

STAT0029 STATISTICAL DESIGN OF INVESTIGATIONS

Department of Statistical Science
University College London

Contents

1	Introduction	4
1.1	Why design?	4
1.2	Three examples	4
1.3	Experimentation versus sampling	7
2	Designed experiments	8
2.1	Introduction	8
2.1.1	Principles of experimental design	8
2.1.2	Planning of experiments	9
2.2	Simple comparative experiments	10
2.2.1	Two independent samples	10
2.2.2	Matched pairs	15
2.3	Completely randomised design	18
2.3.1	One-way analysis of variance	18
2.3.2	Checking model assumptions	25
2.3.3	Design considerations	26
2.4	Randomised block design	26
2.4.1	Two-way analysis of variance	27
2.4.2	Analysis	29
2.4.3	Partition of the treatment s.s.	31
2.4.4	Interaction	32
2.4.5	Multiple observations per cell	33
2.4.6	Unbalanced data	34
2.5	Latin square design	35
2.5.1	Latin square design: theory	35

2.6	Contrasts	37
2.7	Factorial experiments	40
2.7.1	Factorial designs	40
2.7.2	The 2^2 design	41
2.7.3	Partition of a treatment s.s. into single d.f. components	43
2.7.4	The 2^n design	45
2.8	Random effects	46
2.8.1	One-way random-effects model	47
2.8.2	Two-factor random-effects model	48
2.8.3	Two-way mixed-effects model	50
2.8.4	More general models	51
2.9	Nested designs	51
2.10	Split-plot designs (<i>Optional</i>)	54
2.11	Incomplete blocks (<i>Optional</i>)	55
2.12	The general linear model	56
2.12.1	Theory of the linear model: outline	57
2.12.2	Example 12: regression of means in the one-way model	58
2.12.3	Example 13: parallel regression and the analysis of covariance	60
2.13	Departures from assumptions	62
2.13.1	Transformation of variables	63
2.13.2	The Box-Cox transformation (<i>Optional</i>)	64
2.13.3	Generalised linear models	64
2.14	Optimal design	65
3	Observational studies	67
3.1	Introduction	67
3.2	Observational studies <i>versus</i> experiments	67
3.3	Sources of bias in observational studies	69
3.3.1	Confounding variables	69
3.3.2	Other sources of bias	70
3.3.3	The effect of bias	71
3.4	Planning of observational studies	72
3.5	Design, analysis and interpretation	73
3.5.1	Matching	73
3.5.2	Adjustment	77

3.5.3	Instrumental variables	78
3.5.4	Criteria for judging causality	78
3.6	Epidemiological observational studies	79
3.6.1	Cohort studies	79
3.6.2	Case-control studies	79
3.6.3	Comparison of cohort and case-control studies	82
4	Surveys and sampling	83
4.1	Introduction	83
4.2	Sample design	83
4.3	Basic sampling theory	86
4.3.1	Simple random samples	86
4.3.2	Binary values	88
4.3.3	Stratified samples	89
4.3.4	Confidence intervals	90
4.3.5	Optimal sample sizes	90
4.3.6	Cluster and multi-stage sampling	92
4.3.7	Examples	93
4.4	Ratio and regression estimators	96
4.4.1	Bivariate sampling	96
4.4.2	Estimation of a ratio	97
4.4.3	Ratio estimators	98
4.4.4	Regression estimators (<i>Optional</i>)	99
4.5	Data collection (<i>Optional</i>)	100

1 Introduction

1.1 Why design?

Most of modern statistical science is concerned with the *interpretation of information (data) in the presence of uncertainty*. The interpretation is usually based on *probability models* which are intended to represent, in a useful way, the mechanism that generated the data. In these models, the probabilistic framework is the key to dealing with uncertainty. By using the data to make inferences about the values of model parameters (for example, to derive a confidence interval for a mean, or test a hypothesis that the slope of a regression line is zero), we hope to be able to answer questions of interest.

All of this presupposes, however, that the data at hand are suitable for their intended purpose. Consider the following example (fictitious, but perhaps not too unrealistic): a sociologist wants to determine whether or not alcoholism is associated with social deprivation. She contacts her local branches of Alcoholics Anonymous and asks for volunteers to take part in a survey. Based on the information they provide, she calculates that only 15% of them can be classed as ‘socially deprived’. She concludes from this that there is not really any evidence of an association.

Do you agree with this interpretation —and if not, why not?

This course deals with the problem of ensuring that your data *are* suitable for their intended purpose. The problem can be viewed in two different ways: either you could start with a substantive question and ask ‘What data do I need to gather to answer this question with a specified degree of precision?’, or you could start with some data and ask ‘What questions do I have a chance of answering given these data?’. Invariably, there are constraints on the type or amount of data that are available or can be collected; in this case one might be interested in making the best possible use of the available resources. Quite often, if the data are collected or gathered in an intelligent way, the subsequent statistical analysis can be done very simply. To some extent, the experimental design and the method of analysis go hand in hand. We will therefore discuss both in this course.

1.2 Three examples

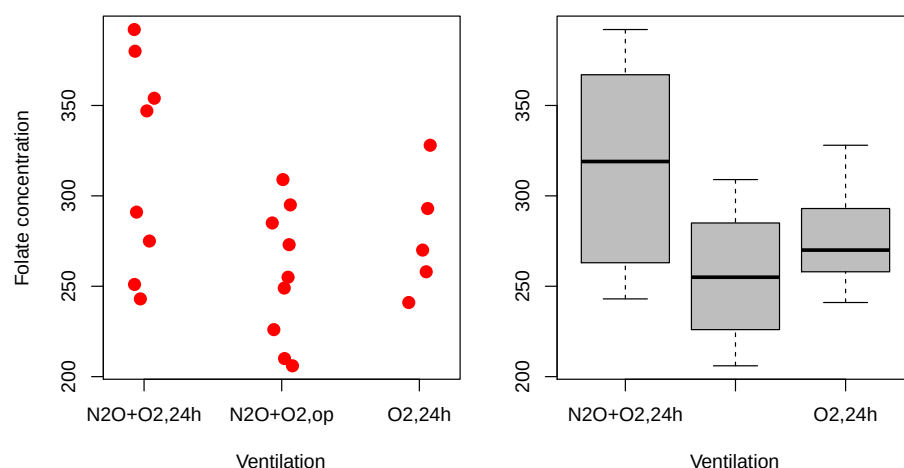
Example 1: an experiment. Folate is a B vitamin that is naturally present in foods, and is also available as a dietary supplement. It is essential for certain bodily functions and it has to be supplied through the diet to meet daily requirements. Humans require folate to produce healthy red blood cells and prevent anemia. Data are available on folate concentration in patients receiving three different methods of ventilation during anesthesia. The example is taken from Altman’s *Practical Statistics for Medical Research* (1991), and the data are available in the R-package ISwR.

Higher values for folate concentration (microgram per liter) are good. The ventilation methods

(treatments) are

- N2O+O2,24h: 50% nitrous oxide and 50% oxygen, continuously for 24 hours.
- N2O+O2,op: 50% nitrous oxide and 50% oxygen, only during operation.
- O2,24h: no nitrous oxide but 35%-50% oxygen for 24 hours.

Scientific question: how do the different treatments compare in terms of folate levels?



Discussion questions:

- How do you decide which patients gets which treatment? Does it matter?
- How would you analyse these data?
- What use will be made of these data?

Example 2: an observational study. In a famous study published in the *British Medical Journal* in 1950, Hill and Doll sampled 1357 cases of lung cancer from four hospitals in the UK along with 1357 hospital-based controls. The two groups were compared as regards smoking history, to yield the following data:

	Cases	Controls	Total
Smoker	1350	1296	2646
Non-smoker	7	61	68
Total	1357	1357	2714

This is a *case-control study* in which individuals are chosen based on *outcome* (cancer or no cancer), and asked about *exposure* (smoking history). The study is *retrospective*.

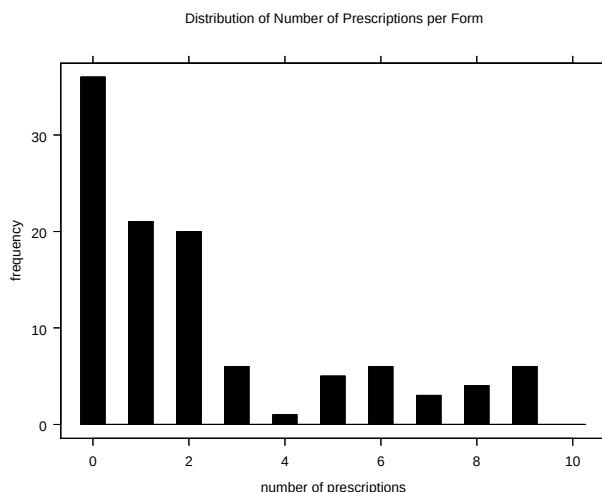
An experiment involves *manipulation*. In Example 1, e.g., patients were allocated to treatments. A case-control does not involve manipulation. Individuals are not allocated to exposure groups —hence *observational study*.

Scientific question: is there an association between smoking and lung cancer?

Discussion questions:

- How much difference would it make to the precision of the conclusions if the sample size had been smaller?
- How much difference would it make if there had been fewer cancer cases?
- How would you test for an association between cancer and smoking?
- What interpretations are there for an association?

Example 3: a sample survey. The *population* consists of thousands of pharmaceutical record forms in Pembury Hospital in Tunbridge Wells, Kent, England. The *sample* consists of 108 forms chosen from patients discharged between 1 July and 31 December 1976. The variable of interest is the number of prescriptions recorded on each was noted:



Note that the results are *descriptive* — hopefully reflecting characteristics of larger population.

Discussion questions

- How would you choose the sample?

- Suppose that the sample of 108 forms was chosen from 5,000. Would it make any difference if there were 10,000 to choose from?

1.3 Experimentation versus sampling

In a controlled **experiment** the system under study is set up and controlled by the investigator. Typically, one of a variety of **treatments** is applied to each individual and the **response** measured. If the allocation of treatments involves **randomisation**, then it will be possible to conclude with some confidence that any clear-cut differences in response between treatments are a consequence of the treatments (causation). Notice that since experiments are run under controlled conditions, they are often unrepresentative of ‘real world’ conditions (there are exceptions, for example in industrial quality control). However, they provide clear-cut conclusions under experimental conditions.

In an **observational study**, the investigator is not able to control the system. Whilst it may be possible to detect clear effects (*e.g.* differences in response between groups of individuals) in such situations, interpretation of these effects is difficult. Explanatory variables that would provide the ‘real’ explanation of the differences may not have been measured and may not even be known to the investigator.

A **sample survey** is a particular type of observational study in which a sample is drawn from a well-defined population by methods, usually involving randomisation, under the control of the investigator. Conclusions can be drawn with confidence about descriptive properties of the population, but interpretation of relationships between variables is again potentially problematic. Compare this with the characteristics of a controlled experiment: there, the conclusions are more clear-cut but possibly less representative.

2 Designed experiments

2.1 Introduction

We will chiefly be concerned with **comparative experiments**, in which we wish to compare the effects of two or more **treatments** (procedures, varieties, *etc.*) on some **response variable** (as opposed to an interest in the absolute effect of a treatment).

Examples:

Field of study	Effect of different ...	on ...
Agriculture	fertilizers/varieties	crop yield
Medical research	drugs	disease status
Industry	processes	yield of a chemical product
	machines/operators	output of product
Materials science	additives/processes	breaking strain
Psychology	learning methods	attainment of a skill

The treatments are applied to **experimental units** (*e.g.* patients in a hospital, plots of land). A common characteristic of experiments in research is that results vary in different replications of the experiment, hence the need for a statistical approach to the analysis of the results. The methods normally take either (or both) of the following forms: *(i)* estimating the effects of treatments, *(ii)* investigating hypotheses regarding their effects. In a frequentist approach *(i)* would normally be accompanied either by standard errors of estimates or confidence intervals and *(ii)* accomplished by significance tests.

2.1.1 Principles of experimental design

Independent from potential treatment effects, there will always be natural variation across experimental units. This noise is usually called **experimental error**, even though the variation is not due to error as such. The main objective in **experimental design** is to reduce the effect of experimental error, thereby increasing the **precision** of the results. However, even if the results are of high precision, there may still be systematic errors (**bias**) present. For example, in comparing two varieties of a crop, one variety might be planted in a more fertile area than the other. Bias can be avoided by allocating the treatments to the units at random, a process known as **randomisation**. Even if precautions are taken against known **environmental** or **nuisance** variables, we still cannot be sure that *everything* has been taken into account; randomisation guards against unforeseen sources of bias.

Methods of increasing precision

1. Increase the number of **replications** for each treatment (i.e. the number of experimental units receiving each treatment).
2. Remove some of the natural variation by eliminating or adjusting for any known or suspected environmental or nuisance variables. Techniques:

Blocking: here we arrange the experimental material so that comparisons are made within more homogeneous blocks as defined by environmental factors. These blocks may be **complete** (all treatments are applied within each block) or **incomplete** (not all treatments are applied within each block).

Adjustment: when an environmental variable is quantitative, an alternative to blocking is to adjust for it using regression techniques. The general idea is to try to allow for all environmental variables envisaged and then randomise to avoid bias from other, unknown sources (**restricted randomisation**).

3. If several variables are of interest, we can reduce the number of experimental runs required to achieve a specified level of precision by changing more than one variable at a time (as opposed to ‘one-at-a-time’ experiments). This is the basis of **factorial experiments** (see §2.7).

2.1.2 Planning of experiments

An experiment should be planned in order to discover natural relationships of interest. A general strategy is as follows:

Pre-experimental planning:

- Formulate and state the problem in clear and unambiguous terms
- Select the response variable (what to measure) and treatments, and determine the levels or ranges of the latter
- Identify any other potential nuisance variables

The experiment itself:

- Choice of design: considerations of sample size (= number of replicates), run order (randomisation), how to handle nuisance variables (block, adjust, hold constant or randomise, depending on objectives of the experiment and whether the nuisances can be measured or controlled)
- Perform the experiment: monitor carefully — possibly conduct a few trial runs to check for unexpected problems and get some practice

- Carry out appropriate statistical analysis (for objective conclusions): apply a model, carry out hypothesis tests, compute confidence intervals *etc.*, carry out model diagnostics
- Draw conclusions, make recommendations, including need for further experimentation *etc.*

In the present course we focus mainly on experiments investigating the effect of categorical factors upon a response, and review the associated standard analysis of variance methods. However, if the treatment variables are quantitative then **regression models** may be used instead (see §2.12).

Example 1: continued

- *Response variable* — folate concentration. Why folate concentration? Are there alternative response variables?
- *Treatments* — three methods of ventilation. Why these three treatments in particular?
- *Experimental design* — How are treatments allocated? Is there some form of randomisation? If treatments are allocated randomly, can we constrain a randomisation to have an equal number of patients per group (**balanced design**)? Is an unbalanced design a problem for the analysis?
- *Conduct of experiment* — we will not cover this in the present course, but this does not mean that it is not important. For example, precise instructions might be necessary for when concentration is measured.
- *Analysis* — standard method is the one-way ANOVA (§2.3.1).

The moral: even ‘simple’ experiments may involve complex subject-matter knowledge. It may be dangerous for the statistician to assume that all that is being produced is a set of numbers, the details being irrelevant. Teamwork may be necessary; in an ideal situation the statistician will be involved from the outset.

2.2 Simple comparative experiments

2.2.1 Two independent samples

Suppose there are two treatment groups to be compared and a random sample of size n_i is available from the i th group ($i = 1, 2$). Let y_{ij} be the j th observation in the i th treatment group, where $j = 1, \dots, n_i$, $i = 1, 2$. It is agreed that the substantive question of interest can be answered by considering the data to be observed values of independent random variables (r.v.’s) Y_{ij} , drawn independently from distributions with means μ_1 (for group 1) and μ_2 (for group 2), and by trying

to decide whether or not the underlying group means μ_1 and μ_2 are equal. This can be done by testing the hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ (or a one-sided alternative).

Statistical model: to proceed with this problem, we can introduce a probability model. Specifically, suppose that the Y s are not only independent but also normally distributed: $Y_{1j} \sim N(\mu_1, \sigma^2)$, $Y_{2j} \sim N(\mu_2, \sigma^2)$. Note that in this model, the **within-group variance** is the same (σ^2) in each group. The model can also be written as

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

for $j = 1, \dots, n_i$, $i = 1, 2$, where the ϵ_{ij} are independent $N(0, \sigma^2)$ r.v.'s.

Analysis: this is a standard problem, for which we can use the **two-sample t -test**. We review the procedure by way of revision.

Let the two group means be $\bar{Y}_1$ and $\bar{Y}_2$. Then

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

The common variance σ^2 can be estimated by the **pooled sample variance**

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

The estimated standard error (s.e.) of $\bar{y}_1 - \bar{y}_2$ is $s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Under H_0 , the reference distribution for

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (1)$$

is a $t(\nu)$ distribution with degrees of freedom $\nu = n_1 + n_2 - 2$. Thus to perform the test, compute t and refer to percentage points of the $t(n_1 + n_2 - 2)$ distribution.

Suppose now that instead of testing a hypothesis about the difference in means, it is of interest to estimate this difference (see Section 2.1). This can be done by constructing a **confidence interval** for the difference $\mu_1 - \mu_2$. Given that $\bar{y}_1 - \bar{y}_2$ is an unbiased estimator for this difference, a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}(n_1 + n_2 - 2)s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (2)$$

Here $t_p(\nu)$ is the upper $100p\%$ point of the $t(\nu)$ distribution satisfying $P(T > t_p(\nu)) = p$, where $T \sim t(\nu)$.

Example 4 In an experiment to compare the effectiveness of two drugs, A_1 and A_2 , each drug was administered to a different group of patients and the number of days to recovery were recorded as follows:

A_1	79	84	108	114	120	103	122	120		
A_2	91	103	90	113	108	87	100	80	99	54

Recall that—in general—the sample variance is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2,$$

which can be computed as $s^2 = 1/(n-1) \sum_{i=1}^n y_i^2 - n/(n-1) \bar{y}^2$. For the above data we have $n_1 = 8$, $\sum_i y_{1i} = 850$, $\sum_i y_{1i}^2 = 92250$, $n_2 = 10$, $\sum_i y_{i2} = 925$, $\sum_i y_{i2}^2 = 88109$. Hence we have $\bar{y}_1 = 106.25$, $s_1^2 = 276.79$, $\bar{y}_2 = 92.50$, $s_2^2 = 282.94$. **Exercise:** verify this! Note that the two sample s.d.'s, 16.64 and 16.82, are close, consistent with the assumption of equal s.d.'s (a formal F test may be carried out to verify this if necessary).

