MATH11400 Statistics 1 2010–11

Homepage http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/

Comparisons and Regression

In this final section, we look at hypothesis tests and confidence intervals in situations where the data has more structure that just a single sample, based on the assumption that data are Normally distributed.

9.1 Introduction

In previous sections we have concentrated on cases where the data could be modelled as a single random sample from a parametric distribution with one or more unknown parameters. This type of model is applicable when there are no systematic differences between the experimental units, so any differences in the data values from unit to unit are attributable to random variation. Our focus was on using EDA to identify an appropriate population distribution, on estimating the parameters of this distribution and evaluating the accuracy of the estimates, and on evaluating the evidence for or against a simple hypothesis about the parameter values.

In reality, data are usually collected in order to make comparisons between groups, and/or to study how the main variable of interest (the *response* variable) depends one one or more *explanatory* variables. These are really different versions of the same question – if data arise in different groups (age groups, groups receiving different treatments, groups from different countries, etc.), then we can think of each data item being accompanied by a *label* indicating the group to which that item belongs, so by studying the way in which the response variable depends on this label, we are comparing groups. In this case, the explanatory variable is discrete or *categorical*, and often referred to as a *factor*.

In the remainder of the course we consider these problems of comparison and dependence. The response variable will be influenced by both systematic and random variation, and we can think of the task of the statistician being to separate these two effects.

9.2 A general framework for tests and confidence intervals using the t distribution

Suppose that we have a statistical model in which we have a parameter of interest θ, and an estimator $\widehat{\theta}$ of this parameter, with the property that

$$\widehat{\theta} \sim N(\theta, \sigma^2_{\widehat{\theta}}). \tag{1}$$

The variance $\sigma^2_{\widehat{\theta}}$ is unknown but we have an estimator of it, which we denote $S^2_{\widehat{\theta}}$, and we suppose that

$$rS^2_{\widehat{\theta}}/\sigma^2_{\widehat{\theta}} \sim \chi^2_r \tag{2}$$

for some r (the degrees of freedom), and that

$$\widehat{\theta} \quad \text{and} \quad S^2_{\widehat{\theta}} \quad \text{are independent.} \tag{3}$$

Results in Section 6 of the course tell us quite a lot about this situation. From (1) and §6.2(c) we know that

$$\frac{\theta - \widehat{\theta}}{\sigma_{\widehat{\theta}}} \sim N(0,1) \tag{4}$$

and then (2), (3) and §6.10 imply that

$$\frac{\theta - \widehat{\theta}}{S_{\widehat{\theta}}} \sim t_r. \tag{5}$$

This is a very general result, with far-reaching consequences, because the assumptions are so general. It is mainly used to construct hypothesis tests and confidence intervals.

Hypothesis tests. It follows from (5) that in testing the hypothesis $H_0 : \theta = \theta_0$ (where θ_0 is a fixed number), we can use the test statistic

$$T = \frac{\theta - \widehat{\theta}}{S_{\widehat{\theta}}}$$

and know that under H_0, $T \sim t_r$.

Confidence intervals. It also follows from (5) that

$$P\left\{ -t_{r;\alpha/2} < \frac{\theta - \widehat{\theta}}{S_{\widehat{\theta}}} < t_{r;\alpha/2} \right\} = 1 - \alpha,$$

and after the usual manipulations (e.g. see §7.3), this means that (c_L, c_U) is a $100(1-\alpha)\%$ confidence intervals for θ, where

$$c_L = \widehat{\theta} - t_{r;\alpha/2} S_{\widehat{\theta}} \quad \text{and} \quad c_U = \widehat{\theta} + t_{r;\alpha/2} S_{\widehat{\theta}}.$$

9.3 Example: a single sample from $N(\mu, \sigma^2)$

Suppose that we are in the familiar situation of assuming that $X_1, X_2, \ldots, X_n$ are a simple random sample from $N(\mu, \sigma^2)$, with both parameters unknown. Then we can see that this is an example of the general set-up of §9.2, with

$$\theta = \mu, \quad \widehat{\theta} = \overline{X}, \quad \sigma^2_{\widehat{\theta}} = \sigma^2/n, \quad S^2_{\widehat{\theta}} = S^2/n, \quad r = n - 1.$$

The confidence interval for $\theta \equiv \mu$ in §9.2 is the same as that in §7.3, and the hypothesis test in §9.2 is the same as that in §8.6 (the one-sample t test).

