Dr Isabella Gollini

Probability and Statistics

10 The testing of hypotheses and t-tests

10.1 A conceptual framework for hypothesis testing

Suppose that we are considering a set of data $x_1, x_2, \ldots, x_n$. We may go through a sequence of steps that characterize a test procedure.

1. We begin by setting up a *statistical model*, which provides a theoretical framework for analysing the data. For example, we may assume that $x_1, x_2, \ldots, x_n$ is a random sample of size n from a $N(\mu, \sigma^2)$ distribution, where the mean μ and the variance σ^2 are unknown.

2. We shall usually consider two competing hypotheses, a *null hypothesis* H_0 and an *alternative hypothesis* H_1. A hypothesis is typically an assertion about the true value of a parameter or parameters. For example, given some particular value μ_0, we might have
$$H_0 : \mu = \mu_0$$
and
$$H_1 : \mu \neq \mu_0.$$
The null hypothesis H_0 might correspond to what some theory predicts or might represent some set of circumstances which, in the absence of evidence to the contrary, we wish to assume holds — it is the default option. The alternative hypothesis H_1 represents the family of possible departures from the null hypothesis that we wish to envisage.

   In the example above, we have specified a *two-sided alternative hypothesis* $H_1 : \mu \neq \mu_0$. We could specify a *one-sided alternative hypothesis* $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$. We shall come across examples where a one-sided alternative hypothesis is appropriate. The form of the alternative hypothesis will influence the test procedure.

3. We construct a test statistic $T(x_1, x_2, \ldots, x_n)$, a function of the observations, whose distribution under H_0 is known, at least approximately. In a way that we shall need to specify more precisely, we shall reject the null hypothesis H_0 if $T(x_1, x_2, \ldots, x_n)$ is large enough.

   In the example of Section 9.3 we had
$$T(x_1, x_2, \ldots, x_n) = \left| \frac{(\bar{x} - \mu_0)}{\frac{s}{\sqrt{n}}} \right| .$$

4. The statistic T usually has a continuous distribution, so that, given any α with $0 < \alpha < 1$, we can find a constant k_α such that
$$\Pr(T(x_1, x_2, \ldots, x_n) \geq k_\alpha | H_0) = \alpha.$$

Values of k_α can usually be found from within a statistical package or from tables of percentage points of the distribution of T (or of a related statistic).

5. Given α, if t is the observed value of the test statistic on a particular occasion, we *reject* H_0 *at the* $100\alpha\%$ *significance level* if and only if $t \geq k_\alpha$. Conventionally, we use $\alpha = 0.05, 0.01$ or 0.001, i.e., tests at the 5%,1% or 0.1% significance level.

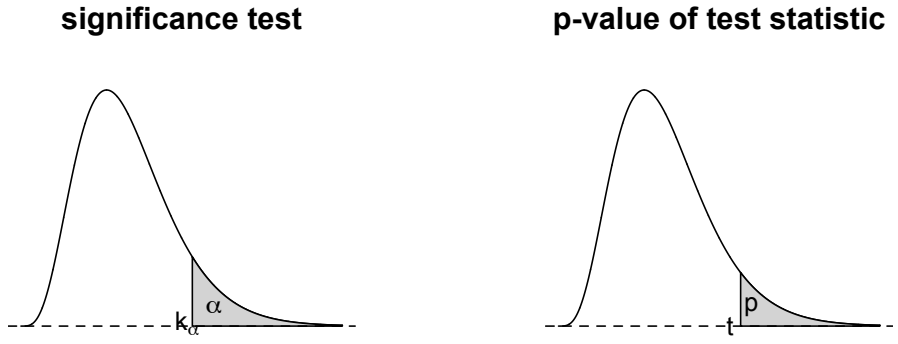

Figure 1: Illustration for tests of significance and p-values of test statistics

6. An alternative way of describing this procedure is in terms of p-values. The probability under H_0 of obtaining a value of the test statistic T that is greater than or equal to the observed value t,

$$p \equiv \Pr(T \geq t | H_0),$$

is the *p-value* (or *significance level*) of the observed value t.

