, when H_0 is true.

Example Growth of chicks (continued)

```
> tapply(chicks$weight_gain, chicks$feed, mean)

       1        2
65.00000 79.83333

> tapply(chicks$weight_gain, chicks$feed, sd)

       1        2
21.65641 23.86979
```

The pooled estimate of σ is $s = \sqrt{\frac{2 \times 21.66^2 + 5 \times 23.87^2}{3 + 6 - 2}} = \sqrt{\frac{3787.2}{7}} = \sqrt{541.0} = 23.26$.

Hence an observed 95% CI for $\mu_1 - \mu_2$ is

$$\begin{aligned} 65.00 - 79.83 \pm t_7^{0.975} \times 23.26\sqrt{\frac{1}{6} + \frac{1}{3}} &= -14.83 \pm 2.365 \times 16.447 \\ &= -14.83 \pm 38.90 = (-53.72, 24.06). \end{aligned}$$

For a test of $H_0 : \mu_1 = \mu_2$,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{-14.83}{16.447} = -0.902$$

which can be compared with a t distribution on 7 df. The 0.025 and 0.975 quantiles of t_7 are -2.36 and 2.36, so the null hypothesis is not rejected. All of this method can be implemented in R:

```
> t.test(weight_gain ~ feed, data = chicks, var.equal = TRUE)

        Two Sample t-test

data:  weight_gain by feed
t = -0.9019, df = 7, p-value = 0.3971
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -53.72318  24.05651
sample estimates:
mean in group 1 mean in group 2
       65.00000        79.83333
```

Not assuming equal variances: $\sigma_1 \neq \sigma_2$

When comparing the means of two normal populations, it is possible to drop the assumption that $\sigma_1 = \sigma_2$. The resultant confidence intervals and tests are still based on the t distribution, but now as an acceptable approximation rather than as an exact result.

An observed 95% CI for $\mu_1 - \mu_2$ is given by

$$\bar{x}_{n_1} - \bar{y}_{n_2} \pm t_\nu^{0.975}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

while a test statistic for testing $H_0 : \mu_1 = \mu_2$ is given by

$$t = \frac{\bar{x}_{n_1} - \bar{y}_{n_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

which has (approximately) a t distribution with ν df, when H_0 is true. The best approximations are obtained using a value for the degrees of freedom (ν) given by the following messy function of s_1, s_2, n_1 and n_2:

$$\nu = \frac{(V_1+V_2)^2}{\frac{V_1^2}{n_1-1}+\frac{V_2^2}{n_2-1}}, \quad \text{where } V_1 = \frac{s_1^2}{n_1}, \quad V_2 = \frac{s_2^2}{n_2}.$$

This is what is used in R as the default setting. A much simpler, conservative approximation is obtained using $\nu = \min(n_1 - 1, n_2 - 1)$.

Example Growth of chicks (continued)

Using the better approximation above gives $V_1 = \frac{(21.66)^2}{3} = 156.39$; $V_2 = \frac{(23.870)^2}{6} = 94.96$.

$\nu = \frac{(156.33+94.96)^2}{(156.33)^2/2+(94.96)^2/5} = 4.50$. An observed 95% CI for $\mu_1 - \mu_2$ is:

$$\begin{aligned} 65.00 - 79.63 &\pm t_{4.50}^{0.975}\sqrt{\frac{23.87^2}{6}+\frac{21.66^2}{3}} \\ = -14.83 &\pm 2.659 \times 15.854 \\ = -14.83 &\pm 42.16 = (-57.0,\ 27.3) \end{aligned}$$

and a test of $H_0: \mu_1 = \mu_2$,

$$t = \frac{65.00 - 79.83}{\sqrt{\frac{21.66^2}{3}+\frac{23.87^2}{6}}} = \frac{-14.83}{15.854} = -0.935$$

which is not significant. In R:

```
> t.test(weight_gain ~ feed, data = chicks)

        Welch Two Sample t-test

data:  weight_gain by feed
t = -0.9357, df = 4.503, p-value = 0.3968
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -56.97338  27.30671
sample estimates:
mean in group 1 mean in group 2
       65.00000        79.83333
```

Note that R uses fractional degrees of freedom for the test.