In the remainder of the section, we will look at some more interesting examples, the application of the results in §9.2 to linear regression and to comparison of two populations. These two special cases are examples of a larger much more general class of methods in statistics for studying dependence and comparison.

9.4 Linear regression

In this section, we return to Linear Regression, as seen in Section 4. But, now we have laid the theoretical groundwork, we can perform inference on the parameters of the model, for example tests of hypotheses and confidence intervals; before we only discussed estimation of the parameters.

As mentioned in the introduction, we are thinking of a kind of comparison of populations. Instead of discretely separated groups, what we are comparing are the populations of potential Y values for different values of x, so it is a *continuous* variable (x) that distinguishes our populations of interest. In this course, we only consider one-dimensional x, and a linear dependence of Y on x, but the subject of Regression becomes much more general later.

Here we will use the assumptions and notation of §4:

$$Y_i = \alpha + \beta x_i + e_i,$$

for unknown parameters α, β, together with the assumption that the e_i, $i = 1, 2, \ldots, n$ are i.i.d. $N(0, \sigma^2)$. This is equivalent to saying that $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, independently for each $i = 1, 2, \ldots, n$.

9.5 Distribution of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$
Lemma (partly seen in §4.10)

(i) $\mathrm{E}(\hat{\beta}) = \beta$, $\mathrm{Var}(\hat{\beta}) = \sigma^2/ss_{xx}$ and $\hat{\beta} \sim N(\beta, \sigma^2/ss_{xx})$

(ii) $\mathrm{E}(\hat{\alpha}) = \alpha$, $\mathrm{Var}(\hat{\alpha}) = \sigma^2(1/n + \bar{x}^2/ss_{xx})$ and $\hat{\alpha} \sim N(\alpha, \sigma^2(1/n + \bar{x}^2/ss_{xx}))$

(iii) $(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$ and $\hat{\sigma}^2$ is independent of $(\hat{\alpha}, \hat{\beta})$.

Proof (i) For a given data set, the y_i are observed values of the random variables Y_i corresponding to the given values of x_i. If we took repeated independent sets of samples of the Y_i keeping the set of values of the predictor variable fixed at $x_1, \ldots, x_n$, then the value of $\hat{\beta}$ would vary from sample to sample as we would get different sets of values for $y_1, \ldots, y_n$. Thus, considered as a random variable,

$$\hat{\beta} = \frac{\sum_1^n Y_i x_i - (\sum_1^n Y_i \sum_1^n x_i)/n}{\sum_1^n x_i^2 - (\sum_1^n x_i)^2/n} = \sum_1^n Y_i \frac{(x_i - \bar{x})}{ss_{xx}} = \sum_{i=1}^n b_i Y_i$$

where, for given fixed values of $x_1, \ldots, x_n$, the $b_i = (x_i - \bar{x})/ss_{xx}, \quad i = 1, \ldots, n$ are fixed constants and the Y_i are independent Normally distributed random variables with mean $\mathrm{E}(Y_i) = \alpha + \beta x_i$ and variance $\mathrm{Var}(Y_i) = \sigma^2$.