The p-value is the value of α such that $t = k_\alpha$, i.e., the smallest value of α such that H_0 is rejected at the $100\alpha\%$ significance level. So we can reject H_0 at the $100\alpha\%$ significance level if and only if $p \leq \alpha$.

The p-value may be regarded as a measure of the weight of evidence against H_0. The larger the observed value of T and the smaller the corresponding p-value, the greater the evidence against H_0. In simple terms, if the p-value is small then either H_0 is true and an extreme outcome of the sampling experiment has occurred or H_0 is false.

There is never any perfect test procedure available, guaranteed always to come up with the right conclusion. There are always two types of error associated with any test, the error of rejecting H_0 when it is true (a *Type I error*) and the error of not rejecting H_0 when H_1 is true (a *Type II error*), as illustrated in Table 1.

Our main focus will be on on the Type I error, but we should at least bear in mind that there is also a Type II error. If we carry out a test procedure at the $100\alpha\%$ significance level then α is the probability of a Type I error, the probability of rejecting H_0 when it is true.

True hypothesis	outcome of test procedure: do not reject H_0	outcome of test procedure: reject H_0
H_0	desired result	Type I error
H_1	Type II error	desired result

Table 1: Hypothesis and outcome of test

- If the null hypothesis H_0 is true, and we carry out a large number of sampling experiments and tests of H_0 at the $100\alpha\%$ significance level then in the long run we shall reject H_0 on a proportion α of occasions — even though H_0 is true.
- Usually the alternative hypothesis H_1 is what is known as a *composite* one, made up of a number of possible parameter values, e.g., $\{(\mu, \sigma^2) : \mu \neq \mu_0\}$. The Type II error probability will not take a single value but will be a function of (μ, σ^2).

In general we would like to have both types of error probability small, but in practice some compromise is needed. The smaller we make the significance level α, i.e., the Type I error probability, the larger are the Type II error probabilities.

We also refer to the *power* of a test, the probability of rejecting H_0 when H_1 is true, which we would like to be as large as possible — as close to 1 as possible. The power is just one minus the Type II error probability. The smaller we make the significance level α, the Type I error probability, the smaller also is the power.

To some extent, the analogy with an English court of law may be helpful. The null hypothesis H_0 is the presumption of innocence of the accused; the alternative hypothesis H_1 is that the accused is guilty. There is an asymmetry between the role of the two hypotheses. The verdict is "guilty" if there is evidence evidence "beyond reasonable doubt" that the accused is guilty. Otherwise, the verdict is "not guilty." But errors do occur.

10.2 One sample t-tests

We saw in Section 9.3 an example of what is known as a *one sample t-test* or *single sample t-test* , or just as a *t-test* . We now consider another example.

Example

Twelve plants of a certain variety, grown under uniform conditions and treated with a new brand of fertilizer, attain the following heights (in cm).

25, 28, 24, 23, 27, 30, 24, 21, 28, 30, 26, 27

From extensive previous experimentation, it is known that plants of the same variety, grown in similar conditions but treated with a standard fertilizer, attain a mean height of 24.9 cm.

Given the observed data, is there sufficient evidence to conclude that the new fertilizer produces plants with a greater mean height than the standard fertilizer?

We set up the statistical model that the data are a random sample from a $N(\mu, \sigma^2)$ distribution, where μ and σ^2 are unknown. We test the null hypothesis $H_0 : \mu = 24.9$ against the one-sided alternative hypothesis $H_1 : \mu > 24.9$.

- We adopt this one-sided alternative hypothesis because we are only envisaging the possibility the new fertilizer has led to an increase in the mean height of plants.

To test the null hypothesis H_0 we use the test statistic

$$t = \frac{(\bar{x} - 24.9)}{\frac{s}{\sqrt{n}}},$$

which under H_0 has the t_{n-1} distribution. We shall reject H_0 if $\bar{x}$ is large enough or, equivalently, if t is large enough.