5.5.2 Paired samples

Example Weight gain of chicks — modified experiment.

Suppose that we are again wanting to compare feeds 1 and 2, but this time we have six sets of twins. The most appropriate design would then be to randomly assign one chick from each set of twins to feed 1, and the other chick to feed 2. This is because we expect twins to be similar in many unmeasured characteristics and we can compare the feeds more accurately because we have accounted for a major source of variation.

The following data were obtained, where each row refers to a different set of twins. Find a 95% CI for the difference in mean weight gains of the two feeds, and carry out a test of whether the means are equal.

Feed 1	Feed 2
44	42
55	61
68	81
85	95
90	97
97	103

Solution:
Here there are really two categorical explanatory variables, the (6) sets of twins and the (2) feeds. However, if we take differences between the weight gains within each set of twins we take account of both of the explanatory variables, and end up with a set of 6 observations on which we can apply the same methods as we did with the soil pH example.

The six differences are: -2, 6, 13, 10, 7, and 6, for which $\bar{x} = 6.67$ and $s = 5.05$. Hence an observed 95% CI for the mean of the differences is

$$6.67 \pm t_5^{0.975} \times \frac{5.05}{\sqrt{6}} = 6.67 \pm 2.571 \times 2.062 = 6.67 \pm 5.30 = (1.37,\ 11.98)$$

and the test statistic is

$$t = \frac{6.67}{5.05/\sqrt{6}} = \frac{6.67}{2.062} = 3.235$$

which is larger than $t_5^{0.975} = 2.571$, and so the null hypothesis of equal means is rejected. In R:

```
> Feed.1 <- c(44, 55, 68, 85, 90, 97)
> Feed.2 <- c(42, 61, 81, 95, 97, 103)
> diff <- Feed.2 - Feed.1
> t.test(diff)

        One Sample t-test

data:  diff
t = 3.2359, df = 5, p-value = 0.02305
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  1.370741 11.962592
sample estimates:
mean of x
 6.666667
```

If the data had been analysed without taking the pairing in to account, the variability would have been estimated to be considerably larger, resulting in a less significant result. Let's try it in R:

```
> Feed <- c(Feed.1, Feed.2)
> group <- c(rep(1, 6), rep(2, 6))
> t.test(Feed ~ group)

        Welch Two Sample t-test

data:  Feed by group
t = -0.514, df = 9.837, p-value = 0.6186
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -35.63370  22.30037
sample estimates:
mean in group 1 mean in group 2
       73.16667        79.83333
```

The mean difference is of course the same, but the confidence interval is much wider and the P-value is much larger. It is certainly worth including the pairing in the analysis.

5.6 Exercises

1. A randomly selected sample of $n = 12$ students at a university was asked "how much did you spend on textbooks this semester?" The responses, in dollars, were

   200 175 450 300 350 250 150 200 320 370 404 250

   (a) Draw a boxplot of the data. Briefly discuss whether the assumptions for doing a confidence interval for the population mean (see page 4) appear to be satisfied.

   (b) Calculate the mean, the standard error of the mean and use these to find a 95% confidence interval for the population mean. (Be sure to use the appropriate quantiles.)

   (c) Write a sentence interpreting the confidence interval.

2. Suppose we spin a coin on its edge, and record whether it falls heads or tails. For some coins, depending on how the faces are weighted, this is not a fair game. For one coin, we record 68 heads in 100 spins. To determine if it is a fair game, we carry out a hypothesis test. Set up the model and hypotheses, and without doing any computations say if you expect a 'small' or 'large' P-value. Check it by calculating the P-value in R.

3. Environmental assessment: has sewage dumping caused the dissolved oxygen in a lake to be so low that fish are endangered?

   Measurements taken at eight locations gave the following values in parts per million:

   5.1 4.9 5.6 4.2 4.8 4.5 5.3 5.2

   (a) Environmental scientists claim that a mean value less than 5.0 parts per million is likely to lead to fish death.

   Rewrite the claim as the test of a hypothesis, clearly stating the null and alternative hypotheses. Carry out the test and state your conclusions.

   (b) Industrial scientists claim that a mean value greater than 5.0 parts per million is safe for fish.