The pooled sample variance here is

$$s^2 = \frac{1937.50 + 2546.50}{7 + 9} = \frac{4484.00}{16} = 280.25.$$

The observed difference is $\bar{y}_1 - \bar{y}_2 = 106.25 - 92.50 = 13.75$, with associated standard error $\sqrt{280.25(\frac{1}{8} + \frac{1}{10})} = 7.94$. To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, the test statistic is $t = 13.75/7.94 = 1.73$ (“the observed difference in means is 1.73 times its standard error.”). The 2.5% point of $t(16)$ is 2.12, so the result is not significant at the 5% level — there is no evidence of a difference in the effectiveness of the two drugs. Note the wording of the conclusion: it would be wrong to conclude that “the two drugs are equivalent”.

Note and warning. This result surprised medics, who thought that A_2 would be better. Investigation revealed that the treatment allocation was not random since A_2 , being more unpleasant, was given mainly to the worse cases. Non-random treatment allocation can lead to bias.

Using R: Example 4. Using the function `t.test` for a two-sample t -test:

```
> t.test(A1, A2, var.equal= TRUE)

      Two Sample t-test

data:  A1 and A2
t = 1.7316, df = 16, p-value = 0.1026
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.083737 30.583737
sample estimates:
mean of x mean of y
  106.25    92.50
```

Design considerations for the two-sample test. The example above highlights the need for randomisation even in simple experiments — here, there was an unrecorded nuisance variable (‘severity of illness’) that caused bias in the results. Randomisation could have been used to overcome this, for example by tossing a coin for each patient.

Another issue to consider is the number of observations to take. This may be limited by costs and/or other practicalities (for example, if a disease is very rare then it may take a long time to accumulate a large sample of patients). Otherwise statistical theory based on length of confidence intervals, power of tests, minimum cost *etc.* may be applied. For example:

Specified length of confidence interval. Suppose that we require that the length of the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is at most L . Suppose that σ is known and that equal sample sizes $n_1 = n_2 = n$ are to be taken. Then from equation (2) applied with $t_{\alpha/2}(n_1 + n_2 - 2)$ replaced by $z_{\alpha/2}$ we have $2z_{\alpha/2}\sigma\sqrt{2/n} \leq L$, from which we obtain

$$\boxed{n \geq \frac{8z_{\alpha/2}^2\sigma^2}{L^2}} \quad (3)$$

In practice σ will usually be unknown, in which case some initial estimate must be made (perhaps using a pilot study, or by looking at similar experiments from the literature). Furthermore, the t -distribution should then be used, which also depends on n (degrees of freedom will be $2n - 2$ in this case). Provided n is not too small, use of (3) should give a reasonable answer which can be checked by replacing $z_{\alpha/2}$ with the corresponding point of the $t(2n - 2)$ distribution; if the resulting confidence interval is too long, adjust n upwards and recheck.

Specified power of test. Suppose we are to perform a one-sided test of $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ at level α , and that it is required to detect an alternative of $\mu_1 - \mu_2 = \delta \neq 0$ with probability $1 - \beta$. This is the **power** of the test at $\mu_1 - \mu_2 = \delta$. Since $\delta \neq 0$, $\beta = \text{P}(\text{Type II error}) = \text{P}(\text{Accept } H_0 \text{ when it is false})$.

Again, suppose that σ is known and $n_1 = n_2 = n$. Then from equation (1) based on the normal distribution, we require $\text{P}(T > z_\alpha | \mu_1 - \mu_2 = \delta) \geq 1 - \beta$. But the distribution of T under H_1 is $N(\delta\sqrt{n/2}/\sigma, 1)$. Hence we have $\text{P}(T > z_\alpha) = \text{P}(Z > z_\alpha - \delta\sqrt{n/2}/\sigma)$ where $Z \sim N(0, 1)$. We therefore require $z_\alpha - \delta\sqrt{n/2}/\sigma \leq z_{1-\beta} = -z_\beta$. Rearranging, we obtain

$$n \geq \frac{2\sigma^2(z_\alpha + z_\beta)^2}{\delta^2}$$

If σ is unknown we may proceed as for confidence intervals. If the test is two-sided, then the above can be used as an approximation, on replacing α by $\alpha/2$.

You might reasonably ask ‘what if we do not want to take equal numbers of observations in each group?’ The answer is that having equal numbers is, in a certain sense, optimal. To see this we can proceed as follows: suppose that you have to allocate N observations to two groups. With n_1 observations in the first group and $n_2 = N - n_1$ in the second, we have $\text{Var}(\bar{Y}_1 - \bar{Y}_2) = \sigma^2(n_1^{-1} + n_2^{-1})$. If we can minimise this then we are, in a sense, obtaining the most precise information possible about the difference in underlying means. The minimisation must be done, however, subject to the constraint $n_1 + n_2 = N$. We therefore introduce a Lagrange multiplier and minimise the function

$$f(n_1, n_2, \lambda) = \sigma^2(n_1^{-1} + n_2^{-1}) + \lambda(n_1 + n_2 - N).$$

Setting the partial derivatives w.r.t. n_1 and n_2 to zero gives the equations

$$-\sigma^2 n_1^{-2} + \lambda = 0, \quad -\sigma^2 n_2^{-2} + \lambda = 0$$

from which we deduce that $n_1 = n_2$, as expected.

Similar principles apply for the more complex experimental and survey designs considered later and we do not give details. Software is available to carry out such computations.

Some practical issues

- Although construction of a t -test itself does not require knowledge of σ , the formula for the sample size does. In practice we use some previous estimate, but this itself will be subject to error and the formula does not take this into account.

- There is usually no agreed standard for choosing δ (in the context of clinical trials, the experimenters will usually choose it to represent a “clinically relevant difference”).
- The levels α and β are themselves arbitrary.
- Even if you design an experiment perfectly, things can still go wrong — patients may withdraw (“drop-outs”) from clinical trials, or experimental runs may fail so that you do not get all of the observations you hoped for.
- It may be required that the results be robust to a number of analyses. This generally requires a larger sample.
- There is no explicit mention of costs of sampling — this may place an upper limit on $N = n_1 + n_2$ and hence affect the choice of α and β .
- The sample size calculations, as well as the results of the tests, rely on the assumed statistical model. Randomisation can be used to guard against problems of bias (and, often, to ensure that observations can be regarded as independent or nearly so). However, the experimenter can do little to guarantee that the underlying distributions are normal. In Section 2.3.2 we will discuss methods for checking normality, and give some suggestions for what to do if it does not hold.

Example 4 (continued): Suppose we wish the length of the 95% confidence interval for $\mu_1 - \mu_2$ to be 20 days and have a preliminary estimate of 16 for σ . Then a normal approximation gives $n \geq 19.67$ ($n \geq 21$ using the t distribution).

Alternatively, suppose that we wish to detect a clinically relevant difference of 15 days with probability 0.9. For a two-sided test, a normal approximation gives $n \geq 23.91$ ($n \geq 24$ using the t distribution).

2.2.2 Matched pairs

Suppose now that there are two treatments to be compared and that, instead of obtaining two independent samples, the observations are in the form of n **matched pairs**. Let y_{ij} be the response to treatment i in pair j ($j = 1, \dots, n, i = 1, 2$). The main question of interest is to determine whether or not the treatments have different effects and, as before, it is agreed that this can be answered by testing hypotheses about, or estimating, the difference in underlying means.

Statistical model: the paired structure of the observations means that we need a different model to that of the previous section — we need to allow for the fact that there may be systematic differences in the response for different pairs. One way to view this is to consider that the responses are no longer independent within pairs. Another, which is in fact equivalent, is explicitly to

model the ‘pair effect’: we consider the data y_{ij} to be realised values of independent random pairs (Y_{1j}, Y_{2j}) with $Y_{ij} \sim N(\mu_i + \beta_j, \sigma_i^2)$. Here μ_i is the mean effect of treatment i and β_j is the effect of the j th pair, with $\sum_j \beta_j = 0$. The model can also be written as

$$Y_{ij} = \mu_i + \beta_j + \epsilon_{ij}.$$

This is an **additive model**.

Analysis: the analysis of this model is simplified by noting that $\{Y_{1j} - Y_{2j} : j = 1, \dots, n\}$ are independent normal random variables, each with mean $\mu_1 - \mu_2$. We can therefore eliminate the **nuisance parameters** $\{\beta_j\}$ by working with the **differences** $\{d_j = y_{1j} - y_{2j}\}$. If we assume that the variances of the differences are all equal (to σ_d^2 , say), then we can test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ (or one-sided alternative) by applying a **one-sample t -test** to the differences. We compute

$$t = \frac{\bar{d}}{s_d \sqrt{1/n}}$$

for which the reference distribution is $t(n-1)$ under H_0 . To perform the test, compute t and refer to percentage points of the $t(n-1)$ distribution. Note that, unlike the case of two independent samples, no assumptions are required about the individual variances σ_1^2 and σ_2^2 .

If we are interested in estimation rather than testing, a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by $\bar{d} \pm t_{\alpha/2}(n-1)s_d\sqrt{1/n}$.

Example 5: to investigate whether one week of special training has a beneficial effect on athletes’ performance, five athletes’ times were recorded before and after the training. The results were as follows:

	Athlete				
	1	2	3	4	5
Before	10.3	11.6	9.4	10.6	9.7
After	10.1	11.6	9.5	10.0	9.4
Difference	0.2	0.0	-0.1	0.6	0.3

For these data we have $n = 5$, $\sum_{j=1}^n d_j = 1.0$ and $\sum_{j=1}^n d_j^2 = 0.5$. Hence we find $\bar{d} = 0.2$, and $s_d^2 = 0.075$. To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ (one-sided alternative implies that we are investigating for a positive effect of the special training, which in turn implies that we are not even prepared to consider the possibility that it is detrimental — is this reasonable?), the test statistic is

$$t = \frac{0.2}{\sqrt{0.075/5}} = 1.633$$

The upper 5% point of $t(4)$ is 2.13 so we accept H_0 at the 5% level — these data provide no evidence that the course has been beneficial.

Using R: Example 5. Using the function `t.test` for matched pairs:

```
> t.test(before, after, alternative = "greater", paired = TRUE)

Paired t-test

data:  before and after
t = 1.633, df = 4, p-value = 0.0889
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.06109686      Inf
sample estimates:
mean of the differences
          0.2
```

Question: what would change in the analysis if the alternative hypothesis was $H_1 : \mu_1 \neq \mu_2$?

Analysis: to pair or not to pair? In Example 5, an alternative experimental strategy would have been to take a sample of ten athletes and to record their times after giving the training to five of them (randomly selected, of course). In this case the results could be analysed using the two-sample t -test of the previous section. Would this be better or worse than the paired design considered above?

The following summarises the reasoning in Rice (1995). For a paired design with $D_j = Y_{1j} - Y_{2j}$, $j = 1, \dots, n$, we have $Y_{1j} \not\perp Y_{2j}$ and

$$E(\bar{D}) = \mu_1 - \mu_2 \quad \text{and} \quad \text{Var}(\bar{D}) = \frac{1}{n}(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2),$$

where ρ denotes the correlation of members of a pair. For the alternative design, take ten athletes, give training only to n of them and use a two-sample t -test. Here we have

$$E(\bar{Y}_1 - \bar{Y}_2) = \mu_1 - \mu_2 \quad \text{and} \quad \text{Var}(\bar{Y}_1 - \bar{Y}_2) = \frac{1}{n}(\sigma_1^2 + \sigma_2^2).$$

Pairing seems more efficient if measurements are positively correlated within paired experimental units (because then $-2\rho\sigma_1\sigma_2 < 0$).

On the other hand ... the inference for the paired design is based on $t(n-1)$, whereas the inference unpaired design based on $t(2n-2)$. The first distribution has heavier tails than the second.

2.3 Completely randomised design

2.3.1 One-way analysis of variance

We now move on to the case where there are $r \geq 2$ treatment groups to be compared. Suppose there are N experimental units in total and they are allocated to the treatments at random. Let y_{ij} be the j th observation in the i th treatment group, where $j = 1, \dots, n_i$, $i = 1, \dots, r$. We must have $N = \sum_i n_i$. We wish to decide whether or not the group means $\mu_1, \mu_2, \dots, \mu_r$ are equal *i.e.* to test the hypothesis $H_0 : \mu_1 = \mu_2 = \dots = \mu_r$ against $H_1 : \text{at least one pair of group means is unequal}$.

Statistical model: in this situation we might start by considering the data as observed values of random variables Y_{ij} satisfying the linear model

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad (i = 1, \dots, r; j = 1, \dots, n_i), \quad (4)$$

where the ϵ_{ij} are independent $N(0, \sigma^2)$ r.v.'s. Note that we are assuming a common within-groups variance. Equivalently, if we define $\alpha_i = \mu_i - \mu$, where $\mu = \sum_i n_i \mu_i / N$ is the weighted average of the μ_i , we can rewrite the model in additive form as

$$\boxed{Y_{ij} = \mu + \alpha_i + \epsilon_{ij}}. \quad (5)$$

Here μ is intended to represent the ‘overall mean’, and α_i is an ‘additional’ effect of the i th treatment. This form of the model is often convenient for interpretation. However, there is a problem, since the model now has $r + 1$ parameters to describe just r individual group means. Clearly, these parameters are not all uniquely defined — the model is not **identifiable**. To overcome this problem we must impose a constraint on the parameters. The constraint that ensures the intended interpretation is the **sum-to-zero parameterisation** defined by $\sum_i n_i \alpha_i = 0$ (other constraints are also possible — see later). In terms of the α_i , the null hypothesis is $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_r = 0$.

Analysis: partition of the total sum of squares. A very convenient way to study this model, at least for the purpose of testing H_0 , is to write each observation as a sum of deviations:

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}), \quad (6)$$

where $\bar{y}_{1.}, \dots, \bar{y}_{r.}$ denote the individual group means and $\bar{y}_{..} = \sum_{i=1}^r n_i \bar{y}_{i.} / \sum_{i=1}^r n_i$ is the overall mean. Notice the notational convention here: a dot is used to indicate a suffix over which averaging has taken place.

The three terms on the right-hand side of (6) may be thought of as the ‘sample’ analogues of μ , α_i and ϵ_{ij} respectively — indeed, we will see below that the first two are the least-squares

estimates of μ and α_i . We now consider the **total variation** of the data about the overall mean $\bar{y}_{..}$:

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} \{(\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})\}^2$$

Expanding out the bracket on the right-hand side here, it can be shown (see exercises) that the cross-product term vanishes to yield

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^r n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 .$$

The first term on the right-hand side in some sense represents variation among the different group means: it is called the **between-groups sum of squares** (s.s) and denoted below by S_G :

$$S_G = \sum_{i=1}^r n_i (\bar{y}_{i.} - \bar{y}_{..})^2 .$$

The second term is a sum of contributions representing the variation of observations within each individual group about the group means: it is called the **residual** or **error** or **within-groups** sum of squares (s.s.):

$$S_E = \sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^r (n_i - 1) s_i^2$$

Denoting the total variation by S_T therefore, we have

$$\begin{aligned} S_T &= S_G + S_E \\ \text{or Total s.s.} &= \text{Between-groups s.s.} + \text{Within-groups (error, residual) s.s.} \end{aligned}$$

Furthermore, we have

$$E(S_E) = (N - r)\sigma^2, \quad E(S_G) = (r - 1)\sigma^2 + \sum_{i=1}^r n_i \alpha_i^2, \quad (7)$$

the final sum being zero if and only if H_0 holds.

Notice that both of the expectations in (7) depend on σ^2 , but that only $E(S_G)$ depends on the $\{\alpha_i\}$. Thus it seems intuitively reasonable to assess the plausibility of $H_0 : \alpha_1 = \dots = \alpha_r = 0$ by

considering the value of S_G relative to that of S_E . This can be done by computing the **variance ratio**, or **mean-square ratio**

$$F = \frac{S_G/(r-1)}{S_E/(N-r)} , \quad (8)$$

which should be approximately 1 under H_0 but larger than 1 otherwise.

The decomposition of the total variation is an algebraic result that does not depend on the assumed probability model. The results for $E(S_E)$ and $E(S_G)$ depend on the assumptions of independence and constant variance. To formulate a test of H_0 , however, we need to be able to say something about the distributions of S_E and S_G . To do this, we need to make use of the normality assumption and appeal to the following remarkable result:

Theorem 2.1 Cochran's Theorem (simplified version). *Let $\mathbf{Z} = (Z_1 \dots Z_\nu)'$ be a vector of independent standard normal random variables, and suppose we write $\sum_{i=1}^\nu Z_i^2 = \sum_{r=1}^m \mathbf{Z}' \mathbf{A}_r \mathbf{Z}$ for some set of positive semidefinite matrices $\mathbf{A}_1, \dots, \mathbf{A}_m$. Define $Q_r = \mathbf{Z}' \mathbf{A}_r \mathbf{Z}$, and denote the rank of $\mathbf{A}_r$ by ν_r . Then $Q_1, \dots, Q_r$ are independent chi-squared random variables, with $\nu_1, \dots, \nu_m$ degrees of freedom respectively, if and only if $\sum_r \nu_r = \nu$.*

The proof requires a reasonable grasp of linear algebra and matrix manipulation, and will not be given here. Notice, however, that Q_r can also be written as $\sum_{i,j} A_r(i,j) Z_i Z_j$ — it is a *quadratic form* in the Z s. Requiring the $\{\mathbf{A}_r\}$ to be positive semidefinite is equivalent to requiring that the $\{Q_r\}$ are never negative. $\square$

We can apply this theorem to our current problem by defining $Z_{ij} = (Y_{ij} - \mu)/\sigma$ so that, under H_0 , the $\{Z_{ij}\}$ are iid standard normal random variables. In this case, we have

$$\begin{aligned} \sum_{i=1}^r \sum_{j=1}^{n_i} Z_{ij}^2 &= \sum_{i=1}^r \sum_{j=1}^{n_i} \left(\frac{Y_{ij} - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^r \sum_{j=1}^{n_i} [(Y_{ij} - \bar{Y}_{..}) + (\bar{Y}_{..} - \mu)]^2 \\ &= \frac{S_T}{\sigma^2} + \frac{N(\bar{Y}_{..} - \mu)^2}{\sigma^2} \\ &\quad \text{(cross-product vanishes as usual)} \\ &= \frac{S_G}{\sigma^2} + \frac{S_E}{\sigma^2} + N\bar{Z}_{..}^2 . \end{aligned}$$

Now consider the term $N\bar{Z}_{..}^2 = N^{-1} \left(\sum_i \sum_j Z_{ij} \right)^2$. This is a quadratic form in the vector $\mathbf{Z} = (Z_{11} \ Z_{12} \ \dots \ Z_{rn_r})'$ — it can be written as $\mathbf{Z}' \mathbf{A}_1 \mathbf{Z}$, where $\mathbf{A}_1$ is an $N \times N$ matrix with elements all equal to N^{-1} . Since all the rows of this matrix are equal, its rank is 1. Similarly (but more tediously), we can show that the terms S_G/σ^2 and S_E/σ^2 can be written as quadratic forms with ranks $r-1$ and $N-r$ respectively. Cochran's theorem now tells us that $S_E/\sigma^2 \sim \chi^2(N-r)$,

$S_G/\sigma^2 \sim \chi^2(r-1)$ and that S_E, S_G are *independent*. Therefore, under H_0 , the mean-square ratio (8) has the $F(r-1, N-r)$ distribution.