From the results in §6 we can immediately deduce that $\hat{\beta}$ has a Normal distribution, since it is a linear combination of independent Normally distributed random variables. To calculate the mean and variance for $\hat{\beta}$ we first note that

$$\mathrm{E}(\hat{\beta}) = \mathrm{E}(\textstyle\sum_1^n b_i Y_i) = \sum_1^n b_i \,\mathrm{E}(Y_i) = \sum_1^n b_i(\alpha + \beta x_i) = \alpha \sum_1^n b_i + \beta \sum_1^n b_i x_i$$

and

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\textstyle\sum_1^n b_i Y_i) = \sum_1^n b_i^2 \,\mathrm{Var}(Y_i) \text{ (as the } Y_i \text{ are independent) } = \sigma^2 \sum_1^n b_i^2.$$

But $$\textstyle\sum_1^n b_i = \dfrac{1}{ss_{xx}} \displaystyle\sum_1^n (x_i - \bar{x}) = \frac{1}{ss_{xx}} 0 = 0$$

and $$\textstyle\sum_1^n b_i x_i = \dfrac{1}{ss_{xx}} \displaystyle\sum_1^n (x_i - \bar{x}) x_i = \frac{\sum_1^n x_i^2 - \bar{x} \sum_1^n x_i}{ss_{xx}} = \frac{\sum_1^n x_i^2 - n\bar{x}^2}{ss_{xx}} = \frac{ss_{xx}}{ss_{xx}} = 1$$

and $$\textstyle\sum_1^n b_i^2 = \dfrac{1}{(ss_{xx})^2} \displaystyle\sum_1^n (x_i - \bar{x})^2 = \frac{ss_{xx}}{(ss_{xx})^2} = \frac{1}{ss_{xx}}.$$

Thus $$\mathrm{E}(\hat{\beta}) = \alpha\, 0 + \beta\, 1 = \beta \quad \text{and} \quad \mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{ss_{xx}}$$

(ii) Derivation of the distribution for $\hat{\alpha}$ is very similar to that for $\hat{\beta}$. We start by noting that $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} = \sum_1^n Y_i/n - \bar{x}\sum_1^n b_i Y_i = \sum_1^n Y_i(1/n - b_i\bar{x}) = \sum_{i=1}^n a_i Y_i$ where $a_i = (1/n - b_i\bar{x}),\ i = 1, \ldots, n$. This in turn means $\hat{\alpha}$ has a Normal distribution and gives $\mathrm{E}(\hat{\alpha}) = \alpha\sum_1^n a_i + \beta\sum_1^n a_i x_i$ and $\mathrm{Var}(\hat{\alpha}) = \sigma^2\sum_1^n a_i^2$. Finally, using the facts that $\sum_1^n b_i = 0$, $\sum_1^n b_i x_i = 1$ and $\sum_1^n b_i^2 = 1/ss_{xx}$, one can easily deduce that $\sum_1^n a_i = 1$, $\sum_1^n a_i x_i = 0$ (so $\mathrm{E}(\hat{\alpha}) = \alpha$) and $\sum_1^n a_i^2 = (1/n + \bar{x}^2/ss_{xx}) = \sum x_i^2/n\, ss_{xx}$ (so $\mathrm{Var}(\hat{\alpha}) = \sigma^2(1/n + \bar{x}^2/ss_{xx}) = \sigma^2\sum x_i^2/n\, ss_{xx}$).

(iii) The proof of part(iii) is similar to that in §6.5, and is omitted.

9.7 Example - the Leaning Tower of Pisa

We previously met this example in §4.

Some of the basic arithmetic:
$\sum_i x_i = 1053$, $\sum_i y_i = 9018$, $\sum_i x_i^2 = 85475$, $\sum_i y_i^2 = 6271714$, $\sum_i x_i y_i = 732154$.
So $\overline{x} = 81$, $\overline{y} = 693.6923$, $ss_{xx} = 182$, $ss_{yy} = 15996.77$, $ss_{xy} = 1696$
and then $\hat{\beta} = 9.319$, $\hat{\alpha} = -61.121$, $\hat{\sigma}^2 = 17.481$, $S^2_{\hat{\alpha}} = 631.51$, $S^2_{\hat{\beta}} = 0.096047$.
Finally, $\hat{\sigma} = 4.181$, $S_{\hat{\alpha}} = 25.130$, $S_{\hat{\beta}} = 0.3099$.