   Rewrite the question above as the test of a hypothesis, clearly stating the null and alternative hypotheses, carry out the test and state your conclusions.

   (c) Contrast the two results.

4. Identify the null and alternative hypotheses in each of the following situations using *descriptive* statements e.g. H_0: new drug is not better, H_1: new drug is better.

   (a) A construction engineer wishes to determine if a new cement mix has a better bonding quality than the mix currently in use. The new mix is more expensive, so the engineer would not recommend it unless the experimental evidence suggests that it does have a better bonding quality. The bonding quality is to be observed from several cement slabs prepared with the new mix.

   (b) A State Labour Department wishes to determine whether the rate of unemployment in the State varies significantly from the forecast of 10% made two months ago. The Bureau of Statistics will obtain some data from a sample of 1000 randomly selected individuals.

   (c) During a flu epidemic, 20% of Melbourne's population suffer from flu attacks. A doctor theorizes that regular users of vitamin C are less susceptible to a flu attack. She intends to sample 500 regular users of vitamin C and determine how many suffered from the flu.

   (d) When a certain machine is in adjustment it produces, on average, 1% defective items. The quality control engineer wants to determine if the machine is in adjustment.

   (e) An agronomist believes that plants grown from a new strain of seed are likely to be more resistant to a disease than an existing variety. She plans to expose both types of plants to the disease, count the number of incidences of the disease, and use the data to establish her conjecture.

5. From Hand, et.al., in *Acta Opthalmologica*, 'On corneal thickness and interocular pressure II': data were given for eight people who each had one eye affected with glaucoma and one not affected. Corneal thickness (in microns) of both eyes was measured. Is there any difference?

Affected:	488	478	480	426	440	410	458	460
Not affected:	484	478	492	444	436	398	464	476

6. The following (sorted) data are the recorded speeds (in kilometres per hour) of a random sample of 20 cars obtained for one street as part of a study to evaluate the effectiveness of reducing the speed limit in local streets from 60 to 50 kilometres per hour.

```
46, 43, 46, 53, 50, 57, 45, 57, 58, 53, 46, 46, 48, 49, 53, 49, 53, 37, 42, 38
```

   For (b) and (c) you may assume that the distribution of speed is normal.

   (a) Produce a "five-number summary" of the data using `quantile()`, and confirm the summary by creating a boxplot of the data.

   (b) The local council wishes to have a 95% confidence interval for the mean speed of all cars in that street. Calculate the 95% confidence interval.

   (c) Find the value of the test statistic for a test of the (null) hypothesis that mean speed (μ) = 50 versus the two-sided alternative ($\mu \neq 50$).

   Give a P-value for this test, and state your conclusion. How useful is this test?

   (d) It is discovered that the first 10 observations are for cars travelling east and the last 10 for cars travelling west. Test the hypothesis that the average speed is the same for the two directions, without assuming equal variances for the two groups. Perform the test again, assuming equal variances. Calculate the variance for each group to examine this assumption.

   (e) The council would also like to have a 95% confidence interval for the proportion of cars (p) that exceed the 50 kilometre per hour limit on this street.

      i. Find $\hat{p}$, the proportion of cars in the sample whose speed was greater than 50.

      ii. Find a 95% confidence interval for p.

7. A psychologist performed an experiment to see if sleep affects the ability to recall information:

   - 40 subjects were randomly allocated to two groups. There were 25 in group A and 15 in group B.
   - Group A were shown a wildlife documentary at 7 am, had their normal day (with no sleep) and then were given a 50 question multiple choice test at 7 pm, on details from the documentary.
   - Group B were shown the same wildlife documentary at 7 pm, had their normal night (with sleep) and then were given the test at 7 am.

   Here is a summary of the number of questions correctly answered by individuals in each group.

Group	Sleep	N	Mean	StDev
A	No	25	35.6	1.32
B	Yes	15	37.2	1.84

   (a) State appropriate null and alternative hypotheses.

   (b) You may assume that it is valid to use a t test with a pooled estimate of s. Show that the pooled estimate of σ is $s = 1.532$

   (c) Calculate an appropriate t statistic (show working). Find the P-value using the R function `pt(q, df)` where `q` is the t statistic. State a conclusion in the context of this study.