Analysis: the ANOVA table. The *Analysis of variance (ANOVA) table* includes the sum of squares, the mean square (m.s.), the mean-square ratio (m.s.r.), and the F test:

Source of variation	Sum of squares	d.f.	m.s.	m.s.r.	Expected m.s.
Between groups	$S_G = \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2$	$r-1$	$\frac{S_G}{r-1}$	$F = \frac{S_G/(r-1)}{S_E/(N-r)}$	$\sigma^2 + \frac{\sum_i n_i \alpha_i^2}{r-1}$
Within groups	$S_E = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$	$N-r$	$\frac{S_E}{N-r}$		σ^2
Total	$S_T = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$	$N-1$			

The expected m.s. column is not included in practice.

To perform the test, compute the m.s.r. F and reject H_0 at level α if $F \geq F_\alpha(r-1, N-r)$, the upper $100\alpha\%$ point of the F distribution. The procedure can also be derived as a likelihood ratio test (see course STAT0008) and hence enjoys certain optimality properties — at least, if the underlying model (5) holds.

You are asked to demonstrate the following useful results in the exercises:

- The test procedure is equivalent to the two-sample t -test when $r = 2$.
- For hand calculation, S_G and S_T can be calculated as

$$S_G = \sum_{i=1}^r \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N} \quad \text{and} \quad S_T = \sum_{i=1}^r \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}, \quad \text{say}, \quad (9)$$

where $y_{..}$, is the grand total and $y_{i.}$, is the i th group total. S_E can then be calculated by subtraction. The quantity $y_{..}^2/N$ is sometimes called a *correction factor*.

- **Coding** the data $y' = ay + b$ does not affect the m.s.r. This can be used to reduce large numbers to something manageable (both in hand calculations and computer implementations).

Analysis: least-squares estimation of μ and α_i . The development so far has focused on testing for differences between the groups. If, instead, we want to estimate the parameters in the model (5) then we could proceed by minimising the sum of squared deviations

$$R = \sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2$$

w.r.t. μ and the α_i to give

$$\frac{\partial R}{\partial \mu} = -2 \sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i) \quad \text{and} \quad \frac{\partial R}{\partial \alpha_i} = -2 \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i).$$

Set these $r + 1$ derivatives to zero to give the **normal equations**

$$\begin{aligned} N\bar{y}_{..} - N\hat{\mu} - \sum_{i=1}^r n_i \hat{\alpha}_i &= 0 \\ n_i \bar{y}_{i.} - n_i \hat{\mu} - n_i \hat{\alpha}_i &= 0 \end{aligned}$$

for the least-squares estimators $\hat{\mu}$ and $\hat{\alpha}_i$. There is no unique solution to these equations. Remember, however, that we are imposing the constraint $\sum_i n_i \alpha_i = 0$ to ensure identifiability. Substituting this constraint into the first equation enables us to solve the entire system:

$$\hat{\mu} = \bar{y}_{..}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$$

These are also the **maximum likelihood** estimates of the model parameters. We have already seen that the within-group variance σ^2 can be estimated as $s^2 = S_E/(N - r)$.

Confidence interval for the difference between two group means. Suppose we wish to construct a confidence interval for the difference $\mu_i - \mu_j$ (or equivalently for $\alpha_i - \alpha_j$). An unbiased estimator for this quantity is $\bar{y}_{i.} - \bar{y}_{j.}$, with estimated s.e. $s\sqrt{(n_i^{-1} + n_j^{-1})}$. Therefore a $100(1 - \alpha)\%$ confidence interval for $\mu_i - \mu_j$ is

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm t_{\alpha/2}(N - r) s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Confidence interval for the error variance. Since $S_E/\sigma^2 \sim \chi^2(N - r)$ a $100(1 - \alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{S_E}{\chi_{\alpha/2}^2(N - r)}, \frac{S_E}{\chi_{1-\alpha/2}^2(N - r)} \right)$$

Example 6 (Clarke & Kempson) In an experiment to compare four different fuel mixtures, mixtures were randomly assigned to each of 25 motor vehicles and the fuel consumption of each (in km/L) was measured under controlled conditions. The results were as follows:

A	13.31	14.04	13.68	13.75	13.12	14.11	13.96
B	14.28	14.47	14.03	15.62	15.10		
C	15.04	14.77	15.13	15.45	14.98	15.51	
D	14.66	13.93	15.05	14.21	14.42	14.30	14.25

The calculations are most easily presented in tabular form:

Treatments	n_i	$y_{i.}$	$\sum_j y_{ij}^2$	$y_{i.}^2/n_i$
A	7	95.97	1316.5907	1315.7487
B	5	73.50	1082.1346	1080.4500
C	6	90.88	1376.9344	1376.5291
D	7	100.82	1452.8760	1452.0961
$N = 25 \quad y_{..} = 361.17 \quad S = 5228.5357 \quad 5224.8239$				

Total s.s. = $S - y_{..}^2/N = 5228.5357 - (361.17)^2/25 = 5228.5357 - 5217.7508 = 10.7849$.

Between-groups s.s. = $\sum_{i=1}^r y_{i.}^2/n_i - y_{..}^2/N = 5224.8239 - 5217.7508 = 7.0731$.

The ANOVA table is therefore as follows:

Source of variation	s.s.	d.f.	m.s.	m.s.r.
Between groups	7.0731	3	2.3577	13.339
Within groups	3.7118	21	0.1768	
Total	10.7849	24		

The 0.1% point of F with 3 and 21 degrees of freedom is 7.94. Since the observed value of F is greater than this, the result is significant at the 0.1% level. There is very strong evidence that mean fuel consumption differs under the four different mixtures. The four mean values are 13.71, 14.70, 15.15 and 14.40, indicating that treatment C is the best (km/L).

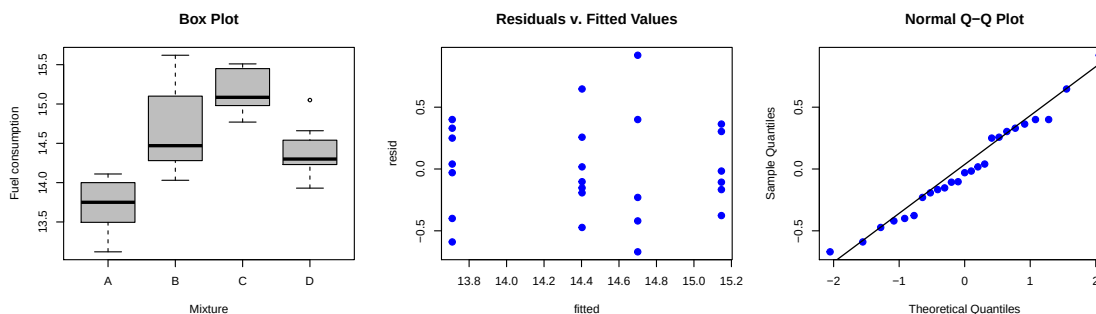
A 99% confidence interval for the difference between treatment C and A is $(15.14 - 13.71) \pm 2.831 \times 0.4204 \sqrt{\frac{1}{6} + \frac{1}{7}}$, giving the interval (0.774, 2.099).

Using R: Example 6. Using the function `aov` for a one-way ANOVA:

```
> fit <- aov(fuel.con ~ mixture, fuel.data)
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mixture	3	7.0731	2.3577	13.339	4.288e-05 ***
Residuals	21	3.7119	0.1768		

Some diagnostic plots (see also §2.3.2 and §2.13):



Analysis: identifiability. We have already seen that the additive model (5) contains more parameters than we can estimate, which means we must impose an arbitrary constraint in order to obtain a unique solution. The constraint used so far is $\sum_i n_i \alpha_i = 0$. This has the effect that the $\{\alpha_i\}$ can be interpreted as deviations about an overall mean. Another possibility would be to set $\alpha_1 = 0$. In this case, the parameter μ would be the expected response under treatment 1, and the remaining α s would represent differences from this first treatment. Such a constraint would be appropriate, for example, if the first treatment represented some kind of baseline against which the other treatments were to be compared. The result is sometimes called a **corner-point parameterisation** of the model.

Using R: parametrisation The corner-point parameterisation is the default for fitting an ANOVA using `aov`. It can be set explicitly by

```
> options(contrasts=c("contr.treatment", "contr.treatment"))
```

The argument `contrasts` is defined by a vector of length two. The first entry is the function to be used with unordered factors and the second entry is the function to be used with ordered factors. Given this parameterisation, the coefficients for Example 6 are

```
> round(coef(summary.lm(fit)),4)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.7100     0.1589  86.2779   0.0000
mixtureB      0.9900     0.2462   4.0215   0.0006
mixtureC      1.4367     0.2339   6.1422   0.0000
mixtureD      0.6929     0.2247   3.0831   0.0056
```

Exercise: use this output to compute a 99% confidence interval for the mean difference between treatments C and A. Compare the answer with that obtained earlier.

The sum-to-zero constraint $\sum_i \alpha_i = 0$ can be set in R by

```
> options(contrasts=c("contr.sum", "contr.sum"))
```

Note that this only implies $\sum_i n_i \alpha_i = 0$ when the design is balanced in the sense that $n_i = n \forall i$.

Analysis: identifiability continued. Whatever constraints we impose, we will always have the relation $\mu_i = \mu + \alpha_i$ ($i = 1, \dots, r$). Suppose now that we wish to estimate the value of a linear combination $\sum_i \lambda_i \alpha_i$, where $\sum_i \lambda_i = 0$ (or example, if $\lambda_1 = 1$ and $\lambda_2 = -1$ then we get the difference $\alpha_1 - \alpha_2$). We have $\sum_i \lambda_i \alpha_i = \sum_i \lambda_i (\mu_i - \mu) = \sum_i \lambda_i \mu_i$, which does not depend on the constraints. Linear combinations of this form are called **contrasts**.

2.3.2 Checking model assumptions

The assumptions of the additive model (5) are firstly that the observations are from r independent normal populations, and secondly that the r population variances are equal. Some comments on these:

1. The F test is fairly **robust** to departures from normality if the sample sizes are sufficiently large (central limit effect), but not for very small samples. To check, plot the **residuals** $r_{ij} = y_{ij} - \bar{y}_{i\cdot}$ as Q-Q plot (see earlier plots for Example 6). Remedies: possibly **transform** the data or use a different distributional model (see §2.13) or use a **distribution-free** analysis (nonparametric method).
2. Independence, and homogeneity within groups, can be controlled to some extent using randomisation. However, even in this case it is useful to plot residuals against other variables in case of unexpected problems. For example, if an experiment is carried out by taking measurements over the course of a day then it may be useful to plot the residuals in time sequence — it may turn out, for instance, that observations taken in the morning are generally less than those taken in the afternoon. If such an effect is identifiable, a more sophisticated analysis may be required — for example, using a randomised blocks layout (see §2.4).
3. Modest failure of the ‘constant variance’ assumption is not too serious if the group sizes are similar. The assumption can be checked by inspection of the within-groups sample variances $s_1^2, \dots, s_r^2$, or using residual plots. If necessary, the hypothesis $H_0 : \sigma_1^2 = \dots = \sigma_r^2$ can be tested formally using **Bartlett’s test** (be careful though—this test is very sensitive to departures from normality).

If the variances appear non-constant, a possible remedy is to transform the data using a **variance-stabilising** transformation (see §2.13). Such a transformation may also result in better normality as well. For example, if the model is multiplicative: $Y_{ij} = \mu_i \epsilon_{ij}$ (so that s.d. $\propto$ group mean) then transform to an additive model by taking $\log Y_{ij}$.

2.3.3 Design considerations

We saw in §2.2.1 that in the two-sample case, it is optimal to allocate half of the experimental units to each treatment group. With more than two groups, it seems intuitively reasonable that the same principle applies and that we should, if possible, allocate equal numbers of observations to each group to obtain a balanced design. Indeed, if this is done then, for a given sample size N , we maximise the power of the F -test to detect departures from H_0 . Other advantages include some robustness to departures from the assumption of constant variance; ease of interpretation of the parameters in (5) (for example, the constraint $\sum_i n_i \alpha_i = 0$ becomes $\sum_i \alpha_i = 0$, and the α s are simply deviations from an overall mean); and the fact that all of the effects are estimated with equal precision.

However, suppose that the purpose of the experiment is to compare each of $r - 1$ experimental treatments to a single control. In this case, rather than maximising the power of the overall F -test, we may wish to estimate the treatment-control contrasts as precisely as possible. It can be shown that if group 1 is the control group then the variance of a typical contrast $\bar{y}_i - \bar{y}_1$ ($i = 2, \dots, r$) is minimised by allocating $N/(1 + \sqrt{r - 1})$ units to the control (notice that this *does* give a balanced design in the two-group case, but not otherwise). Heuristically, what is happening here is that the control group has a special place in the context of the experiment, and hence it is worth spending more resources in estimating the control mean precisely. This example shows that with more than two groups, the optimal allocation depends on the context.

2.4 Randomised block design

As in the paired t -test, we may encounter situations in which subsets of the experimental units are similar to each other but different from other subsets. Such similar subsets are called **blocks**. The presence of blocks can be exploited. When the experimental material is heterogeneous, we try to group experimental units into homogeneous blocks using some ‘environmental’ factor. The treatments are then applied to the units within each block at random. More generally we might use more than one blocking, giving a higher-order model.

Example 7: A class of 20 schoolchildren was divided into groups of four children each, according to ability. Within each group the children were randomly assigned to four tutors A, B, C and D. The end-of-term marks for the class were as follows:

		Group				
		1	2	3	4	5
Tutor	A	62	62	52	54	50
	B	56	52	57	50	41
	C	50	54	47	41	41
	D	58	51	51	47	48

In this experiment, the ability groups form the blocks and the tutors are the ‘treatments’, giving a two-dimensional structure. This is a two-way layout: a **randomised complete block design** (‘complete’ because each block contains each treatment). We will see below that such a design can be analysed using two-way analysis of variance.

The idea of blocking originated from field trials in agriculture, in which a block was a set of plots of presumed similar fertility. Using each treatment within each block eliminates a source of variation. The technique is now widely used in experimental design. Examples of blocks are:

- Centres in a multi-centre trial (multiple patients per centre).
- Cars in a fuel consumption experiment (multiple runs per car).
- Patients in a cross-over trial (multiple episodes of treatment for each patient).
- Fermentation tanks in a plant (multiple runs per tank).

2.4.1 Two-way analysis of variance

Suppose there are $r \geq 2$ treatments and $b \geq 2$ blocks. Let y_{ij} be the observation for treatment i in block j ($i = 1, \dots, r, j = 1, \dots, b$). The main problem is to decide whether or not the treatment means are equal.

Statistical model: The data are observed values of independent random variables Y_{ij} satisfying the linear model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (10)$$

for $i = 1, \dots, r, j = 1, \dots, b$, where μ = overall effect, α_i = effect of the i th treatment, β_j = effect of the j th block and the ϵ_{ij} are independent $N(0, \sigma^2)$ r.v.’s. The constraints are $\alpha_{\cdot} = \beta_{\cdot} = 0$. Note that we are assuming a common within-cells variance.

Analysis: partition of the total sum of squares. As in the one-way case, each observation can be written as the sum of deviations:

$$y_{ij} = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})$$

Likewise the total variation may be partitioned as

$$\begin{aligned} \sum_{i=1}^r \sum_{j=1}^b (y_{ij} - \bar{y}_{\cdot\cdot})^2 &= \sum_{i=1}^r \sum_{j=1}^b \{(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})\}^2 \\ &= b \sum_{i=1}^r (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + r \sum_{j=1}^b (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^r \sum_{j=1}^b (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2 \\ \text{or } S_{\text{Total}} &= S_{\text{Treat}} + S_{\text{Block}} + S_{\text{Error}} . \end{aligned}$$

Furthermore,

$$\begin{aligned} E(S_{\text{Error}}) &= (r-1)(b-1)\sigma^2 & E(S_{\text{Treat}}) &= (r-1)\sigma^2 + b \sum_{i=1}^r \alpha_i^2, \\ E(S_{\text{Block}}) &= (b-1)\sigma^2 + r \sum_{j=1}^b \beta_j^2 \end{aligned}$$

(*Exercise:* check that all the cross-product terms vanish and verify the expectations).

For the inference about the effect of the treatments, the mean-square ratio of interest is

$$F = \frac{S_{\text{Treat}}/(r-1)}{S_{\text{Error}}/[(r-1)(b-1)]}.$$

We have $S_{\text{Error}} \sim \sigma^2 \chi^2((r-1)(b-1))$ and, under $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_r = 0$, Cochran's theorem can be used to show that $S_{\text{Treat}} \sim \sigma^2 \chi^2(r-1)$ independently of S_{Error} . Therefore, under H_0 , the corresponding mean-square ratio has the $F(r-1, (r-1)(b-1))$ distribution.

Analysis of variance (ANOVA) table. The s.s. partition and expected mean squares are set out in the table below.

Source of variation	s.s.	d.f.	Expected m.s.
Between treatments	S_{Treat}	$r-1$	$\sigma^2 + \frac{b \sum_i \alpha_i^2}{r-1}$
Between blocks	S_{Block}	$b-1$	$\sigma^2 + \frac{r \sum_j \beta_j^2}{b-1}$
Residual (error)	S_{Error}	$(r-1)(b-1)$	σ^2
Total	S_{Total}	$rb-1$	

The following useful results can be derived (see also the exercises):

- The test procedure is equivalent to the paired sample t -test when $r = 2$.
- For hand calculation, the following formulae can be used:

$$S_{\text{Treat}} = \frac{\sum_{i=1}^r y_{i.}^2}{b} - CF, \quad S_{\text{Block}} = \frac{\sum_{j=1}^b y_{.j}^2}{r} - CF \quad \text{and} \quad S_{\text{Total}} = \sum_{i=1}^r \sum_{j=1}^b y_{ij}^2 - CF,$$

where CF (correction factor) $= y_{..}^2 / rb$. These formulae generalise straightforwardly from the one-way case.

- Again, coding the data $y' = ay + b$ does not affect the variance ratios.

Analysis: Least-squares estimation of μ , α_i and β_j . As before, to estimate the parameters we can minimise the sum of squared deviations

$$R = \sum_{i=1}^r \sum_{j=1}^b (y_{ij} - \mu - \alpha_i - \beta_j)^2$$

w.r.t. μ , α_i and β_j , and subject to the constraints $\alpha_{\cdot} = \beta_{\cdot} = 0$, to give the least-squares estimators. These are (see exercises)

$$\hat{\mu} = \bar{y}_{\cdot\cdot}, \quad \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}, \quad \hat{\beta}_j = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}.$$

Confidence interval for the difference between two treatment means. The estimated s.e. of the mean difference $\bar{y}_{i\cdot} - \bar{y}_{j\cdot}$ between treatments i and j is $s\sqrt{2/b}$, where $s^2 = S_{\text{Error}}/(r-1)(b-1)$, the error m.s. Therefore a $100(1-\alpha)\%$ confidence interval for $\alpha_i - \alpha_j$ is

$$(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) \pm t_{\alpha/2}((r-1)(b-1))s\sqrt{2/b}$$

Confidence interval for the error variance. Since $S_{\text{Error}} \sim \sigma^2 \chi^2((r-1)(b-1))$ a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{S_{\text{Error}}}{\chi_{\alpha/2}^2((r-1)(b-1))}, \frac{S_{\text{Error}}}{\chi_{1-\alpha/2}^2((r-1)(b-1))} \right)$$

Inference about blocks: the randomised block design is intended as a means of extracting information about the effect of the treatment. However, as with the paired sample t -test, blocking is only effective in increasing precision if there are genuine differences between blocks. It may therefore be of interest to test the null hypothesis of no block effect, since if this holds then it will be preferable to ignore the blocks and analyse as a one-way layout. It is clear from the discussion above that, mathematically at least, tests for block effects, and inference about the block parameters $\{\beta_j\}$, can be dealt with in exactly the same way as for the treatments. If the model (10) holds then this can certainly be done, and testing for block effects is common in modern statistical practice. Some authors have argued, however, that a formal analysis of the block effects should not be undertaken since randomisation has only been carried out within blocks. If you subscribe to this argument, it may still be worth looking at the ratio $S_{\text{Block}}/S_{\text{Error}}$ informally to see whether blocking was worthwhile.

There are, of course, other situations in which we are genuinely interested in the effect of two factors upon a response. Such situations can be dealt with using **factorial designs** (see §2.7).

2.4.2 Analysis

To simplify the calculations, it is helpful to work with data coded as $y' = y - 50$ in this example. The necessary information can be presented in a table:

						$y'_{i.}$	$y'^2_{i.}$
A	12	12	2	4	0	30	900
B	6	2	7	0	-9	6	36
C	0	4	-3	-9	-9	-17	289
B	8	1	1	-3	-2	5	25
$y'_{.j}$	26	19	7	-8	-20	24	1250
$y'^2_{.j}$	676	361	49	64	400	1550	

Total, $y'_{.} = 24$, squares, $\sum \sum y'^2_{ij} = 744$, $CF = 24^2/20 = 28.8$

$S_{\text{Total}} = 744 - 28.8 = 715.2$, $S_{\text{Treat}} = \frac{1250}{5} - 28.8 = 221.2$, $S_{\text{Block}} = \frac{1550}{4} - 28.8 = 358.7$

The ANOVA table is therefore

Source of variation	s.s.	d.f.	m.s.	m.s.r.
Between groups	358.7	4	89.67	7.95
Between tutors	221.2	3	73.73	6.54
Residual (error)	135.3	12	11.28	
Total	715.2	19		

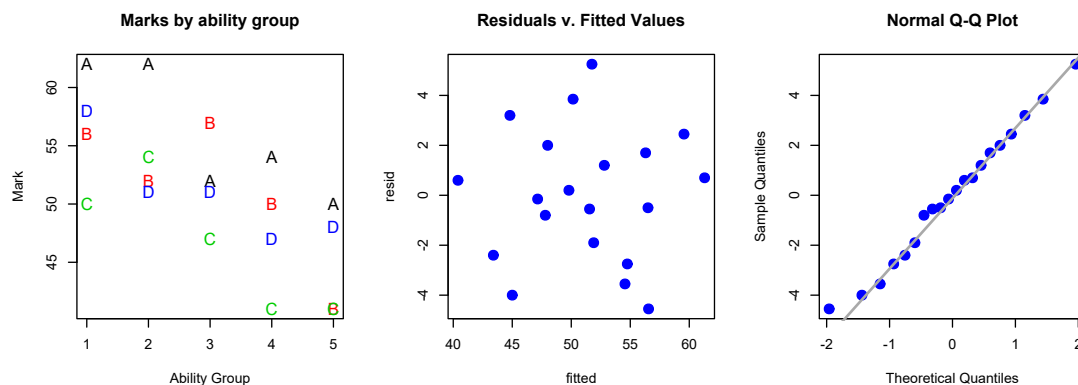
The 1% point of $F(3, 12)$ is 5.95. We therefore reject the hypothesis of no difference between the tutors at the 1% level — there is strong evidence of a real difference between the tutors.

Also, the 1% point of $F(4, 12)$ is 5.41 strongly indicating a real difference between the ability groups, as expected (**question:** do you think it was necessary to test this?). Thus blocking has been effective in reducing the error variation.

Using R: Example 7.