In the present case, $n = 12$, $\bar{x} = 26.083$ and $s = 2.778$:

```
height <- c(25, 28, 24, 23, 27, 30, 24, 21, 28, 30, 26, 27)

n <- length(height)
n

## [1] 12

xbar <- mean(height)
xbar

## [1] 26.08333

s <- sd(height)
s

## [1] 2.778434
```

The calculated value of the test statistic, which under H_0 has the t_{11} distribution, is

$$t = \frac{(26.083 - 24.9)}{\frac{2.778}{\sqrt{12}}} = 1.475.$$

If F denotes the distribution function of the t_{11} distribution then in this case, using Table 9 of *Lindley and Scott*, with some interpolation, the p-value is given by

$$p = \Pr(t \geq 1.475) = 1 - F(1.475) = 1 - 0.916 = 0.084 \quad \text{(to 3 decimal places)}.$$

We do not reject the null hypothesis at the 5% significance level, i.e., there is no strong evidence that the new brand of fertilizer produces plants with a greater mean height.

Alternatively we could compare our calculated t-value with the percentage points of the t-distribution with 11 degrees of freedom. From Table 10, $t_{11}(5) = 1.796$. Since our calculated value 1.475 is smaller than this percentage point, we do not reject the null hypothesis at the 5% significance level.

We are using a one-sided alternative hypothesis and, correspondingly, what is known as a *one-tail test*, where the p-value $p = \Pr(t \geq 1.475)$. For the calculated t-value to be significant at the 5% level it would have to exceed the value of the percentage point $t_{11}(5)$.

In the example of Section 9.3 we were using a two-sided alternative hypothesis and, correspondingly, what is known as a *two-tail test*.

In the present example, if we had instead used the two-sided alternative hypothesis $H_1 : \mu \neq 24.9$ then we would have used a two-tail test, where the p-value is given by $p = \Pr(|t| \geq 1.475) = 2\Pr(t \geq 1.475) = 0.168$, twice the value for the one-tail test. For the calculated t-value to be significant at the 5% level it would have to exceed the value of the percentage point $t_{11}(2.5) = 2.201$. Figure 2 illustrates the difference between the p-value of a one-tail and a two-tail test.

In `R`, you can use the function `t.test`. Type `?t.test` if you want to access help files. Set the argument `mu = 24.9` to specify that the hypothesized mean is 24.9, and to get the *one-tail t-test* with the hypothesis $H_1 : \mu > 24.9$ use the argument `alternative = "greater"`.

```
t.test(height, mu = 24.9, alternative = "greater")

##
##  One Sample t-test
##
## data:  height
## t = 1.4754, df = 11, p-value = 0.08408
## alternative hypothesis: true mean is greater than 24.9
## 95 percent confidence interval:
##  24.64292      Inf
## sample estimates:
## mean of x
##  26.08333
```

If we had used a one-sided alternative hypothesis such as $H_1 : \mu < 24.9$ then we would reject H_0 if $\bar{x}$ were small enough or, equivalently, if t were small enough. The p-value of the calculated test statistic would be a probability in the left-hand tail of the t-distribution. In `R` the argument `alternative = "less"` would be used.