To be continued on the board.....

9.8 Confidence Intervals and Hypothesis Tests using the `summary` command in R

Consider the simple Normal linear regression model $Y_i = \alpha + \beta x_i + e_i$, where the e_i are i.i.d. $\mathrm{N}(0, \sigma^2)$. Assume the predictor values $x_1, \ldots, x_n$ are contained in an **R** data vector called `xdata` and the response values $y_1, \ldots, y_n$ are contained in an **R** data vector called `ydata`, and assume we want our analysis to be contained in the **R** object `xyoutput`.

We have already seen how to produce the output using the **R** command

```
> xyoutput <- lm(ydata ~ xdata)
```

and how to perform exploratory data analysis, estimation and assessment of fit using the follow-up commands `plot`, `coef`, `fitted`, and `residuals`.

For confidence intervals and hypothesis tests, most of the necessary information can be obtained with the `summary` command. For example

```
> summary(xyoutput)
```

produces the following output, where the formulae shown in each box is replaced in the actual output by its numerical value.

```
Call:
lm(formula = ydata ~ xdata)

Residuals:
   Min     1Q   Median  3Q    Max

Coefficients:
                  Estimate     Std. Error      t value  Pr(>|t|)
(Intercept)   $\boxed{\hat\alpha = \bar y - \hat\beta \bar x}$   $\boxed{S_{\hat\alpha} = \hat\sigma\sqrt{\frac{1}{n} + \frac{\bar x^2}{ss_{xx}}}}$   $\boxed{\hat\alpha / S_{\hat\alpha}}$   $\boxed{2(1 - F_{t_{n-2}}(|\hat\alpha / S_{\hat\alpha}|))}$
xdata         $\boxed{\hat\beta = ss_{xy}/ss_{xx}}$   $\boxed{S_{\hat\beta} = \hat\sigma/\sqrt{ss_{xx}}}$   $\boxed{\hat\beta / S_{\hat\beta}}$   $\boxed{2(1 - F_{t_{n-2}}(|\hat\beta / S_{\hat\beta}|))}$
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
Residual standard error: $\boxed{\hat\sigma = \sqrt{\frac{ss_{yy} - ss_{xy}^2/ss_{xx}}{n-2}}}$ on $\boxed{n-2}$ degrees of freedom
```

Thus the output:
(i) first of all it reproduces the formula used to produce the output, so you can check exactly which model is being analysed;
(ii) then it produces a 5-number summary of the residual values (or lists in full the numerical values of the residuals if there are only a few of them);
(iii) then it lists the relevant values for constructing confidence intervals and performing hypothesis tests on the linear model coefficients α and β – first for α (the intecept in the model) and then for β (the coefficient of x in the model);
(iv) then it lists numerical values relevant to estimating σ^2 (or, more precisely, σ);
(v) and finally it gives information on the R-Squared and F-statistic values (which I've omitted as they are not covered in this unit).

In particular, the values on the line beginning `(Intercept)` are:
(i) $\hat\alpha$ (the estimate of α),
(ii) $S_{\hat\alpha}$ (the standard error, which estimates the standard deviation $\hat\alpha$),
(iii) $t_{obs} = \hat\alpha / S_{\hat\alpha}$ (the observed test statistic for testing $H_0{:}\alpha = 0$ vs. $H_1{:}\alpha \neq 0$),
(iv) $P(|W| > |t_{obs})$, where $W \sim t_{n-2}$ (the p-value of the data for the test).