```
> options(contrasts=c("contr.sum", "contr.sum"))
> fit <- aov(mark ~ group + tutor, marks.data)
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	4	358.70	89.67	7.9534	0.002261 **
tutor	3	221.20	73.73	6.5395	0.007194 **
Residuals	12	135.30	11.28		



2.4.3 Partition of the treatment s.s.

Continuing with Example 7. Suppose that, tutors B, C and D are trainees and that A has several years' experience. In this case, we may be interested in comparing B, C and D between themselves, and in comparing A with the other three. In terms of the model (10), the quantities of interest are $\alpha_1 - (\alpha_2 + \alpha_3 + \alpha_4)/3$ (to compare A with the rest) and $\{\alpha_i - (\alpha_2 + \alpha_3 + \alpha_4)/3 : i = 2, 3, 4\}$ (to compare the others between themselves). Notice that all of these quantities take the form $\sum_i \lambda_i \alpha_i$ with $\sum_i \lambda_i = 0$ — they are contrasts.

One way to compare tutor A with the other three would be to test the hypothesis $H_0 : \alpha_1 = 0$ (this is the same as the hypothesis of interest, because of the constraint $\sum_i \alpha_i = 0$). This could be done by carrying out a t -test on the estimate $\hat{\alpha}_1$ (we would need to calculate the standard error of $\hat{\alpha}_1 = \bar{y}_1. - \bar{y}..$, taking into account the fact that $\bar{y}_1.$ and $\bar{y}..$ are not independent). However, this would not tell us anything about how tutors B, C and D compare among themselves.

An alternative approach is to partition $S_{\text{Treat}} = 221.2$ into (i) a sum of squares for the variation due to a binary grouping tutor A versus tutors B, C, and D taken together, and (ii) a sum of squares for the variation among tutors B, C, and D.

To get (i), re-code the data so that there are two treatment groups: A and {B,C,D}. Next run a two-way ANOVA. This leads to the following result

Source of variation	s.s.	d.f.	m.s.	m.s.r.
Between groups	153.6	1	153.6	
Between tutors				
Residual (error)				
Total				

The resulting between-group s.s. corresponds to the difference between tutor A and the rest. (ii) can now be derived by subtraction: $221.2 - 153.6 = 67.6$. Combining (i) and (ii), we get

Source of variation	s.s.	d.f.	m.s.	m.s.r.
Between A and the rest	153.6	1	153.6	13.6
Among B, C, D	67.6	2	33.8	3.0
Residual (error)	135.3	12	11.3	

F-tests can now be used (this will be shown later). Conclusion: no evidence of a difference between B, C, and D, but strong evidence that tutor A has an effect.

2.4.4 Interaction

The linear additive model (10) implies that the treatment effect is independent of the level of the block effect. But this may not always be true. In Example 7 there might be little difference between the tutors with higher ability pupils, but a substantial difference with lower ability pupils. This is known as an **interaction** effect. We can include an adjustment or interaction term in the model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}.$$

If $\gamma_{ij} \neq 0$ then the treatment and block effects are not additive, and interpretation becomes more difficult — α_i is now the *average* effect of treatment i over all blocks.

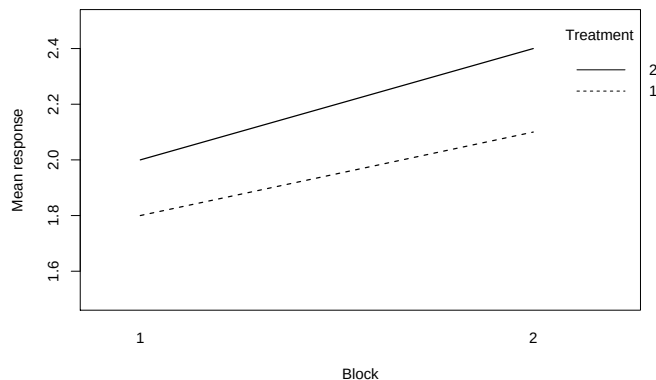
To illustrate the idea, suppose we have the following data:

	Block 1	Block 2
Treatment 1	1.80	2.10
Treatment 2	2.00	2.40

Treatment effect in block 1 is $2.00 - 1.80 = 0.20$
 Treatment effect in block 2 is $2.40 - 2.10 = 0.30$

} An apparent interaction

Alternatively, an *interaction plot* may be used for a graphical display:



In the absence of interaction, the lines in such a plot are expected to be parallel.

If we fit an additive model to these data, we obtain $\hat{\mu} = \bar{y}_{..} = 2.075$, $\hat{\alpha}_1 = \bar{y}_{1.} - \bar{y}_{..} = -0.125$, $\hat{\alpha}_2 = \bar{y}_{2.} - \bar{y}_{..} = +0.125$, $\hat{\beta}_1 = \bar{y}_{.1} - \bar{y}_{..} = -0.175$, $\hat{\beta}_2 = \bar{y}_{.2} - \bar{y}_{..} = +0.175$. The residuals $\{r_{ij} = y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j\}$ are

	Block 1	Block 2
Treatment 1	0.025	-0.025
Treatment 2	-0.025	0.025

However, notice that under the interaction model, we have $y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j = \gamma_{ij} + \epsilon_{ij}$. Each residual is therefore effectively estimating a combination of interaction effect and error variation — there is no way of telling how much of 0.025 is due to each component. In order to do this, we need an independent estimate of σ^2 , which is not possible with just one observation for each treatment / block combination. It *is* possible, however, if we take more than one observation per cell, since then we can pool the within-cells sample variances to estimate σ^2 .

2.4.5 Multiple observations per cell

To overcome the interaction problem, suppose that we have a two-way layout with $m > 1$ observations for each block / treatment combination (for example, in Example 7, if the class was bigger we might be able to allocate $m = 2$ students from each group to each tutor — using randomisation to perform the allocation, of course). A statistical model that allows for interaction is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \quad (11)$$

with $i = 1, \dots, r$, $j = 1, \dots, b$, $k = 1, \dots, m$. In this model, γ_{ij} is the interaction effect of the i th row and j th column. The α and β terms are called the **main effects**. The $\{\alpha_i\}$ are the treatment effects, averaged over the b blocks *etc.*. The constraints are $\alpha_{.} = \beta_{.} = \gamma_{i.} = \gamma_{.j} = 0$. Hypotheses to test are $\alpha_1 = \dots = \alpha_r = 0$ (no treatment effect), $\beta_1 = \dots = \beta_b = 0$ (no block effect) and $\gamma_{ij} = 0$ for all i, j (no interaction).

Analysis: partition of the total sum of squares. Each observation can be written as

$$y_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

and the total s.s. may be partitioned as

$$\begin{aligned} \sum_{i=1}^r \sum_{j=1}^b \sum_{k=1}^m (y_{ijk} - \bar{y}_{...})^2 &= bm \sum_{i=1}^r (\bar{y}_{i..} - \bar{y}_{...})^2 + rm \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2 \\ &\quad + m \sum_{i=1}^r \sum_{j=1}^b (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^r \sum_{j=1}^b \sum_{k=1}^m (y_{ijk} - \bar{y}_{ij.})^2 \end{aligned}$$

or

$$S_{\text{Total}} = S_{\text{Treat}} + S_{\text{Block}} + S_{\text{Inter}} + S_{\text{Error}}$$

This partition and expected mean squares are set out below.

Source of variation	s.s.	d.f.	Expected m.s.
Between treatments	S_{Treat}	$r - 1$	$\sigma^2 + \frac{bm \sum_i \alpha_i^2}{r-1}$
Between blocks	S_{Block}	$b - 1$	$\sigma^2 + \frac{rm \sum_j \beta_j^2}{b-1}$
Interaction	S_{Inter}	$(r - 1)(b - 1)$	$\sigma^2 + \frac{m \sum_i \sum_j \gamma_{ij}^2}{(r-1)(b-1)}$
Within cells (error)	S_{Error}	$rb(m - 1)$	σ^2
Total	S_{Total}	$rbm - 1$	

Analysis: least-squares estimation. To estimate the model parameters, we can minimise the sum of squares

$$R = \sum_{i=1}^r \sum_{j=1}^b \sum_{k=1}^m (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2$$

w.r.t. μ , α_i , β_j and γ_{ij} . This yields

$$\hat{\mu} = \bar{y}_{...}, \quad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \quad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \quad \hat{\gamma}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

For hand calculation, we have

$$S_{\text{Total}} = S - CF, \quad S_{\text{Treat}} = \frac{\sum_{i=1}^r y_{i..}^2}{bm} - CF, \quad S_{\text{Block}} = \frac{\sum_{j=1}^b y_{.j.}^2}{rm} - CF$$

$$\text{and } S_{\text{Error}} = S - \frac{\sum_{i=1}^r \sum_{j=1}^b y_{ij.}^2}{m},$$

where $S = \sum_{i,j,k} y_{ijk}^2$ and now $CF = y_{...}^2 / rbm$. Then S_{Inter} may be obtained by subtraction.

Confidence intervals for differences between row (column) means and for σ^2 may be constructed in the usual way, using the error m.s. as estimate of σ^2 . An example will be given in §2.7.

2.4.6 Unbalanced data

So far when discussing the randomised block design, we have assumed that each treatment appears the same number of times within each block. What happens if this is not the case? For example, it may be that the original design was for randomised complete blocks but some of the experimental runs failed, or patients left the study. In general, the consequences of this will be some loss

of efficiency in estimation (compared to some theoretical optimum) and complications in the analysis: the total sum of squares cannot be decomposed uniquely into “treatment” and “block” contributions. It is usually easier to analyse such data using a **general linear model** (see §2.12) than to compute an ANOVA table. For the moment, however, we will continue to work with **balanced designs** since their analysis is more straightforward.

2.5 Latin square design

The randomised (complete) block design is appropriate for situations when there is a single treatment variable of interest and a single ‘environmental’ factor that may also affect the response. What happens if there are two such environmental factors? Intuitively, it seems that it may be necessary to consider a three-way layout. However, this may be expensive. It turns out that if the number of levels of each environmental factor is the same as for the treatment then a balanced two-way layout can be produced with three factors in a **Latin square** arrangement.

A Latin square is an $p \times p$ array filled with p different symbols, each occurring exactly once in each row and exactly once in each column. Sudoku is one of the best known examples of a Latin square.

With respect to ANOVA, using a Latin square reduces costs and, providing there are no interactions, controls for both environmental factors simultaneously.

Example 8: Four tyre brands, A, B, C and D, are to be compared. Four tyres of each type are put on each of four cars. There are 4 positions (front/rear $\times$ nearside/offside). Tyre wears are recorded and tabulated as follows (in coded form):

		Car			
		I	II	III	IV
Position	I	-1(C)	-2(D)	0(A)	-5(B)
	II	1(B)	-1(C)	-2(D)	0(A)
	III	4(A)	1(B)	-3(C)	-4(D)
	IV	0(D)	1(A)	0(B)	-4(C)

Note that every letter (treatment) appears once and only once in each row and in each column, leading to a balanced design.

2.5.1 Latin square design: theory

Randomisation. In a Latin square design, there is only one observation for each combination of the environmental factors. Hence, it is not possible to randomise within blocks. Instead, however, randomisation can be achieved by choosing a Latin square at random from all possible Latin

squares of the required size p . For example, when $p = 4$ there are four **standard squares** in which the first row and column are both in alphabetical order:

A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
B	C	D	A	B	D	A	C	B	A	D	C	B	A	D	C
C	D	A	B	C	A	D	B	C	D	A	B	C	D	B	A
D	A	B	C	D	C	B	A	D	C	B	A	D	C	A	B

None of these squares can be obtained by permuting the rows and/or columns of any of the others. Each standard square can, however, be used to generate other squares by permuting rows and columns. To choose a square at random:

1. Choose a basic square type at random.
2. Randomise the order of the rows.
3. Randomise the order of all but the first column.