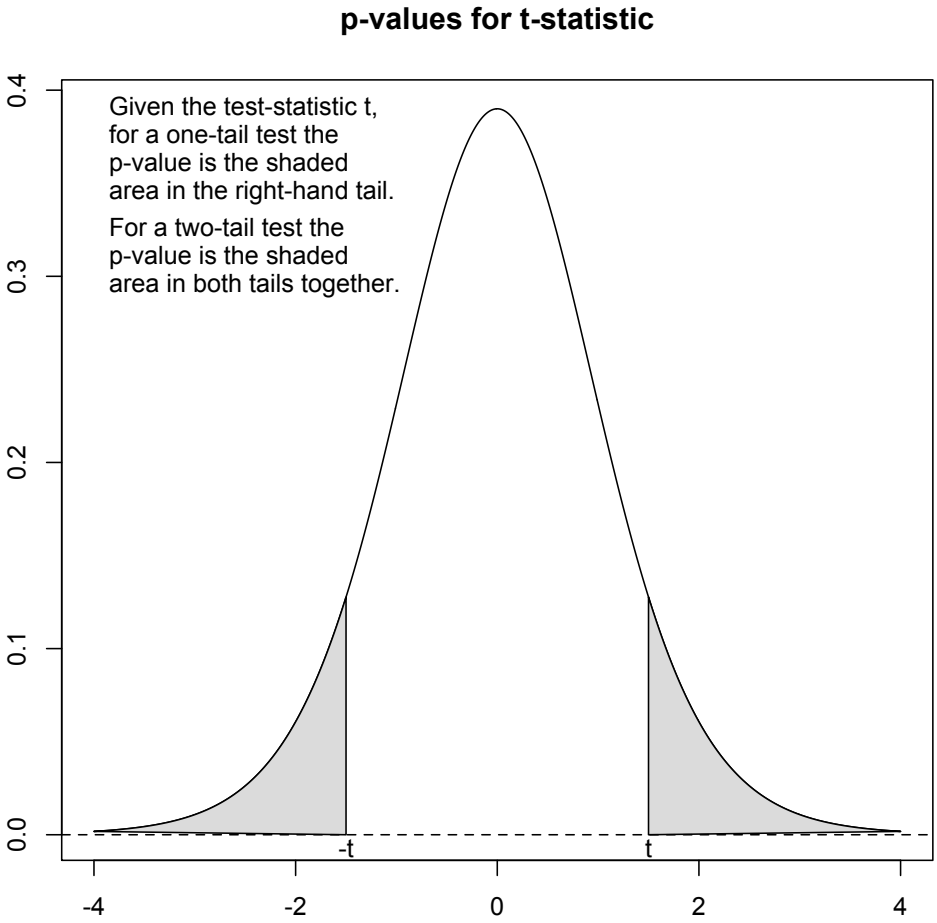

Figure 2: One-tail and two-tail p-values for the t_{11} distribution

10.3 Two-sample t-tests

Suppose that we wish to compare two populations, Population 1 and Population 2. We take a random sample $x_1, x_2, \ldots, x_n$ of size n from Population 1 and, independently of the first sample, a random sample $y_1, y_2, \ldots, y_m$ of size m from Population 2.

We assume that the first sample comes from a $N(\mu_1, \sigma^2)$ distribution and that the second sample comes from a $N(\mu_2, \sigma^2)$ distribution. So we suppose that the two populations may have different means, μ_1 and μ_2, but have the same variance σ^2, where μ_1, μ_2 and σ^2 are unknown.

- Another possible model envisages that the two population variances may differ, so that for Population 1 we assume a $N(\mu_1, \sigma_1^2)$ distribution and for Population 2 a $N(\mu_2, \sigma_2^2)$ distribution. However, it is often reasonable to assume the same variance σ^2 for both populations, and this is commonly done. The t-test that we develop is valid only under the assumption of equal variances for the two populations.

We shall test the null hypothesis that the two population means are equal,

$$H_0 : \mu_1 = \mu_2 \ .$$

The alternative hypothesis may be two-sided,

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

or one-sided,

$$H_1 : \mu_1 > \mu_2$$

or

$$H_1 : \mu_1 < \mu_2 \ .$$

To obtain an appropriate test statistic, note first that

$$\bar{x} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$$

and

$$\bar{y} \sim N\left(\mu_2, \frac{\sigma^2}{m}\right)$$

Hence, since $\bar{x}$ and $\bar{y}$ are independently distributed,

$$\bar{x} - \bar{y} \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

and, under H_0,

$$\frac{\bar{x} - \bar{y}}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim N(0,1) \ . \tag{1}$$

In practice, σ^2 is unknown and we replace it by an estimate s^2, the *pooled estimate of variance*, based on the data from both samples:

$$s^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} \ , \tag{2}$$

where s_1^2 and s_2^2 are the sample variances for the samples from Population 1 and Population 2, respectively. The pooled estimate s^2 as defined by Equation (2) is a weighted average of the individual sample variances s_1^2 and s_2^2. It is an unbiased estimate of σ^2. Equivalently,

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{n + m - 2} \ . \tag{3}$$

It turns out that for the pooled estimate s^2 of Equation (2)/(3),

$$\frac{(n + m - 2)s^2}{\sigma^2} \sim \chi^2_{n+m-2} \ .$$

Note that the degrees of freedom $n+m-2$ associated with s^2 is the sum of the individual degrees of freedom for s_1^2 and s_2^2,

$$n + m - 2 = (n - 1) + (m - 1).$$

Replacing σ^2 in Equation (1) by the pooled estimate s^2, we find that, under H_0,

$$\frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2} \ . \tag{4}$$

The statistic of Equation (4) is the two-sample t-statistic that we use for testing H_0. As for the one sample t-test, depending upon whether we have a two-sided or one-sided alternative hypothesis, we use a two-tail or one-tail test, respectively.