The result of a hypothesis test of $H_0{:}\alpha = 0$ vs. $H_1{:}\alpha \neq 0$ can then be deduced immediately from the corresponding p-value. Moreover, the endpoints for a $100(1-\gamma)\%$ confidence interval for α can be calculated using the values of $\hat\alpha$, $S_{\hat\alpha}$ and the appropriate t-distribution percentage point $t_{n-2;\gamma/2}$.

The values on the line beginning `xdata` are the corresponding quantities for estimating, constructing confidence intervals or performing hypothesis tests on β:
i.e. (i) $\hat\beta$, (ii) $S_{\hat\beta}$, (iii) $t_{obs} = \hat\beta / S_{\hat\beta}$, and (iv) $P(|W| > |t_{obs})$, where $W \sim t_{n-2}$.
A $100(1-\gamma)\%$ confidence interval for β can be obtained in a similar manner to that for α.

9.9 The Leaning Tower of Pisa example in R

In our previous analysis (§4) we had typed `pisafit<-lm(tilt~year,data=pisa)` to carry out the linear regression. Applying the `summary(pisafit)` command using this previous result produces the output below. You can (and should) check that the values shown correspond to the appropriate values calculated in your notes when we constructed confidence intervals and performed hypothesis tests on α and β.

From the output we can, for example, immediately read off the least squares estimate $\hat{\beta} = 9.3187$ and its standard error $S_{\hat{\beta}} = 0.3099$. We can also see that the p-value for testing $H_0{:}\beta = 0$ versus $H_1{:}\beta \neq 0$ is extremely small (6.5×10^{-12}) and so there is very strong evidence that β is not zero and the mean tilt does vary significantly with the year.

```
Call:
lm(formula = tilt ~ year)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9670 -3.0989  0.6703  2.3077  7.3956

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -61.1209    25.1298  -2.432   0.0333 *
year          9.3187     0.3099  30.069   6.5e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.181 on 11 degrees of freedom
Multiple R-Squared: 0.988,      Adjusted R-squared: 0.9869
F-statistic: 904.1 on 1 and 11 DF,  p-value: 6.503e-12
```

9.10 Comparison of two groups

We now look at the situation of the comparison of two groups – that is of an explanatory variable that is *binary* – takes just two values. We have two groups of data, and suspect there are systematic differences between the groups which may affect the data. So we assume that the data values for each group come from distinct populations. The question of interest is whether there are systematic differences between the populations in the value of some quantity of interest (corresponding to differences in the factor values). The groups might be defined by properties of the experimental units (human subjects, etc.) or by different treatments (drug therapies, for example) applied to those units.

For simplicity, we restrict ourselves to the case where the data can be assumed normally distributed, the quantity of interest is the population mean, and to the very specific question of testing whether observed differences are statistically significant.

Sometimes, we may be able to assume that the values in each data set are entirely independent of each other and of those in the other data set. In this case the data can be modelled as two independent random samples from different population distributions. Here the question of interest reduces to whether the means of the two populations differ. This type of model can be analysed using a **two sample t-test**.

Alternatively, when we are studying differences between treatments, the experiment *may* be designed so that each treatment is applied to each experimental unit, so the data consist of pairs of observations on each of n experimental units, with the first observation in each pair corresponding to one factor value and the second corresponding to the other. Whatever the differences between the experimental units, it may be plausible to assume that the change in factor value is associated with a common systematic change in the underlying distribution of the variable being measured. An appropriate model is often that the *differences* between the observations in each pair are independent observations from the same distribution, whose mean corresponds to the systematic change, and the question of interest reduces to whether or not this mean change is zero. This type of model can be analysed using a **paired t-test**.

9.11 Two sample t-test

For the two sample t-test the model assumptions are that the data consists of two independent random samples, where $X_1, \ldots, X_n$ is a random sample of size n from the Normal $N(\mu_X, \sigma_X^2)$ distribution and $Y_1, \ldots, Y_m$ is a random sample of size m from the Normal $N(\mu_Y, \sigma_Y^2)$ distribution. Denote the sample means by $\bar{X} = (X_1 + \cdots + X_n)/n$ and $\bar{Y} = (Y_1 + \cdots + Y_m)/m$.