It is not obvious that this generates all possible unique permutations (why not randomise the order of *all* the columns in step 3?), but in fact it does. To convince yourself of this, try it with a 2×2 square. Given this, the total number of possible squares is clearly $4 \times 4! \times 3! = 576$ — this is a reasonably large ‘population’ within which to interpret the results of the experiment. In practice sometimes ‘unsuitable’ squares are rejected *e.g.* the same letter down a diagonal in an agricultural field trial.

The advantages and disadvantages of Latin square designs are as follows:

Advantages

- Those of a randomised block design (elimination of bias, increased precision)
- Two factors can be eliminated with only p^2 experimental units instead of p^3 (using a 3-way layout), thereby saving experimental material

Disadvantages

- Constrained to have all 3 factors at the same number of levels
- Cannot determine interactions, which may be important
- Tend to have few d.f. for error estimation unless large squares are used — but these are cumbersome

Statistical model: for the Latin square design, the appropriate model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}$$

for $i, j, k = 1, \dots, p$, where the value of k (the treatment) applied with a given (i, j) is determined by the design. The interpretation of the parameters, and of ϵ_{ijk} , should be familiar by now. The constraints are $\alpha_{..} = \beta_{..} = \gamma_{..} = 0$. Notice that the model is additive.

Analysis: the least-squares estimators of the parameters are

$$\hat{\mu} = \bar{y}_{...}, \quad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \quad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \quad \hat{\gamma}_k = \bar{y}_{..k} - \bar{y}_{...}$$

The s.s. partition, d.f. and expected mean squares are as follows:

Source of variation	s.s.	d.f.	Expected m.s.
Between rows	$S_R = p \sum_i \hat{\alpha}_i^2$	$p - 1$	$\sigma^2 + \frac{p \sum_i \alpha_i^2}{p-1}$
Between columns	$S_C = p \sum_j \hat{\beta}_j^2$	$p - 1$	$\sigma^2 + \frac{p \sum_j \beta_j^2}{p-1}$
Between Latin letters	$S_L = p \sum_k \hat{\gamma}_k^2$	$p - 1$	$\sigma^2 + \frac{p \sum_k \gamma_k^2}{p-1}$
Residual (error)	S_{Error} (by subtraction)	$(p-1)(p-2)$	σ^2
Total	S_{Total}	$p^2 - 1$	

Notice that only $(p-1)(p-2)$ degrees of freedom remain after estimating the main effects — this is why interactions cannot be estimated. *Question:* what is the minimum number of d.f. required to estimate interactions?).

Using R: Example 8

```
> fit <- aov(wear ~ position + car + brand, wear.data)
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
position	3	6.188	2.063	2.3023	0.176947
car	3	38.688	12.896	14.3953	0.003784 **
brand	3	30.688	10.229	11.4186	0.006825 **
Residuals	6	5.375	0.896		

The difference between the tyre brands is significant at the 1% level. Also, there are significant differences between cars but not between the positions on the car.

2.6 Contrasts

Linear combinations $\sum_i \lambda_i x_i$ are called **contrasts of x_i s** when $\sum_i \lambda_i = 0$. In an ANOVA setting, we are typically interested in contrasts of treatment effects; for example, $\sum_{i=1}^2 \lambda_i \alpha_i = \alpha_1 - \alpha_2$ in Section 2.3.1, and $\sum_{i=1}^4 \lambda_i \alpha_i = \alpha_1 - (\alpha_2 + \alpha_3 + \alpha_4)/3$ in Section 2.4.3.

Because contrasts play an important role in the next section on factorial experimentation, it is worthwhile to discuss contrasts in more detail first.

Consider, for example, the following model for a two-way ANOVA:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad \text{with} \quad \epsilon_{ij} \sim N(0, \sigma^2),$$

for treatment level $i = 1, \dots, r$, and block $j = 1, \dots, b$. The restrictions are $\alpha_{\cdot} = \beta_{\cdot} = 0$. Define **treatment mean** $\bar{Y}_i = \frac{1}{b} \sum_{j=1}^b Y_{ij}$ for each treatment i . We can define the estimator for α_1 as a contrast of treatment means, and use the contrast to derive the variance of the estimator as a function of σ^2 . Write

$$\hat{\alpha}_1 = \bar{Y}_{1\cdot} - \bar{Y}_{\cdot\cdot} = \bar{Y}_{1\cdot} - \frac{\bar{Y}_{1\cdot} + \dots + \bar{Y}_{r\cdot}}{r} = \frac{(r-1)\bar{Y}_{1\cdot}}{r} - \frac{\bar{Y}_{2\cdot}}{r} - \dots - \frac{\bar{Y}_{r\cdot}}{r}.$$

This is a contrast of treatment means defined by $\lambda = (\frac{r-1}{r}, \frac{-1}{r}, \dots, \frac{-1}{r})$, which is a contrast because $\sum_i \lambda_i = 0$. Given independence across treatments and the common variance assumption:

$$\text{Var}(\hat{\alpha}_1) = \sum_{i=1}^r \lambda_i^2 \text{Var}(\bar{Y}_{i\cdot}) = \frac{\sigma^2}{b} \sum_{i=1}^r \lambda_i^2 = \frac{\sigma^2}{b} \left(\frac{(r-1)^2}{r^2} + \frac{(r-1)}{r^2} \right) = \frac{\sigma^2}{b} \frac{(r-1)}{r}.$$

The example illustrates important properties of contrasts of treatment means: we can estimate a contrast using the observed treatment means, and, when we have an estimate of the residual variance σ^2 , we can also estimate the variance of a contrast.

Using R: Example 7 and using contrasts.

```
> # Fit model and extract parameter estimates:
> options(contrasts=c("contr.sum", "contr.sum"))
> fit <- aov(mark ~ group + tutor, marks.data)
> anova <- summary(fit)
> print(round(coef(summary.lm(fit)),3))
              Estimate Std. Error t value Pr(>|t|)
...
tutor1          4.80         1.300    3.691   0.003
...
> # Estimate SE tutor effect by using contrast:
> r <- 4; b <- 5 ; sigma2 <- anova[[1]][3,3]
> var.alpha <- sigma2/b*(r-1)/r
> cat("SE tutor effect = ", sqrt(var.alpha),"\n")
SE tutor effect =  1.300481
```

Contrast: general case. Assume there are t treatments. Consider a contrast $\bar{L} = \sum_i \lambda_i \bar{T}_i$ of treatment means $\bar{T}_1, \dots, \bar{T}_t$, where $\sum_i \lambda_i = 0$. Suppose each treatment is replicated r times. Then the treatment *totals* are given by $T_i = \sum_{j=1}^r Y_{ij} = r\bar{T}_i$. Notice that, if the observations $\{Y_{ij}\}$ are mutually independent (although not necessarily identically distributed), the totals are too. Working with these totals, let $L = \sum_i \lambda_i T_i$ so that the contrast of interest is L/r . Then

$$\text{Var}(L) = \sum_{i=1}^t \lambda_i^2 \text{Var}(T_i) = \sum_{i=1}^t \lambda_i^2 \text{Var}\left(\sum_{j=1}^r Y_{ij}\right) = \sigma^2 \left(r \sum_{i=1}^t \lambda_i^2\right) = \sigma^2 D, \text{ say.}$$

To test $H_0 : E(L) = 0$ we can compare the t -statistic

$$t = \frac{L}{\text{s.e.}(L)} = \frac{L}{\hat{\sigma} D^{1/2}}.$$

with a $t(\nu)$ distribution, where ν is the degrees of freedom used to calculate the residual m.s. $\hat{\sigma}^2$. Equivalently, we could calculate

$$F = t^2 = \frac{L^2/D}{\hat{\sigma}^2} = \frac{\text{m.s. due to } L}{\text{error m.s.}}$$

and compare this to an F distribution with $(1, \nu)$ degrees of freedom. Since there is only one degree of freedom associated with the numerator here, we can think of L^2/D as a ‘contrast sum of squares’.

Next, consider what happens if we are interested in studying more than one contrast simultaneously. In this case, matters are simplified considerably if the contrasts of interest are **orthogonal**. Two contrasts $L_1 = \sum_i \lambda_i T_i$, $L_2 = \sum_i \xi_i T_i$ are said to be orthogonal if $\sum_i \lambda_i \xi_i = 0$ (for example, $\bar{T}_3 - \bar{T}_2$ and $\bar{T}_1 - \frac{1}{2}(\bar{T}_2 + \bar{T}_3)$ are two orthogonal contrasts). Generalising this, contrasts $L_1, \dots, L_k$ are **mutually orthogonal** if every pair is orthogonal.

One advantage of orthogonal contrasts is that they are uncorrelated random variables:

$$\text{Cov}(L_1, L_2) = \sum_{i=1}^t \sum_{j=1}^t \lambda_i \xi_j \text{Cov}(T_i, T_j) = r\sigma^2 \sum_{i=1}^t \lambda_i \xi_i = 0.$$

This allows us to assess the effect of each contrast independently of all the others.

Another advantage is that given a set of $t - 1$ mutually orthogonal contrasts, the treatment s.s. can be written as a sum of contributions from each of them (the proof of this requires matrix methods and is omitted):

$$\text{Treatment s.s.} = \frac{L_1^2}{D_1} + \dots + \frac{L_{t-1}^2}{D_{t-1}}.$$

Each contribution here is a single d.f. sum of squares; hence the treatment sum of squares can be partitioned uniquely into contributions from each of the contrasts (actually, we have already seen something similar to this in the analysis of Example 7 in §2.4). There are many sets of $t - 1$ contrasts; we choose the ones that are relevant to the problem in hand.

2.7 Factorial experiments

So far, we have focused upon problems in which there is a single treatment and one or two nuisance factors. What if there is more than one treatment of interest? This occurs very often in industrial experimentation. In general the most efficient designs in such situations are **factorial designs**, in which each replicate contains all possible combinations of the **levels** (i.e. unique values) of the different factors.

2.7.1 Factorial designs

Suppose we wish to investigate the effects of factors $A_1, A_2, \dots, A_n$, and that A_i has r_i levels. If all levels of each factor are combined with all levels of every other factor in the experiment then we have an $r_1 \times r_2 \times \dots \times r_n$ factorial design. In such a design, the factors are sometimes said to be **crossed**, and the design may be described as **completely cross-classified**. All the factors are regarded as treatments, as opposed to environmental factors, although the design may also include blocks (for example, if replicates are run on different days and there is reason to believe that results may vary from day to day). As well as investigating the effects of categorical factors upon a response, factorial designs can be used to assess the effect of continuous variables, by suitable selection of levels (e.g. “Low” and “High”).

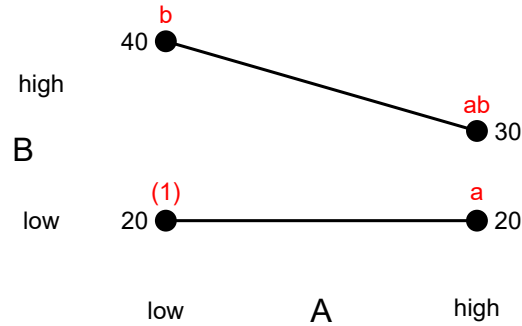
Intuitively, you might think that to investigate the effect of an individual factor, the most appropriate procedure would be to hold everything else constant and vary just that one factor. This is called a *one-factor-at-a-time experimentation*. Say there are factors A, B, and C. If levels of B and C are fixed while A is varied to find the ‘best’ level, then possible interaction is ignored. For example, if the change in mean response with change of A depends on the level of B, then this is not picked up.

Factorial designs have several advantages over one-at-a-time experimentation:

1. Can assess interaction between factors.
2. If there is no interaction it is still worthwhile using a factorial design because it saves on experimental material. The effect of a factor can be assessed in a factorial design with the same precision as if the whole experiment were devoted to that factor. For example, in an experiment to determine the effect of nitrogen (at two levels, ‘None’ and ‘Some’) on yield, half of the plots receive nitrogen and half do not, regardless of how many other factors are included in the analysis. See also the illustration at the end of the next section.
3. By including several factors, we extend the range of validity of the results. For example, when comparing different fertilisers we can simultaneously account for different plant varieties.

2.7.2 The 2^2 design

In the simplest factorial design there are two factors A and B, each at a low and a high level. Say the observations are given by



For example, at the low level of A and the low level of B, we observe the value 20. The effect of A, can be assessed by the difference between the average at the high level of A and the average at the low level of A:

$$\text{Main effect of A:} \quad \frac{30 + 20}{2} - \frac{40 + 20}{2} = -5$$

Similar for the effect of B:

$$\text{Main effect of B:} \quad \frac{40 + 30}{2} - \frac{20 + 20}{2} = 15$$

Note that at the high level of B, the effect of A is 30-40, and at the low level of B, the effect of A is 20-20. The interaction effect AB, can be assessed by the average of these two differences:

$$\text{Interaction effect of AB:} \quad \frac{(30 - 40) - (20 - 20)}{2} = -5$$

The above effects are called the **factorial effects** and we use letters A, B, AB to denote them. There is a direct link between the factorial effects and the effects in the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij} \quad \text{with} \quad \epsilon_{ij} \sim N(0, \sigma^2),$$

where the levels of A are $i \in \{high, low\}$, the levels of B are $j \in \{high, low\}$, and the sum-to-zero constraint is used. It follows that $A = 2\alpha_1$, $B = 2\beta_1$, and $AB = 2\alpha\beta_{11}$. For example, effect α_1 is the deviation from the mean μ for the low level of A. Two times this deviation makes up the effect of going from the high level of A to the low level of A. In other words, in the model we are interested in deviations about an overall mean, whereas in the 2^2 design we are interested in differences between the factors.

Using R: illustration factorial effects. In R, the notation $A:B$ is used to denote the interaction between A and B; $A:B:C$ is a three-way interaction between A, B and C and so forth. The notation $A*B$ is used as shorthand for $A + B + A:B$ (main effects, together with all their interactions).

```
> # Data:
> dta <- data.frame(Y=c(20,40,20,30), A = factor(c("low","low","high","high")),
                    B = factor(c("low","high","low","high")))
> # For sum-to-zero parameterisation:
> options(contrasts = c("contr.sum", "contr.sum"))
> # Model, coefficients, and factorial effects:
> p <- coef(summary.lm( aov(Y ~ A*B, dta) ))[,1]
> print(2*p[2:4])
```

```
  A1    B1 A1:B1
-5    15    -5
```

Factorial effects: notation. In factorial design designs it is conventional to represent the mean response for different treatment combinations symbolically, using lower-case letters to denote the high level of each factor.

For the 2^2 design, ‘ ab ’ represents the mean response for the high level of A and the high level of B, ‘ a ’ represents that for the high level of A and the low level of B, and ‘ b ’ represents that for the low level of A and high level of B. The mean response when both factors are at their low levels is denoted by ‘(1)’. Suppose, that each treatment combination is replicated r times in the experiment. We can display the mean responses in a 2×2 table as follows:

		B	
		low	high
A	low	(1)	b
	high	a	ab

We can now express various quantities of interest using our symbolic notation as follows:

$$\begin{array}{ll}
 \text{Overall mean:} & \frac{(1) + a + b + ab}{4} \\
 \text{Main effect of A:} & \frac{a + ab}{2} - \frac{(1) + b}{2} = \frac{-(1) + a - b + ab}{2} \\
 \text{Main effect of B:} & \frac{b + ab}{2} - \frac{a + (1)}{2} = \frac{-(1) - a + b + ab}{2} \\
 \text{AB interaction:} & \frac{ab - b}{2} - \frac{a - (1)}{2} = \frac{(1) - a - b + ab}{2}
 \end{array}$$

Notice that the BA interaction is equal to the AB interaction: $(ab - a)/2 - (b - (1))/2 = (ab - a - b + (1))/2$.

The **table of signs** summarises the above computation:

Factorial effect	Treatment combination				Divisor
	(1)	<i>a</i>	<i>b</i>	<i>ab</i>	
Mean	+	+	+	+	4
A	-	+	-	+	2
B	-	-	+	+	2
AB	+	-	-	+	2

As the table of signs shows: **the factorial effects A, B and AB are contrasts of the means of the treatment combinations**. This is an essential feature of factorial designs and it can be used to derive tests for the factorial effects (see the next section).

2.7.3 Partition of a treatment s.s. into single d.f. components

As well as estimating the effects in a factorial experiment, we might be interested in testing hypotheses about them. This could be done using an ANOVA table, exactly as in §2.4.5 — each of A, B and AB would have an associated sum of squares and a single degree of freedom. It is, however, instructive to consider what can be achieved using contrasts.

In the 2^2 design, we can see that the factorial effects A, B and AB are mutually orthogonal contrasts, whence the theory in Section 2.6 applies. Tests for each effect can be based either on t -statistics or upon F -ratios (although some authors have argued that to use F -ratios and ANOVA tables in this situation is unnecessarily complicated). Notice that in this particular situation, the $\{\lambda_i\}$ are all either $-1/2$ or $1/2$ and $\sum \lambda_i^2 = 1$; hence $D = r$, the number of replicates, here. This makes the analysis particularly simple in this case.

Example 9 (taken from Clarke & Kempson): in an experiment to investigate the quality of output from a manufacturing process, it was decided to investigate two factors: the speed of running a machine (factor A — two speeds) and the supplier of a particular material (factor B — two sources). Five replicates of a 2^2 design were run on different days; on each day, a batch was produced under each factor combination and the average number of defectives in each batch was recorded. The data were as follows (denoting the ‘higher’ levels of A and B by a and b respectively):

		Day					Total
		I	II	III	IV	V	
Treatment	(1)	5.3	5.7	5.1	5.3	5.6	27
combination	a	11.8	13.0	12.6	12.1	11.5	61
	b	20.0	19.0	20.3	19.5	20.2	99
	ab	26.7	24.1	25.7	26.0	25.5	128

Nota bene, using averages as data is not an essential feature of a factorial design; see the exercises for an example with data consisting of single observations.

In the above experiment, the replicates of the 2×2 factorial are arranged in a randomised block design, with days determining the block structure. We start by considering the treatment combinations as four different treatments. The analysis is then identical to that for a randomised block design and, for hand calculation, can be set out as for Example 7. The ANOVA table is

Source of variation	s.s.	d.f.	m.s.	m.s.r.
Between treatments	1165.750	3	388.58330	795.06
Between days	0.655	4	0.16375	0.34
Residual (error)	5.865	12	0.48875	
Total	1172.270	19		

Note that the day effect is not significant. We now partition the treatment s.s. into its factorial components by combining the treatment totals with the table of signs. Recall that $\lambda_i = \pm 1/2$ for the 2^2 design, and that $D = r = 5$ in this instance. We therefore have:

Factorial effect	Mean responses				$L = \sum \lambda_i T_i$	s.s. = L^2/D
	(1)	a	b	ab		
A	-27	61	-99	128	31.5	198.45
B	-27	-61	99	128	69.5	966.05
AB	27	-61	-99	128	-2.5	1.25

(**Exercise:** check that the treatment sums of squares add up). The full ANOVA table can now be constructed:

Source of variation	s.s.	d.f.	m.s.	m.s.r.
A	198.450	1	198.45000	406.04
B	966.050	1	966.05000	1976.57
AB	1.250	1	1.25000	2.56
<i>Between treatments</i>	<i>1165.750</i>	<i>3</i>	<i>388.58330</i>	<i>795.06</i>
Between days	0.655	4	0.16375	0.34
Residual (error)	5.865	12	0.48875	
Total	1172.270	19		

There is no evidence of an interaction between source of supply and running speed. Both source of supply and running speed are highly significant, indicating very strong evidence that both affect batch quality.

An alternative way to analyse the treatment effects would be using t -tests on the effect estimates or totals. For example, the estimated effect for B is $69.5/5 = 13.9$. The t -statistic for the effect total is $69.5/(\hat{\sigma}^2 D)^{1/2} = 69.5/\sqrt{0.48875 \times 5} = 44.459$ (squaring this gives the F -statistic for B in the ANOVA table above). Compared to a $t(12)$ distribution this is highly significant. Notice, however, that in this case it was necessary to compute the initial ANOVA table in order to account for the block effect and find the residual mean square.

Using R: Example 9. Define `Treat` as variable with 4 levels representing the 4 treatment combinations. Ignoring the factorial structure we get