If, instead of carrying out a hypothesis test, we wish to estimate the difference $\mu_1 - \mu_2$, we naturally use the estimate $\bar{x} - \bar{y}$. Adopting a similar approach to the one in Section 9.1, a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\left(\bar{x} - \bar{y} - t_{n+m-2}(50\alpha)s\sqrt{\frac{1}{n} + \frac{1}{m}} \ , \ \bar{x} - \bar{y} + t_{n+m-2}(50\alpha)s\sqrt{\frac{1}{n} + \frac{1}{m}}\right) \ . \tag{5}$$

Example

Twenty cows were used in an experiment to test two types of feed, Feed A and Feed B. Half of the cows, chosen at random, were fed Feed A over a certain period of time and the other half were fed Feed B. Two of the cows on Feed B were wrongly fed for part of the period and they had to be removed from the experiment. The weight gains in kilograms of the remaining cows over the period of the experiment are listed below, subdivided between the cows on Feed A and the cows on Feed B.

Feed A: 15, 13, 15, 10, 13, 19, 14, 19, 13, 16
Feed B: 20, 17, 14, 15, 13, 18, 14, 19

Is there any evidence that the two feeds have differing effects on weight gain?

We assume that the weight gains for Feed A come from a $N(\mu_A, \sigma^2)$ distribution and that the weight gains for Feed B come from a $N(\mu_B, \sigma^2)$ distribution. We test the null hypothesis

$$H_0 : \mu_A = \mu_B$$

against be two-sided alternative

$$H_1 : \mu_A \neq \mu_B \ .$$

To begin with we calculate the basic sample statistics. The average weight gain has been greater for the cows on Feed B, but is the difference between Feed A and Feed B significant?

	Feed A	Feed B
Sample size	10	8
Sample mean	14.7	16.25
Sample variance	7.789	6.786

Table 2: Sample statistics of the weight gains for the two experimental groups

```
feed_A <- c(15, 13, 15, 10, 13, 19, 14, 19, 13, 16)
n <- length(feed_A)
n

## [1] 10

xbar <- mean(feed_A)
xbar

## [1] 14.7

s2_1 <- var(feed_A)
s2_1

## [1] 7.788889

feed_B <- c(20, 17, 14, 15, 13, 18, 14, 19)
m <- length(feed_B)
m

## [1] 8

ybar <- mean(feed_B)
ybar

## [1] 16.25

s2_2 <- var(feed_B)
s2_2

## [1] 6.785714
```

Using Equation (2), the pooled estimate s^2 of the population variance σ^2 is given by

$$s^2 = \frac{9\,(7.789) + 7\,(6.786)}{16} = 7.350\ .$$

Using Equation (4), the t-statistic for testing H_0 is given by

$$t = \frac{14.7 - 16.25}{\sqrt{7.350\left(\frac{1}{10} + \frac{1}{8}\right)}} = -1.21$$

with 16 degrees of freedom.

Because we are using a two-sided alternative hypothesis, we use a two-tail test. If F denotes the distribution function of the t_{16} distribution then in this case, using Table 9 of *Lindley and Scott*, with some interpolation, the p-value is given by

$$p = \Pr(|t| \geq 1.21) = 2(1 - F(1.21)) \approx 2(1 - 0.878) = 0.244.$$

Since $p > 0.05$, we do not reject the null hypothesis at the 5% significance level, or, since $p > 0.2$, even at the 20% level. There is no strong evidence that the two feeds have differing effects on weight gain.

Alternatively we could compare our calculated t-value with the percentage points of the t-distribution with 16 degrees of freedom. From Table 10, $t_{16}(2.5) = 2.120$. Since our calculated value, $|t| = 1.21$, is smaller than this percentage point, we do not reject the null hypothesis at the 5% significance level. As $t_{16}(10) = 1.337$, we do not even reject the null hypothesis at the 20% significance level.