The null hypothesis of interest is $H_0 : \mu_X = \mu_Y$, or equivalently $H_0 : \mu_X - \mu_Y = 0$. The standard estimators of μ_X and μ_Y are $\bar{X}$ and $\bar{Y}$, so it is natural to base our analysis on the value of $\bar{X} - \bar{Y}$. From §6.4, $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$ and $\bar{Y} \sim N(\mu_Y, \sigma_Y^2/m)$, so from §6.3 $\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \sigma_X^2/n + \sigma_Y^2/m)$.

The situation is slightly more complicated than the single sample case. If we can assume the X and Y distributions have the same variance, then we can combine our estimates of σ_X^2 and σ_Y^2 into a single **pooled** estimate, and the resulting test statistic does have a standard t-distribution. If we cannot make this assumption, then a result due to **Welch** shows that the distribution of the test statistic can be approximated by a t-distribution with non-integer degrees of freedom. We deal with these two cases in turn below.

9.11.1 Pooled two sample t-test

Here we are prepared to make the extra model assumption that $\sigma_X^2 = \sigma_Y^2 = \text{ (say) } \sigma^2$.

Denote the sample variances by $S_X^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ and $S_Y^2 = \sum_{j=1}^m (Y_j - \bar{Y})^2/(m-1)$. Since both of these are independent estimates of the common variance σ^2, we can combine them into the *pooled* estimate

$$S_p^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{n+m-2} = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

Since $(n-1)S_X^2$ and $(m-1)S_Y^2$ are independent and have χ_2 distributions, by §6.8, $(n-1)S_X^2 + (m-1)S_Y^2 \sim \chi_{n+m-2}^2$. So we are again in the general situation of §9.2, with

$$\theta = \mu_X - \mu_Y, \quad \widehat{\theta} = \overline{X} - \overline{Y}, \quad \sigma_{\widehat{\theta}}^2 = \sigma^2(1/n + 1/m), \quad S_{\widehat{\theta}}^2 = S_p^2(1/n + 1/m), \quad r = n+m-2.$$

Thus the test statistic becomes

$$T = (\bar{X} - \bar{Y})/S_p\sqrt{\frac{1}{n} + \frac{1}{m}} \quad \text{and } T \sim t_{n+m-2} \text{ when } H_0 \text{ is true.}$$

9.11.2 Welch two sample t-test

In the general case, when $\sigma_X^2 \neq \sigma_Y^2$, the natural estimators of the population variances are the corresponding sample variances. Put

$$\hat{\sigma}_X^2 = S_X^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1) \quad \text{and} \quad \hat{\sigma}_Y^2 = S_Y^2 = \sum_{j=1}^{m} (Y_j - \bar{Y})^2/(m-1).$$

The test statistic is then: $\quad T = (\bar{X} - \bar{Y}) \Big/ \sqrt{\dfrac{\hat{\sigma}_X^2}{n} + \dfrac{\hat{\sigma}_Y^2}{m}}.$

A result due to Welch shows that: $\quad T \simeq t_\nu$ when H_0 is true. Note that this is only an approximation. The degrees of freedom are not necessarily integer, and are computed as:

$$\nu = \frac{\left(\dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}\right)^2}{\dfrac{1}{n-1}\left(\dfrac{S_X^2}{n}\right)^2 + \dfrac{1}{m-1}\left(\dfrac{S_Y^2}{m}\right)^2}.$$

When the X and Y distributions have similar unimodal shapes, the approximation to the distribution of the test statistic is reasonably good for $n \geq 5$ and $m \geq 5$. Note also that, when the sample quantities (sample sizes and sample variances) are similar for the two samples, then the degrees of freedom ν will be close to the value $n + m - 2$ used in the pooled test.