```
> fit1 <- aov(Defect ~ Day + Treat)
> summary(fit1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Day	4	0.66	0.16	0.335	0.8492
Treat	3	1165.75	388.58	795.055	4.607e-14 ***
Residuals	12	5.87	0.49		

Now add the factorial structure to give

```
> fit2 <- aov(Defect ~ Day + Supply * Machine)
> summary(fit2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Day	4	0.66	0.16	0.3350	0.8492
Supply	1	966.05	966.05	1976.5729	1.088e-14 ***
Machine	1	198.45	198.45	406.0358	1.278e-10 ***
Supply:Machine	1	1.25	1.25	2.5575	0.1358
Residuals	12	5.86	0.49		

Note that *if* the interaction between A and B would have been significant, *then* the main effects should not be interpreted on their own; see Sections 2.4.4 and 2.4.5.

2.7.4 The 2^n design

The advantage of the symbolical representation and the usefulness of the table of signs becomes more apparent for more complex factorial designs. Here we briefly discuss the 2^n design.

If we have n factors, each at two levels, then the development is almost identical to the 2^2 case. The main difference is that there are more possible interactions now. Apart from n main effects, there are $\binom{n}{2}$ 2-factor interactions, $\binom{n}{3}$ 3-factor interactions and, in general, $\binom{n}{k}$ k -factor

interactions. With 2^n treatment combinations in total, there are $2^n - 1$ degrees of freedom which can be partitioned into contributions from these different orders of interaction:

$$2^n - 1 = \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}$$

(**revision:** prove this identity using the binomial theorem). Therefore, we now have $2^n - 1$ factorial effects. These can be computed from a table of signs (or using other techniques such as *Yates' method* — see course STAT0012 *Factorial Experimentation*). To create such a table in the general case, start by writing down rows for the main effects — for example, the row corresponding to A has ‘+’ whenever the symbolic treatment label contains an a and ‘−’ otherwise. The signs for interactions are then (remarkably!) obtained simply by multiplying out the signs for the corresponding main effects (look at the table in the previous section to verify this in the 2^2 case). The divisors for the factorial effects are 2^{n-1} .

The sum of squares due to a factorial effect is L^2/D (see earlier), where now $D = r \sum_i \lambda_i^2 = 2^{2-n}r$ (since $\lambda_i = \pm 2^{1-n}$ and the number of treatment combinations is 2^n). Thus

$$\text{s.s. due to a factorial effect} = \frac{2^{n-2} \times (\text{factorial effect total from experiment})^2}{r}.$$

The factorial effects are $2^n - 1$ mutually orthogonal contrasts of the 2^n treatment combinations.

So far, we have implicitly assumed that the r replicates of a factorial experiment are made under identical conditions. Sometimes this is not possible, however — for example, if experimental material is supplied in batches and a single batch does not provide enough material for all the replicates. In this case, different replicates can be carried out within blocks, or Latin squares, as necessary. The easiest way to analyse such an experiment is to start by considering all of the treatment combinations as separate treatments and carry out the analysis in the usual manner to obtain a treatment sum of squares; and then to split this into the $2^n - 1$ factorial effects. If it can be assumed that the higher-order interactions are absent (by a combination of testing and prior knowledge) then these may be pooled into the residual term. In particular, provided at least one interaction term is absent, then it is not necessary for the design to be replicated in order to estimate the error variance.

Finally: do not forget randomisation! **Question:** what, if anything, is the difference between randomisation in a two-way factorial design and in a randomised complete block design?

2.8 Random effects

Up to now we have assumed (implicitly) that the factor levels in our experiments are the only levels of interest i.e. they are **fixed**. Often, however, the specific levels available may be regarded as a sample from a larger population. In this case we could view the factor as a **random factor**. To see whether this makes any difference to the analysis, we can write down and analyse a **random-effects model**.

Example 10 (Montgomery, Exercise 13.9): In an experiment to investigate the capability of a measurement system, ten machine parts were randomly selected, and two randomly selected operators measured each part three times (in random order). The data were as follows:

		Operator 1			Operator 2		
		1	2	3	1	2	3
Part Number	1	50	49	50	50	48	51
	2	52	52	51	51	51	51
	3	53	50	50	54	52	51
	4	49	51	50	48	50	51
	5	48	49	48	48	49	48
	6	52	50	50	52	50	50
	7	51	51	51	51	50	50
	8	52	50	49	53	48	50
	9	50	51	50	51	48	49
	10	47	46	49	46	47	48

We will consider operators as a *factor* here (rather than a blocking), since they are considered to be part of the measurement system rather than an uncontrollable nuisance variable. In this case, we have a 10×2 factorial design with 3 replicates per cell. However, it is likely that we are interested in characterising the measurement process for *any* operators and parts, rather than just those selected for this particular experiment. We should therefore treat both operator and part number as random factors.

2.8.1 One-way random-effects model

Before dealing with Example 10, we will consider a slightly simpler setup in which there are $k \geq 2$ treatment groups, considered to be randomly selected from a population of such groups, and a random sample is available from each group. Let y_{ij} be the j th observation in the i th treatment group, where $j = 1, \dots, n$, $i = 1, \dots, k$ (note: equal group sizes). The main problem is to decide whether or not the group means vary in the population.

Statistical model: As in the fixed-effects case, it seems reasonable to consider the data to be observed values of random variables $\{Y_{ij}\}$, with

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} .$$

However, whereas previously the $\{\alpha_i\}$ were fixed, they are now considered themselves to be random variables — this reflects the fact that the groups in our experiment are just a sample from a larger population. Specifically, we will assume that the $\{\alpha_i\}$ are independent $N(0, \sigma_\alpha^2)$ r.v.'s. In this case, the null and alternative hypotheses are $H_0 : \sigma_\alpha^2 = 0$ and $H_1 : \sigma_\alpha^2 > 0$ respectively.

Analysis: partition of the total sum of squares. The usual partition of the total sum of squares ($S_T = S_G + S_E$) is a purely algebraic result, so it holds here just as in the fixed-effects case. Also, $E(S_E) = k(n-1)\sigma^2$ as before. However, the expected sum of squares for treatments is now $E(S_G) = (k-1)(\sigma^2 + n\sigma_\alpha^2)$ (see exercises). We therefore see that the treatment group m.s. is an unbiased estimator of σ^2 if and only if H_0 is true, otherwise it will overestimate σ^2 . The logic now follows as in the fixed-effects case; the usual F test remains valid, and the ANOVA table can be constructed as before.

Analysis: estimation of the parameters. The least-squares estimate of μ is $\hat{\mu} = \bar{y}_{..}$, as in the fixed-effects case. The variance σ^2 can be estimated as $\text{m.s.}_E = S_E/k(n-1)$, again as in the fixed-effects case. The only other parameter in the model is σ_α^2 , the variance of the random effects. Since we have $E(S_G) = (k-1)(\sigma^2 + n\sigma_\alpha^2)$, a simple estimator of σ_α^2 is

$$\hat{\sigma}_\alpha^2 = \frac{\text{m.s.}_G - \text{m.s.}_E}{n}.$$

(see exercises). In the unbalanced case (n_i in the i th group), n in the denominator here is replaced with

$$n_0 = \frac{1}{k-1} \left[N - N^{-1} \sum_{i=1}^k n_i^2 \right],$$

where $N = \sum_{i=1}^k n_i$.

Estimators derived in this way, by equating sample properties with their expectations, are called **moment estimators**. A disadvantage of the estimator $\hat{\sigma}_\alpha^2$ is that it can produce negative values. Usually this indicates little or no treatment variation, but some caution needs to be exercised.

2.8.2 Two-factor random-effects model

If, as in Example 10, there are two random factors then things get a bit more complicated. Suppose there are $r \geq 2$ ‘row’ categories and $c \geq 2$ ‘column’ categories, both considered to be randomly selected from underlying populations, and suppose there are $m \geq 1$ observations per cell. The main question now is to decide whether or not the row and/or column effects vary in the population.

Statistical model. In its most general form, including interaction terms, this is given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \quad (12)$$

with $i = 1, \dots, r$, $j = 1, \dots, c$, $k = 1, \dots, m$, where the α_i are independent $N(0, \sigma_\alpha^2)$, β_j are independent $N(0, \sigma_\beta^2)$ and γ_{ij} are independent $N(0, \sigma_{\alpha\beta}^2)$. The relevant null hypotheses are $\sigma_\alpha^2 = 0$ (no row effect), $\sigma_\beta^2 = 0$ (no column effect) and $\sigma_{\alpha\beta}^2 = 0$ (no interaction).

Analysis: partition of the total sum of squares. Again, the algebraic partition is exactly as in the fixed-effects case: $S_T = S_R + S_C + S_{RC} + S_E$. The difference is in the expected mean squares, which are now

$$\begin{aligned} E(\text{m.s.}_R) &= \sigma^2 + m\sigma_{\alpha\beta}^2 + cm\sigma_\alpha^2 \\ E(\text{m.s.}_C) &= \sigma^2 + m\sigma_{\alpha\beta}^2 + rm\sigma_\beta^2 \\ E(\text{m.s.}_{RC}) &= \sigma^2 + m\sigma_{\alpha\beta}^2 \\ E(\text{m.s.}_E) &= \sigma^2 \end{aligned}$$

The appropriate mean-square ratios for testing the three hypotheses can be deduced from these expected values. To test the “no-interaction” hypothesis $H_0 : \sigma_{\alpha\beta}^2 = 0$, we compare m.s._{RC} with m.s._E . This is exactly as in the fixed-effects case. However, to test either of the “no main effects” hypotheses $H_0 : \sigma_\alpha^2 = 0$ or $H_0 : \sigma_\beta^2 = 0$, we must compare m.s._R (resp. m.s._C) with the *interaction* m.s. , m.s._{RC} . This is different from the fixed-effects case. An informal explanation of this difference is that in the random-effects case, it is possible that chance differences between groups could arise simply due to the sampling of interaction terms from the population; in the fixed-effects case, this is not possible since all of the interaction terms are constrained to sum to zero.

Analysis: estimation of the parameters. The least-squares estimator of μ is $\hat{\mu} = \bar{y}_{..}$, as in the fixed-effects case. For the variance components, moment estimators may be derived from the expected mean squares, as in the single factor case. For example, we have $\hat{\sigma}^2 = \text{m.s.}_E$ and $\hat{\sigma}_{\alpha\beta}^2 = [\text{m.s.}_{RC} - \hat{\sigma}^2] / m$.

Using R: Example 10. The s.s and d.f. partition are constructed just as in the fixed-effects case. To save some work we will just give the R output, which is as follows:

```
> fit <- aov(Size ~ Operator * Part, capability.data)
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Operator	1	0.417	0.417	0.2778	0.6011
Part	9	99.017	11.002	7.3346	3.216e-06 ***
Operator:Part	9	5.417	0.602	0.4012	0.9270
Residuals	40	60.000	1.500		

The test supplied here for the interaction effect is correct. However, those for the main effects are incorrect since the `aov` command assumes fixed effects and hence compares all mean squares with those for the residuals. (There are R packages that have explicit routines for fitting random-effects models of this type; for example, `lme4`.) If, instead, we compare the main effects with the interaction m.s., we obtain F values of 0.693 and 18.276 on (1, 9) and (9, 9) d.f. respectively. Thus the interaction and operator effects are not significant, but the variation between parts is highly significant.

The moment estimates of the variance components can be derived from the mean squares as

Source	Variance component
Operators	-0.0062
Parts	1.7333
Operators×parts	-0.2993
Error	1.5000

Note the negative estimates! (Reported here only for illustrative purposes - never present a negative value as an estimate of a variance.) If we are prepared to believe that the operators and interaction variation is negligible, then fitting only parts (*i.e.* a one-way random effects model) produces an estimated error variance 1.317 and parts variance of $1.6142 = (11.002 - 1.317)/(2n)$:

```
> fit <- aov(Size ~ Part, capability.data)
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Part	9	99.017	11.002	8.3558	1.802e-07 ***
Residuals	50	65.833	1.317		

2.8.3 Two-way mixed-effects model

We could envisage situations in which there are two factors of interest, but now one of them is considered to be fixed and the other random. Without loss of generality, we consider the ‘row’ factor to be fixed. In such a situation, we may wish to decide whether or not the column effects vary in the population, and whether or not the row means differ.

Statistical model: this is, once again, given by equation (12), but now the fixed α_i are subject to the constraint $\alpha_{\cdot} = 0$, the β_j are independent $N(0, \sigma_\beta^2)$ as before and the γ_{ij} are jointly normally distributed but subject to the constraints $\gamma_{\cdot j} = 0$ (and hence not independent).

Analysis: partition of the total sum of squares. This should be familiar by now! The only difference from the fixed-effects case is in the expected mean squares, which are

$$E(\text{m.s.}_R) = \sigma^2 + m\sigma_{\alpha\beta}^2 + \frac{cm \sum_i \alpha_i^2}{(r-1)} \quad (\text{fixed effect})$$

$$E(\text{m.s.}_C) = \sigma^2 + rm\sigma_\beta^2 \quad (\text{random effect})$$

$$E(\text{m.s.}_{RC}) = \sigma^2 + m\sigma_{\alpha\beta}^2$$

$$E(\text{m.s.}_E) = \sigma^2$$

To test the random effects, $H_0 : \sigma_{\alpha\beta}^2 = 0$ and $H_0 : \sigma_\beta^2 = 0$, we compare m.s._{RC} and m.s._C respectively with m.s._E as in the fixed effects case. However, to test the fixed effect $H_0 : \alpha_1, \dots, \alpha_r = 0$, we must compare m.s._R with the interaction m.s. , m.s._{RC} .

Analysis: estimation of the parameters. The least-squares estimators of μ and α_i are as in the fixed-effects case. For the variance components, moment estimators may be derived from the expected mean squares.

2.8.4 More general models

In experiments with three or more factors, and more complex designs involving random or mixed effects, there are often no exact test statistics for certain effects. In such cases approximate F tests can sometimes be used. Alternatively, and preferably, maximum likelihood methods may be used, usually using restricted or residual maximum likelihood (REML). The issues involved are often quite subtle and specialised software may be required (e.g. MLwiN, or lmer in the R package lme4). These methods will not be pursued further in the present course.

2.9 Nested designs

Up to now the various factors in an experiment have been considered to be **crossed**: every level of one factor occurs with every level of the others, *e.g.* every treatment in each block. Sometimes, however, some factors can only appear within other factors, *e.g.* blocks with sub-blocks. Such factors are said to be **nested**.

Suppose, for example, that we wish to compare two treatments for asthma, formoterol and salmeterol ('molecules'). Formoterol has three formulations (solution, single-dose dry-powder inhaler, and multidose dry-powder inhaler) whereas salmeterol has two (suspension, multi-dose dry powder inhaler). From one point of view, there are five possible treatments here, defined by the combination of molecule and formulation. However, we may be interested in a **hierarchy** of treatments: firstly to compare the molecules, and then to compare the formulations within

molecules. Notice that the factor “formulation” has to be interpreted differently depending on the molecule involved.

It is often the case that one or more factors in a nested design is considered as random (see §2.8) and this will affect the form of the test required.

Two-stage nested design. We consider the situation where factor B is nested within the levels of factor A. In this case the “main effect of B” has no meaning on its own (think about the formulation/molecule example above), so a slightly different statistical model is required.

Statistical model. For a balanced design (b levels of B within each level of A, m replicates of each combination), this is given by

$$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$$

with $i = 1, \dots, a$, $j = 1, \dots, b$, $k = 1, \dots, m$. Here β_{ij} is the effect of the j th level of B within the i th level of A. Notice that there is no main effect for B. The usual constraints apply when the factors are fixed, specifically for β_{ij} we use $\sum_{j=1}^b \beta_{ij} = 0$. Hypotheses to test are $\alpha_1 = \dots = \alpha_a = 0$ and $\beta_{ij} = 0$ for all i, j .

Analysis: partition of the total sum of squares. Each observation can be written as

$$y_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..}) + (y_{ijk} - \bar{y}_{ij.})$$

(compare this with the corresponding decomposition in §2.4.5). The total s.s. may therefore be partitioned as

$$\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2 = bm \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 + m \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..})^2 + \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2$$

or

$$S_T = S_A + S_{B(A)} + S_E .$$

The d.f. and expected mean squares are as follows:

Source	d.f.	Expected mean squares		
		A fixed B fixed	A fixed B random	A random B random
A	$a - 1$	$\sigma^2 + \frac{bm \sum_i \alpha_i^2}{a-1}$	$\sigma^2 + m\sigma_\beta^2 + \frac{bm \sum_i \alpha_i^2}{a-1}$	$\sigma^2 + m\sigma_\beta^2 + bm\sigma_\alpha^2$
B in A	$a(b - 1)$	$\sigma^2 + \frac{m \sum_i \sum_j \beta_{ij}^2}{a(b-1)}$	$\sigma^2 + m\sigma_\beta^2$	$\sigma^2 + m\sigma_\beta^2$
E	$ab(m - 1)$	σ^2	σ^2	σ^2

Note the following:

- The appropriate statistics for testing the effects of the factors depend on whether A and B are fixed or random. If A is random, B must also be random (since the levels of B must be interpreted with reference to the levels of A).
- The s.s. associated with B in A is the sum of the B and A×B (i.e. interaction) s.s. in a decomposition that ignores the nested structure (to see this, notice that the s.s. for A and for E are the same here as in the non-nested case).
- Least-squares estimators of fixed effects are obtained in the usual way.
- When random effects are present, the variance components may be estimated by inspection of the expected mean squares above, as in §2.8.

Example 11 (Montgomery, Example 14.1): A company buys raw material from three different suppliers. The purity of this material varies considerably, and it is of interest to determine whether this variation is attributable to differences between the suppliers. To investigate this, four batches of raw material are selected at random from each supplier and three determinations of purity (in %, presumably) are made on each batch. After coding the data by subtracting 93 from all observations, the results are as follows:

Supplier 1				Supplier 2				Supplier 3			
Batch				Batch				Batch			
1	2	3	4	1	2	3	4	1	2	3	4
1	-2	-2	1	1	0	-1	0	2	-2	1	3
-1	-3	0	4	-2	4	0	3	4	0	-1	2
0	-4	1	0	-3	2	-2	2	0	2	2	1

This arrangement is a two-stage nested design. For the purposes of this experiment, ‘supplier’ is a fixed factor (the company is only interested in these three specific suppliers) but ‘batch within supplier’ is random (the company is also interested in other batches besides those used for this experiment). Note that it would be wrong to analyse these data as a two-way (mixed effects) ANOVA, since the four batches are not the same for the three suppliers.

Using R: Example 11. Rather than construct the ANOVA table manually, we will examine the output from R. The notation used for a factor B nested within A is `B%in%A`. Furthermore, `A/B` can be used to specify `A + B%in%A`:

```
> fit <- aov(Purity ~ Supplier/Batch, purity.data)
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Supplier	2	15.056	7.528	2.8526	0.07736

```
Supplier:Batch  9 69.917   7.769  2.9439 0.01667 *
Residuals      24 63.333   2.639
```

Note the following:

- The tests here are for fixed effects. Thus the test for batches within suppliers is correct (significant differences between batches) but that for suppliers is incorrect. Instead, comparing with the Supplier:Batch m.s. we obtain $F = 7.528/7.769 = 0.97$ on (2, 9) d.f., which is not significant.
- The correct F -statistic for suppliers is much smaller than that reported in the R output; the latter suggested that the supplier effect was almost significant. So, if we incorrectly regarded batches as fixed, we might draw the erroneous conclusion that the variation in purity is partly due to differences between the suppliers. Actually, the correct analysis shows that it is due to the inconsistency of batches within suppliers.
- We might also reach a wrong conclusion if we incorrectly analysed these data as arising from a factorial experiment (`aov(Purity ~ Supplier*Batch, purity.data)`).

2.10 Split-plot designs (*Optional*)

Another example of a type of nested design is the **split-plot** design. This occurs in factorial experiments when we may not be able to completely randomise the order of runs and leads to a generalisation of the factorial design.

Consider the following example, taken from Montgomery (§14.4): a paper manufacturer is interested in the effects of three different pulp preparation methods and four different cooking temperatures. An experiment is to be conducted over a three-day period. On each day, three batches of pulp are produced (one by each method). Each batch is then divided into four samples, each of which is cooked at a different temperature. For each sample, the tensile strength of the paper is measured.

In this experiment, we have three replicates of a 3×4 factorial arrangement, but these take place on different days. Clearly, we should regard the days as blocks (and probably treat the block effect as random). If the order of experimentation within each day is completely randomised, then the data can be analysed as a 3×4 factorial arranged in a randomised block design. However, for reasons of experimental efficiency, it would probably be more practical to deal with one batch of pulp at a time. We would therefore divide each block (day) into three **whole plots**, to which the pulp preparation methods would be randomly assigned. Then, each whole plot would be subdivided into four **subplots** (or **split-plots**), and the temperature levels randomly assigned to these. Thus the preparation methods are the **whole plot treatments** (or **main treatments**) and temperature is the **subplot treatment**. The terminology comes from agriculture, where whole plots are large areas of land and subplots are smaller areas within the large areas.

One way to think of split-plot designs is that there are two levels of blocking: at the first level, whole plot treatments are applied randomly within blocks, and at the second level the subplot treatments are applied within whole plots. In the example above there are three factors in total: day (block), preparation method (main treatment) and temperature (subplot treatment). An ANOVA table, showing the sums of squares and degrees of freedom for each factor and 2- or 3-way interactions, can be computed in the usual way. For testing purposes, however, care is required to identify what mean squares should be compared to test the various hypotheses of interest. The precise comparisons depend on which factors are regarded as fixed and which are random, but the basic principle is to compare the main treatment m.s. with the **whole plot error** m.s. (represented by an interaction between blocks and whole plots); and to compare the subplot treatment m.s. with the residual m.s. For further details, see Section 14.4 of Montgomery.

2.11 Incomplete blocks (*Optional*)

It is not always possible to test all possible treatments within a block (for example, due to constraints of time or batch size). In such situations, an **incomplete block design** must be used instead. Here, each block receives only a subset of the treatments.

The Latin square design is an example of an incomplete block design. In this case only one treatment occurs within each row-column combination. In factorial experiments, provided that not all interactions are of interest, we can arrange the experiment in incomplete blocks by **confounding** (mixing up) the unwanted interactions with the block effects (see course STAT0012 *Factorial Experimentation*). The **balanced incomplete block design** is another such arrangement. Here, ‘balance’ means that every pair of treatments occurs together the same number of times as any other pair. An advantage is that if all the blocks are the same size, then all possible pairs of treatment differences can be estimated with the same precision (this is, of course, only an advantage if all such pairwise comparisons are equally interesting!). The statistical model is just the two-way model (10). However, for incomplete block designs the decomposition of the total sum of squares is generally not unique; usually, the treatment s.s. will have to be adjusted before carrying out the usual F -test.

If a design is unbalanced, the consequences have already been discussed in §2.4. By and large, for most ‘sensible’ designs (in which, for example, the lack of balance is not too severe) there are no major difficulties except that hand calculation of results may be more complicated. Nowadays, however, with modern computing power, the data from a huge range of experimental designs can be analysed easily as a **general linear model**. One could argue that all of those ANOVA tables are now redundant!

2.12 The general linear model

In its general form, the linear model may be written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (13)$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ is the vector of responses, $\mathbf{X}$ is an $n \times p$ **design matrix**, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is the vector of regression coefficients and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ is an error vector with $E(\boldsymbol{\epsilon}) = \mathbf{0}$. The standard assumptions are that $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \sigma^2 \mathbf{I}_{n \times n}$ (homoscedastic, uncorrelated errors) and the $\{\epsilon_i\}$ are normally distributed. We can incorporate qualitative factors having two or more levels by introducing suitable dummy variables as columns of the design matrix $\mathbf{X}$. For example, the one-way model of §2.3.1 can be written in the form (13) with $\mathbf{Y} = (Y_{11}, \dots, Y_{1n_1}, \dots, Y_{k1}, \dots, Y_{kn_k})^T$, $\boldsymbol{\beta} = (\mu, \alpha_1, \dots, \alpha_k)^T$ and

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_0 & \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & 1 & 0 & \cdots & 0 \\ \hline \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \hline 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}. \quad (14)$$

That is, we introduce k dummy variables $x_1, \dots, x_k$, where $x_j = 1$ when the factor is at its j th level and is zero otherwise, and set $x_0 = 1$. Then $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$, which is the one-way model (5). Note that there is a linear dependence between the columns of $\mathbf{X}$, since $\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_k = \mathbf{x}_0$. This dependence corresponds exactly to the lack of identifiability discussed in §2.3.1. When working with the linear model algebraically, the easiest way to remove the dependence is to delete a column of $\mathbf{X}$. For example, if we drop the $\mathbf{x}_1$ column then the ‘intercept’, μ , will represent the mean for the first group and the remaining parameters will represent differences relative to this first group. This is the corner-point parameterisation mentioned previously. So far in the course, however, we have worked mainly with the ‘sum-to-zero’ constraint $\sum_i \alpha_i = 0$. Notice that in this case we can write the final parameter $\alpha_k = -\sum_{i=1}^{k-1} \alpha_i$. In the linear model, sum-to-zero constraints can therefore be implemented by deleting the last column of $\mathbf{X}$ in (14) and replacing each of the last n_k rows with $(1 \ -1 \ -1 \ \cdots \ -1)$.

If the entries of $\mathbf{X}$ are the values of one or more **covariates**, then we obtain a regression model. The principle of the general linear model makes it easy to allow for the effect of covariates within a standard experimental design. We give two examples in §2.12.2 and §2.12.3.

2.12.1 Theory of the linear model: outline

The least squares estimators of β are obtained by minimising the residual sum of squares

$$S^2 = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta).$$

Solving $dS^2/d\beta = \mathbf{0}$ gives the **normal equations** $\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{Y}$. If $\mathbf{X}^T \mathbf{X}$ is nonsingular then the unique solution is

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

Note that $\mathbf{X}^T \mathbf{X}$ is nonsingular if and only if there are no linear dependencies between the columns of $\mathbf{X}$. It follows that, under the standard linear model assumptions,

$$E(\hat{\beta}) = \beta, \quad \text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

and the residual sum of squares is $S_0^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})^T (\mathbf{Y} - \mathbf{X}\hat{\beta})$. The expression for the covariance matrix makes it easy to extract standard errors for the model parameters — thus the linear model could be used to test the hypothesis $H_0 : \alpha_1 = 0$ instead of partitioning the treatment sum of squares in Example 7 (Section 2.4), for example. Notice that the parameter estimates will be uncorrelated if $\mathbf{X}^T \mathbf{X}$ is diagonal, which itself will occur if the columns of $\mathbf{X}$ are orthogonal (effectively, if the covariates are themselves uncorrelated).

Standard errors of contrasts can also be calculated easily, since any set of ℓ contrasts can be written as $\mathbf{L} = \mathbf{\Lambda}\beta$, where $\mathbf{\Lambda}$ is an $\ell \times p$ matrix whose row sums are all zero. The least squares estimate of $\mathbf{L}$ is $\hat{\mathbf{L}} = \mathbf{\Lambda}\hat{\beta}$, and its covariance matrix is $\sigma^2 \mathbf{\Lambda} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{\Lambda}^T$.

The extra sum of squares principle. To test a **linear hypothesis** of the form $H_0 : \mathbf{H}\beta = \phi$, where $\phi = (\phi_1, \dots, \phi_r)$ is a given vector, $\mathbf{H}$ is an $r \times p$ matrix with $r \leq p$ and there are no linear dependences between the equations, we use the sum of squares $S_0'^2 - S_0^2$, where $S_0'^2$ is the residual sum of squares under H_0 . Then $S_0'^2 - S_0^2$ with r degrees of freedom is the “extra sum of squares” accounted for by fitting the unrestricted model, and an exact F -test can be carried out based on the statistic

$$\frac{(S_0'^2 - S_0^2)/r}{S_0^2/(N - p)},$$

where N is the total number of observations. The most common use of this technique is to test whether the last $p - q$ regression coefficients are zero in a multiple regression: in this case the

system of equations $\mathbf{H}\boldsymbol{\beta} = \boldsymbol{\phi}$ is simply $\beta_{q+1} = \beta_{q+2} = \dots = \beta_p = 0$. This provides a very convenient way of, for example, testing for treatment effects in an incomplete block design: we start by fitting a one-way model for the block effects, then fit a model including both block and treatment effects and carry out an F -test based on the change in sum of squares. In the balanced case this does, of course, deliver exactly the same result as the two-way ANOVA already studied.

The class of linear models can be extended to accommodate features such as random effects. For more details of the linear model and its variants, see course STAT0028 *Statistical Models and Data Analysis*.

2.12.2 Example 12: regression of means in the one-way model

In an experiment to determine the yield of a chemical reaction, five replicates were carried out at each of four different temperatures. In coded form, the data were as follows:

Temperature	500°C	0	-3	-8	-5	2
	525°C	-9	0	2	1	-6
	550°C	6	5	-2	1	-2
	575°C	-2	3	2	5	2

This is a one-way layout, and could be analysed as such. However, the levels of the ‘factor’ temperature are quantitative here. It is probably not of much interest to see if the mean yields differ at different temperatures (the chemist probably knows that already) — rather, (s)he will be interested in the form of the *relationship* between yield and temperature. In particular, is the relationship linear?

To address this question we use the extra sum of squares principle. We fit both the full (unrestricted) one-way model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

with, for example, $\sum_i \alpha_i = 0$, and the (restricted) straight-line model

$$y_{ij} = \alpha + \beta x_i + \epsilon_{ij}$$

where x_i is the value of the quantitative factor in the i th group. The restricted model can be obtained from the full model by imposing the constraints $\alpha_4 - \alpha_3 = \alpha_3 - \alpha_2 = \alpha_2 - \alpha_1$, which can be rearranged into a linear hypothesis of the form

$$\begin{pmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

so that $r = 2$, $p = 4$, and $\mathbf{H}$ is a $r \times p$ matrix — in the notation above.

The extra sum of squares is the difference in the residual sums of squares from the two models. This measures departure from linearity.

Equivalently we could define the unrestricted model to be

$$y_{ij} = \mu + \beta x_i + \alpha_i + \epsilon_{ij},$$

with appropriate restrictions, since there are now two linear dependencies.

Using R: Example 12.

```
> Group <- factor(reaction.data$temperature)
> fit1 <- aov(yield ~ Group, reaction.data)
> summary(fit1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	3	97.600	32.533	2.1761	0.1307
Residuals	16	239.200	14.950		

```
> fit2 <- aov(yield ~ temperature, reaction.data)
> summary(fit2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
temperature	1	84.640	84.640	6.0419	0.02434 *
Residuals	18	252.160	14.009		

```
> fit3 <- aov(yield ~ temperature + Group, reaction.data)
> summary(fit3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
temperature	1	84.64	84.64	5.6615	0.03013 *
Group	2	12.96	6.48	0.4334	0.65566
Residuals	16	239.20	14.95		

```
> fit4 <- aov(yield ~ Group + temperature, reaction.data)
> summary(fit4)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	3	97.600	32.533	2.1761	0.1307
Residuals	16	239.200	14.950		

Note that the factor `Group` is not significant in `fit1`, but this does not take into account the quantitative nature of this factor (this does not contradict the other fits: remember that ‘not significant’ is not the same thing as ‘no effect’). `fit2` gives a significant linear relationship between mean yield and temperature. The extra sum of squares is $252.160 - 239.200 = 12.96$, which measures departure from linearity. The test statistic is $(12.96/r)/(239.20/(N-p)) = 0.433$ for $r = 2$, $N = 20$, and $p = 4$, which can be assed by using the distribution $F(2, 16)$.

The above can also be obtained directly from `fit3`. Thus there is no evidence of any departure from linearity here.

Note that `fit4` simply gives the analysis of the one-way model (`fit1`). So - if you want to investigate departure from linearity in R, then make sure that you include the regression term before the factor term!

If we are interested in estimation, we can use `coef(summary.lm(fit2))` and derive the equation for predicting yield from temperature:

$$\hat{Y} = -39.96 + 0.0736x.$$

The usual diagnostic plots should, of course, be carried out.

2.12.3 Example 13: parallel regression and the analysis of covariance

Twenty rats were randomly assigned to four different treatment groups. Each group was given a different food type and, for each rat, the amount of food taken was measured in terms both of calorie intake and weight gain after a certain length of time. The researchers were primarily interested in the effect of food type upon weight gain for a given calorie intake. The data, in coded form, were as follows (x denotes calorie intake, y denotes weight gain):

Food type							
1		2		3		4	
x	y	x	y	x	y	x	y
29	22	15	30	16	12	5	23
20	22	9	32	31	8	25	25
14	20	1	26	26	13	16	28
21	24	6	25	35	25	10	26
6	12	19	31	12	7	24	23

This experiment has the form of a one-way layout (one factor, four treatment groups) but now two response variables are measured. Since the primary interest is in weight gain, we need to ‘adjust’ the y ’s for calorie intake x .

An initial scatter diagram of weight gain against calorie intake (not shown here) suggests that there are differences between groups but that, as a first assumption at least, the relationship may be linear within each group. This can be represented by a full (unrestricted) regression model of the form

$$y_{ij} = \mu + \alpha_i + (\beta + \gamma_i)x_{ij} + \epsilon_{ij}.$$

If the four regression lines are parallel then we can conclude that differences between foods, in terms of effect on weight gain, are the same regardless of calorie intake. It is therefore natural to

see first if the data are consistent with parallel regressions (**parallel regression analysis**) i.e. to test $H_0 : \gamma_1 = \dots = \gamma_4 = 0$, where γ_i is the interaction effect on the slope of the regression of y on x for food type i . This corresponds to a restricted model of the form

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij}$$

Finally, if parallel regressions can be assumed then we can look for differences between the four food types by testing for common intercepts (**analysis of covariance**): now the null hypothesis is $H_0 : \alpha_1 = \dots = \alpha_4 = 0$, where α_i is the effect of the regression for food type i . The corresponding model is

$$y_{ij} = \mu + \beta x_{ij} + \epsilon_{ij}.$$

The terminology dates back to before the time when linear models were routinely used for such calculations, and is perhaps falling into disuse.