We can readily carry out the analysis using `R`. The function `t.test` allows to perform the two-sample t-test. The first two arguments of the function are two vectors containing the first and the second sample data. The argument `var.equal = TRUE` indicates that equal variances are being assumed for the underlying populations.

```
t.test(feed_A, feed_B,
  var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  feed_A and feed_B
## t = -1.2053, df = 16, p-value = 0.2456
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.27616  1.17616
## sample estimates:
## mean of x mean of y
##     14.70     16.25
```

The output gives a more accurate p-value, based on an unrounded t-value, than we obtained earlier. In addition to the t-test, an estimation approach has also been taken,

where a 95% confidence interval for $\mu_A - \mu_B$ is given (-4.28, 1.18), calculated according to the formula of Equation (5).

10.4 Paired comparisons

Example

In a study using identical twins to test the effect of alcohol on intelligence, for each pair of twins, one of them, selected at random, was given an intelligence test while under the influence of a given dose of alcohol. The other twin was given the same test under alcohol-free conditions. The test scores are listed below in Table 3. Do these data provide evidence that alcohol lowers performance on intelligence tests?

Pair	No Alcohol	Alcohol	difference
1	83	78	5
2	74	74	0
3	67	63	4
4	64	66	-2
5	70	68	2
6	67	63	4
7	81	77	4
8	64	65	-1
9	72	70	2

Table 3: Intelligence test scores

The above experimental design is an example of a "matched pairs" or "paired comparison" design. It represents a more sophisticated form of experiment, which aims to provide more powerful test procedures. We have two sets of results, but they are not two independent samples, as was the case for the two-sample t-test.

More generally, suppose that we have n pairs of observations, so that there are two random samples, $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ of size n. The two samples are not independent of each other. On the contrary, the experimental design results in an association between the members of the ith pair (x_i, y_i) for each i.

Another type of situation which can result in a similar data structure and a similar form of analysis is where x_i and y_i are two different observations on the same individual i, possibly before and after some treatment.

The analysis will be based on consideration of the differences

$$d_i = x_i - y_i \qquad (1 \leq i \leq n).$$

We consider $x_1, x_2, \ldots, x_n$ as a random sample from some population with mean μ_X and $y_1, y_2, \ldots, y_n$ as a random sample from some population with mean μ_Y. It follows that $d_1, d_2, \ldots, d_n$ is a random sample from a population with mean μ_D, where the population means are related by the formula

$$\mu_D = \mu_X - \mu_Y \ .$$

Correspondingly, the sample means are related by

$$\bar{d} = \bar{x} - \bar{y} \ .$$

We assume further that $d_1, d_2, \ldots, d_n$ is a random sample from a $N(\mu_D, \sigma_D^2)$ distribution, where μ_D and σ_D^2 are unknown. The hypotheses that we test may be expressed in terms of μ_X and μ_Y or, equivalently, in terms of μ_D. The various possibilities are shown in Table 4 below.

	Null hypothesis H_0	Alternative hypothesis H_1			
		two-sided	or	one-sided	
In terms of μ_X, μ_Y:	$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	or	$\mu_X < \mu_Y$
In terms of μ_D:	$\mu_D = 0$	$\mu_D \neq 0$	$\mu_D > 0$	or	$\mu_D < 0$

Table 4: Hypotheses for paired comparisons

The method of analysis is to carry out a one sample t-test for the random sample $d_1, d_2, \ldots, d_n$ to test $H_0 : \mu_D = 0$, using the test statistic

$$t = \frac{(\bar{d} - 0)}{\frac{s_D}{\sqrt{n}}} = \frac{\bar{d}}{\frac{s_D}{\sqrt{n}}} \tag{6}$$

with $n - 1$ degrees of freedom, where s_D^2 is the sample variance for $d_1, d_2, \ldots, d_n$.