9.12 Paired t-test

For the paired t-test, the data consists n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. Denote the difference between the values in each pair by $W_i = X_i - Y_i$. The model assumption is then that $W_1, \ldots, W_n$ are a random sample from the $N(\delta, \sigma^2)$ distribution, where δ and σ^2 are unknown, and the null hypothesis of interest is $H_0 : \delta = 0$.

Again we are in the general situation of §9.2, with now:

$$\theta = \delta, \quad \widehat{\theta} = \overline{W} = \overline{X} - \overline{Y}, \quad \sigma_{\widehat{\theta}}^2 = \sigma^2/n, \quad S_{\widehat{\theta}}^2 = S_W^2/n, \quad r = n-1.$$

Here, $S_W^2 = \sum_{i=1}^n (W_i - \overline{W})^2/(n-1)$. The test statistic is then

$$T = \frac{\sqrt{n}\,\overline{W}}{S_W} \quad \text{and } T \sim t_{n-1} \text{ when } H_0 \text{ is true.}$$

Again, we reject H_0 if the value of the test statistic is significantly different from zero, where the relevant direction of the difference will depend on the particular alternatives of interest.

Note that the model assumptions do not necessarily require that $X_1, \ldots, X_n$ all have the same distribution. For example, suppose that each $X_i \sim N(\mu_i, \tau^2)$ and that each $Y_i \sim N(\mu_i - \delta, \tau^2)$, where the μ_i may all be different. This corresponds to a situation where the underlying mean value μ_i for each experimental unit may vary from unit to unit, but where the systematic difference in the mean due to the factor is the same for all units. This would still be consistent with the model assumptions above, since it still implies that each $X_i - Y_i \sim N(\delta, \sigma^2)$, where $\sigma^2 = 2\tau^2$.

This type of experimental design may be particularly appropriate if the experimental units are quite variable. In this case, small but consistent systematic differences may show up in an experiment that uses paired observations, but may not be detected by an experiment that uses two independent samples because the small differences in the mean are masked by the high variability between the experimental units.

9.13 t-test procedures in R

Welch two sample t-test

The default two sample t-test in **R** is the Welch test. Assume the two random samples are in data arrays `xdata` and `ydata`. A test of the null hypothesis $H_0 : \mu_X - \mu_Y = 0$ against the two sided alternative $H_A : \mu_X - \mu_Y \neq 0$ can be performed using the command

```
> t.test(xdata,ydata)
```

The output includes the value of the test statistic, the degrees of freedom ν for the approximating t-distribution and the (approximate) p-value.

The option `alternative="less"` can be used to test against the alternative $H_A : \mu_X - \mu_Y < 0$ as in the command

```
> t.test(xdata,ydata,alternative="less")
```

Similarly the option `alternative="greater"` can be used to test against the alternative $H_A : \mu_X - \mu_Y > 0$.

Pooled two sample t-test

Again assume the two random samples are in data arrays `xdata` and `ydata`. Under the model assumption that the population variances are equal, a pooled t-test of the null hypothesis $H_0 : \mu_X - \mu_Y = 0$ agaist the two sided alternative $H_A : \mu_X - \mu_Y \neq 0$ can be performed using the command

```
> t.test(xdata,ydata,var.equal=T)
```

Other alternatives can be specified using the `alternative=`$\cdots$ option as above. Again, the output includes the value of the test statistic, the degrees of freedom and the p-value.

Paired t-test

For the paired t-test, the data is assumed to be in equal-length data arrays `xdata` and `ydata`, where each component of `xdata` will be paired with the corresponding component of `ydata`. A paired t-test of the null hypothesis $H_0 : \delta = 0$ agaist the two sided alternative $H_A : \delta \neq 0$ can then be performed using the command

```
> t.test(xdata,ydata,paired=T)
```

and other alternatives can be specified using the `alternative=`$\cdots$ option. As usual, the output includes the value of the test statistic, the degrees of freedom and the p-value.