If, instead of carrying out a hypothesis test, we wish to estimate the difference $\mu_X - \mu_Y \equiv \mu_D$, we use the estimate $\bar{x} - \bar{y} \equiv \bar{d}$. A $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\left(\bar{d} - t_{n-1}(50\alpha)\frac{s_D}{\sqrt{n}} \ , \ \bar{d} + t_{n-1}(50\alpha)\frac{s_D}{\sqrt{n}} \right) \ . \tag{7}$$

Example (continued)

The sample size is $n = 9$, the x_i are the observations with no alcohol, the y_i are the observations with alcohol and the differences $d_i = x_i - y_i$ $(1 \leq i \leq 9)$ are listed in Table 3. We test the null hypothesis $H_0 : \mu_D = 0$ against the one-sided alternative hypothesis $H_1 : \mu_D > 0$. For this simple set of data it is easy to calculate the sample mean $\bar{d} = 2$ and the sample variance $s_D^2 = 6.25$, so that $s_D = 2.5$. Hence, using Equation (6),

$$t = \frac{2}{\frac{2.5}{\sqrt{9}}} = 2.4$$

with 8 degrees of freedom. We use a one-tail test. From Table 10 of *Lindley and Scott*, we see that $t_8(5) = 1.860$ and $t_8(1) = 2.896$. It follows that we reject H_0 at the 5% level although not at the 1% level. There is strong evidence that alcohol lowers performance.

Using Equation (7), a 95% confidence interval for the underlying mean difference in scores between subjects not under the influence of alcohol and those under the influence of alcohol, $\mu_X - \mu_Y \equiv \mu_D$, is given by

$$\left(2 - t_8(2.5)\frac{2.5}{\sqrt{9}} \ , \ 2 + t_8(2.5)\frac{2.5}{\sqrt{9}} \right) = \left(2 - 2.306\,\frac{2.5}{3} \ , \ 2 + 2.306\,\frac{2.5}{3} \right) = (0.08, 3.92) \ .$$

To carry out the analysis in R, we first have to save the data in two vectors, that we call `NoAlcohol` and `Alcohol`. The 9 pairs of twins are thought of as the experimental units, and there are two measurements made on each pair.

```
NoAlcohol <- c(83, 74, 67, 64, 70, 67, 81, 64, 72)
Alcohol <- c(78, 74, 63, 66, 68, 63, 77, 65, 70)
cbind(NoAlcohol, Alcohol)

##       NoAlcohol Alcohol
##  [1,]        83      78
##  [2,]        74      74
##  [3,]        67      63
##  [4,]        64      66
##  [5,]        70      68
##  [6,]        67      63
##  [7,]        81      77
##  [8,]        64      65
##  [9,]        72      70
```

To carry out the paired comparisons test, use the function `t.test`. You have to specify the names of the vectors in which the sample values are to be found (`NoAlcohol` and `Alcohol` in the present example), `alternative = "greater"` in order to get the one-sided test, and the argument `paired = TRUE`.

Note that the p-value is 0.022, consistent with our earlier conclusion that we reject H_0 at the 5% level but not at the 1% level.

```
t.test(NoAlcohol, Alcohol,
  alternative = "greater",
  paired = TRUE)

##
##  Paired t-test
##
## data:  NoAlcohol and Alcohol
## t = 2.4, df = 8, p-value = 0.02159
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4503766       Inf
## sample estimates:
## mean of the differences
##                       2
```

To find the 95% confidence interval for $\mu_X - \mu_Y$ using R, we have to carry out the paired comparisons test with two-sided alternative hypothesis, as follows.

```r
t.test(NoAlcohol, Alcohol,
  paired = TRUE)

##
##  Paired t-test
##
## data:  NoAlcohol and Alcohol
## t = 2.4, df = 8, p-value = 0.04318
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.07832989 3.92167011
## sample estimates:
## mean of the differences
##                       2
```

If we erroneously use a two-sample t-test, we obtain the following output, which yields a non-significant p-value of 0.258.

- The argument `var.equal = TRUE` has been used instead of `paired = TRUE`.

```r
t.test(NoAlcohol, Alcohol,
  alternative = "greater",
  var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  NoAlcohol and Alcohol
## t = 0.66462, df = 16, p-value = 0.2579
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3.253792       Inf
## sample estimates:
## mean of x mean of y
##  71.33333  69.33333
```