Using R: Example 13. To test $H_0 : \gamma_1 = \dots = \gamma_4 = 0$, we calculate the difference between the residual sums of squares of the first two models as follows:

```
> options(contrasts=c("contr.treatment", "contr.treatment"))
> fit1 <- aov(weight.gain ~ food.type * calories, weightgain.data)
> summary(fit1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
food.type	3	699.40	233.13	14.8606	0.0002414 ***
calories	1	118.34	118.34	7.5432	0.0177169 *
food.type:calories	3	44.21	14.74	0.9393	0.4520537
Residuals	12	188.26	15.69		

The F statistic for the interaction term here indicates no significant departure from a parallel regression model. We can therefore proceed by assuming parallel regressions and testing $H_0 : \alpha_1 = \dots = \alpha_4 = 0$:

```
> fit2 <- aov(weight.gain ~ food.type + calories, weightgain.data)
> summary(fit2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
food.type	3	699.40	233.13	15.043	8.595e-05 ***
calories	1	118.34	118.34	7.636	0.01449 *
Residuals	15	232.46	15.50		

```
> fit3 <- aov(weight.gain ~ calories + food.type, weightgain.data)
> summary(fit3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
--	----	--------	---------	---------	--------

```

calories      1  24.74   24.74  1.5965    0.2257
food.type     3 793.00  264.33 17.0565 4.246e-05 ***
Residuals    15 232.46   15.50

```

The result for `fit2` indicates that the regression coefficient is significant, so there is an underlying linear relationship between gain in weight and calorie intake for each food type, with common slope. `fit3` indicates a highly significant difference in mean gain in weight between the four food types, having adjusted for calorie intake. Note that we are using the `summary` command here as a convenient means of examining different reduced models — the last line in each ANOVA table corresponds to the hypothesis of interest. The reason that `fit2` and `fit3` give different sums of squares is that the effects of calorie intake and food type are not orthogonal.

Finally, the parameter estimates for `fit3` :

```
> summary.lm(fit3)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	14.3046	2.7106	5.277	9.3e-05	***
calories	0.3164	0.1145	2.763	0.01449	*
food.type2	11.3313	2.6529	4.271	0.00067	***
food.type3	-8.8985	2.5828	-3.445	0.00361	**
food.type4	5.6328	2.5003	2.253	0.03967	*

Note that the same *estimates* would be obtained if we interchanged the order of `calories` and `food.type` — it is only the extra sums of squares that differ.

From the output above, we see that the predicted gain in weight using food type i of a rat with calorie intake x is

$$\hat{Y} = \hat{\alpha}_i + 0.3164x,$$

where the intercept terms $\hat{\alpha}_i$ are 14.3046, 25.6359, 5.4061 and 19.9374 respectively.

2.13 Departures from assumptions

The models described so far have usually included three key assumptions: homoscedasticity (constant error variance), additivity of effects, and normality of error. One or more of these assumptions may be inappropriate, either from modelling considerations or empirically from residual plots. There are two main approaches to dealing with this: by suitable transformation of variables or by using a more appropriate model.

2.13.1 Transformation of variables

1. To give *additivity*. For example, if $Y_{ij} = \mu\alpha_i\beta_j\epsilon_{ij}$ (multiplicative model) then additivity can be achieved by taking logs. Notice that this model could equivalently be written as $Y_{ij} = \mu\alpha_i\beta_j + \delta_{ij}$, where $\delta_{ij} = \mu\alpha_i\beta_j(\epsilon_{ij} - 1)$. This has the form of an “interaction only” model which could be fitted using the kinds of techniques discussed already. Notice, however, that the error term in this formulation has variance $\text{Var}(\delta_{ij}) = \mu^2\alpha_i^2\beta_j^2\text{Var}(\epsilon_{ij})$, which is unlikely to be constant — taking logs is usually a much more sensible way of dealing with this particular situation.
2. A remedy for *homoscedasticity* is to use transformations, weighted least squares, or (sometimes) robust linear regression.
3. To give *normality*. For example, in (1) above it is likely that Y , and hence ϵ , is a nonnegative r.v. and so cannot be normal. Taking logs will nearly always improve normality for non-negative variables.

It is not always possible to find a transformation that simultaneously achieves all three goals, but sometimes a transformation that stabilises variance will lead to approximate additivity and normality as well. It turns out that we can often find a transformation that does a remarkably good job of variance stabilisation, by arguing as follows: given a random variable Y with expectation μ , we require some function $g(Y)$ whose variance is roughly the same for all values of μ . Suppose the variance is itself a function of μ (this is the case for many common distributions, as we will see below): then we can write $\text{Var}(Y) = V(\mu)$, say. Now a first-order Taylor series expansion of $g(Y)$ about μ yields

$$g(Y) \simeq g(\mu) + (Y - \mu)g'(\mu) .$$

Taking expectations, we obtain

$$E[g(Y)] \simeq g(\mu) .$$

The variance of $g(Y)$ is defined as $E[g(Y) - E[g(Y)]]^2$. Approximate this by $E[(Y - \mu)g'(\mu)]^2 = E[(Y - \mu)^2][g'(\mu)]^2 = V(\mu)[g'(\mu)]^2$. It follows that if we want $g(Y)$ to have approximately constant variance, we should choose g such that $g'(\mu) = V^{-1/2}(\mu)$. Integrating, we find that an approximate variance-stabilising transformation is

$$g(y) = \int \{V(y)\}^{-1/2} dy .$$

Some particular examples of this result are as follows:

- If $V(\mu) \propto \mu$, the variance-stabilising transformation is $\sqrt{Y}$, the square-root transformation. This applies, for example, if Y has a Poisson distribution with a reasonably large mean.

- If $V(\mu) \propto \mu^2$ (e.g. for the exponential distribution), the variance-stabilising transformation is $\log Y$, in agreement with the earlier discussion.
- When the Y 's are *proportions* obtained from a binomial model, $V(\mu) \propto \mu(1 - \mu)/n$ and the variance-stabilising transformation is found to be $\sqrt{n} \sin^{-1} \sqrt{Y}$, the **angular transform**. Unsurprisingly, this transformation does not work very well when the n s are very small, but it is good when the n 's are large.

2.13.2 The Box-Cox transformation (*Optional*)

Generally, $V(\mu)^{1/2} \propto \mu^{1-\lambda}$, the variance-stabilising transformation is the **power transformation**

$$g(Y) = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log Y & \lambda = 0 \end{cases} . \quad (15)$$

This is called the Box-Cox transformation, after its initial proponents. If we assume that a power transformation of the form (15) will simultaneously achieve the objectives of variance stabilisation, additivity and normality, then the transformation parameter λ may be estimated simultaneously with the other model parameters, by maximum likelihood, for example. This approach has the advantage of giving a range of plausible values for λ , by constructing an approximate confidence interval. The method is implemented in R using the `boxcox` command, which is available as part of the **MASS** library.

Of course, there are also situations where the variance is non-constant but where it is not related to the mean. For example, it may be harder to measure accurately in some blocks than in others so that the precisions of observations vary between blocks. If the setup is such that $\text{Var}(Y_{ij}) = \sigma^2/w_{ij}$, and the relative precisions $\{w_{ij}\}$ are known, then weighted least squares can be used to estimate the model parameters: the usual least-squares objective function is changed to $\sum_{i,j} w_{ij} (Y_{ij} - \mu_{ij})^2$, in an obvious notation.

2.13.3 Generalised linear models

A more radical and usually more satisfying way to deal with departures from assumptions is to change the model — it is better to write down a model that genuinely represents the data than to change the data to fit the model! If the distribution of the response variable is a member of the **exponential family**, we obtain the class of **generalised linear models** (GLMs). The exponential family includes common distributions such as the normal, Poisson, binomial, exponential and gamma distributions. In the normal case, the linear model specifies a linear structure for $E(Y)$. In a GLM this linear structure is specified for some function g of $E(Y)$, called the **link function**. To use a GLM the response distribution and link function must be specified. In particular, this approach does not involve approximations (as does the transformation approach)

for discrete response variables — for example, if Y is binary then we can use logistic regression. Of course, all the experimental designs discussed in this chapter may be used for non-normal data.

2.14 Optimal design

We have discussed some principles for determining optimal sample sizes and optimal allocation of sample units in some simple designs. In this section we discuss some general principles for selecting optimal designs, including not only sample sizes but also possibly the arrangement of the experimental units. Much of the theory of optimal design is concerned with designing for parameter estimation. We give a brief overview in the case of the normal linear model (13). Assume for simplicity that σ is known. In that case the covariance matrix of the estimated regression coefficients is

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1},$$

and we seek to minimise this in some sense. Two criteria that we might consider are as follows:

c-optimality. Let $\mathbf{c} = (c_1, \dots, c_k)^T$ be a vector of coefficients and suppose we are interested in estimating the linear combination $\mathbf{c}^T \boldsymbol{\beta}$ of the regression coefficients. Then

$$\text{Var}(\mathbf{c}^T \hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{c}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{c}$$

and we seek to minimise this quantity with respect to $\mathbf{X}$.

D-optimality. Suppose instead that we are interested in estimating the entire regression parameter vector. A $100(1 - \alpha)\%$ confidence region for $\boldsymbol{\beta}$ is the set for which $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (\mathbf{X}^T \mathbf{X}) (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq \sigma^2 \chi_{\alpha}^2(k)$, which is a k -dimensional ellipsoid centred at $\hat{\boldsymbol{\beta}}$. We might therefore seek to minimise the volume of this ellipsoid, which is proportional to $|\mathbf{X}^T \mathbf{X}|^{-1/2}$. So we choose the design to maximise $|\mathbf{X}^T \mathbf{X}|$. Similarly, if we wish to consider a confidence region for a set of linear combinations, $\mathbf{A}^T \boldsymbol{\beta}$, then this will give rise to minimisation of $|\mathbf{A}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{A}|$ (D_A -optimality).

Other criteria based on $\text{Cov}(\hat{\boldsymbol{\beta}})$ include E -optimality, which aims to maximise the minimum eigenvalue of $\mathbf{X}^T \mathbf{X}$. The motivation for this is to minimise the variance of the least precisely estimated linear combination of all the regression coefficients.

Illustration for the one-way layout. Consider the one-way model (4) of §2.3.1. The design matrix in this case is given by (14) without the first column, giving $\mathbf{X}^T \mathbf{X} = \text{diag}(n_1, n_2, \dots, n_k)$. Therefore

$$\text{Var}(\mathbf{c}^T \hat{\boldsymbol{\beta}}) = \sigma^2 \sum_{i=1}^k \frac{c_i^2}{n_i}.$$

Minimising this with respect to $n_1, \dots, n_k$ subject to $\sum_i n_i = N$ gives $n_i \propto |c_i|$. This is the c -optimal solution. In particular, in the case of two groups if we are interested in the contrast $\mu_2 - \mu_1$ then $c_2 = 1$ and $c_1 = -1$, giving $n_1 = n_2$ (equal group sizes).

For D -optimality, note that $|\mathbf{X}^T \mathbf{X}| = \prod_i n_i$. Maximising this with respect to $n_1, \dots, n_k$ subject to $\sum_i n_i = N$ gives $n_i = N/k$, giving equal group sizes. It can be shown that this is also the E -optimal solution.

3 Observational studies

3.1 Introduction

In the previous chapter we looked at a variety of different experimental designs. In all of these, the experimenter was able to intervene in the studied system by deciding which units received which treatments. Randomisation could therefore be used to ensure that any observed differences between groups could reasonably be attributed to the treatment.

Often, however, experimentation is either impossible or uneconomic. For example:

- Ethical obstacles may prohibit experimentation *e.g.* to assess effect of smoking on lung cancer
- Experimentation may be physically impossible *e.g.* to assess gender differences in academic attainment
- Data on the variables of interest, collected for a different purpose, may already be available
- Experiment may be too costly to mount
- Experiment may take too long, *e.g.* if the effect being studied is rare, or if long-term effects are of interest
- If there are many risk factors then we would need to pre-specify which ones to study (*e.g.* causes of cholera in the nineteenth century)

In such situations, the investigator is restricted to taking selected observations either by gathering new data or using data already gathered. Either way, it is not possible to manipulate the system — only to observe it. Such a study is therefore called an **observational study**, and it is customary to refer to **exposure** to factors of interest rather than to ‘treatment’. Observational studies are often used in human populations. Some examples:

Agent	Response variable(s)
Wearing seat belts	Severity & type of injury
Rise in taxes	Consumer spending & saving
Fluoridation of water	Status of children’s teeth
Smoking of cigarettes	Mortality and morbidity

Although we do not manipulate conditions in observational studies, we may wish to manipulate them in the future — *e.g.* by banning smoking in workplaces.

3.2 Observational studies *versus* experiments

Main differences between observational studies and experiments:

Controlled experiments	Observational studies
–Convenient material	–Representative material
–Treatment groups formed by randomisation	–Exposure groups already determined self-selected, <i>e.g.</i> smokers selected by others, <i>e.g.</i> choice of anaesthetic selected by natural forces, <i>e.g.</i> premature babies
–Sources of variation controlled or randomised ...	–Some sources may be neither controlled nor randomised ...

Sources of variation: In both controlled experiments and observational studies, sources of variation may be considered to fall into one of three classes:

1. Sources whose effects the investigator tries to remove. There are three main methods of doing this:
 - Refinement of measurement techniques, *e.g.* to maintain constant conditions. There is little scope for this in observational studies.
 - Blocking and matching. See §2.4 for blocking in experiments. The corresponding idea in observational studies is called **matching** — attempt to find, for example, pairs of the same age, sex *etc.*
 - **Adjustment**, for example by regression of the response on concomitant variables Z to remove their effects. When used in conjunction with a model for treatment/exposure effects and Z is continuous, the standard method of analysis is the analysis of covariance (*c.f.* §2.12.3).
2. Sources whose effects the investigator handles by randomisation and replication. In experiments, randomisation moves a potential source of variation to the error term, which can be made arbitrarily small by replication. This is rarely available in observational studies, however.
3. Sources whose effects are neither controlled nor randomised. Ideally there should be no sources of this kind in a controlled experiment, but they may be overlooked or we might lose full control (*e.g.* subjects who become ill in a clinical trial might move from the control to treatment group, either through self-selection or on the advice of a clinical investigator). If these effects are related to the treatment differences, they are said to be **confounded** with the treatment effect, giving rise to misleading inferences through **bias**. In a clinical trial, if we carry out an analysis based on the treatments actually received (an ‘as-treated’ or ‘per-protocol’ analysis) then there is a danger of *noncompliance bias*, because differences observed may be associated with factors other than just the treatment. Alternatively we could analyse the results as if individuals actually received the treatment to which they were originally assigned (‘intention-to-treat analysis’). This would still produce biased effects, but

in the conservative direction — the treatment effect would be diluted by the movement of patients between groups.

3.3 Sources of bias in observational studies

3.3.1 Confounding variables

Potential confounding variables are a particular problem in observational studies. This can be illustrated using a simple example.

Example 14: The following is a two-dimensional contingency table relating to the survival of infants and the amount of prenatal care received by infants' mothers.

Amount of Prenatal Care	Died	Survived	Percentage of deaths
Less	20	373	5.1
More	6	316	1.9

Clearly these two variables are highly associated ($p = 0.02$ on a χ^2 test) — but is the amount of prenatal care a contributory factor to survival?

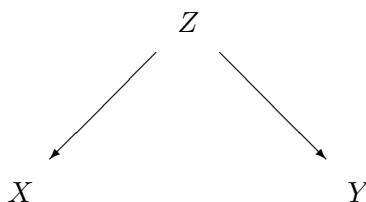
As it happens, the babies were born in two clinics, A and B. This is the three-dimensional contingency table including clinic:

Clinic	Amount of Prenatal Care	Died	Survived	Percentage of deaths
A	Less	3	176	1.7
	More	4	293	1.4
B	Less	17	197	7.9
	More	2	23	8.0

Considering each clinic separately, the amount of prenatal care now seems unrelated to survival! It seems that the important determinant of survival is actually the clinic rather than the amount of prenatal care — but the clinics themselves have different levels of prenatal care, which induces the apparent association in the combined table. Here 'clinic' is a confounding variable. It is in fact possible to construct examples like this, in which the direction of association in the combined table is the reverse of that in *both* of the contributing tables — this is known as **Simpson's paradox** or 'effect reversal'.

The possible presence of confounding variables makes **causal interpretation** difficult in observational studies. If some significant association is found between an exposure variable X and a response Y , a naïve interpretation is that changes in Y are caused, at least in part, by

changes in X . However, it could be that changes in both variables are caused by changes in an unmeasured confounder Z , and that the association is simply a reflection of this common response. In this situation, the causal relationships between the variables would be as indicated in the following diagram:



Even if there is a causal relationship between X and Y (i.e. an arrow pointing from X to Y in the diagram), in general the magnitude of the association will still be affected by the confounder Z . There are two main strategies for dealing with confounders: they can either be controlled in the design of the study (e.g. by holding them constant or by matching — see §3.5.1), or adjusted for in the subsequent analysis (§3.5.2). However, it is rarely possible to be sure that all confounders have been accounted for.

Effects of confounders: to see the possible effects of confounding variables in general, consider the case where two populations are to be compared: ‘exposed’ and ‘unexposed’. Then the exposure variable X is an indicator variable for the ‘exposed’ population and a confounding variable Z can have one of two effects:

1. If the distribution of Z differs systematically between the two populations then Z and X are dependent: if Z also affects the response Y , exposure effects will be biased. This is the case in Example 14.
2. If the distribution of Z is the same in the two populations, then Z and X are independent: in this case, exposure effects will be unbiased. However, note that variation in Z may contribute to the variability in Y (and hence reduce the precision of comparisons between the two populations).

3.3.2 Other sources of bias

A second source of bias in observational studies is **nonresponse** or, more generally, missing data where the missingness could itself be related to the outcome of interest. For example, many individuals with extreme views may be unlikely to respond to a survey about social attitudes. We return to this potentially serious source of bias in §4.2.

A final source of bias is where the exposure variable itself is measured with error. This is called **information bias**. There are two types of information bias: **differential measurement**

error and **non-differential measurement error**. In the former, measurement of the exposure variable is itself biased — for example, people tend to under-report their drinking habits. This will clearly lead to biased estimates of exposure effects. Non-differential measurement error occurs where, although there is no systematic bias in the exposure measurements, they are imprecise. This can still lead to biased estimation of exposure effects. To see this, suppose that the exposure variable X is continuous and that the true relationship between Y and X is linear of the form

$$Y = \alpha + \beta X + \epsilon , \quad (16)$$

where β is the exposure effect and $E(\epsilon) = 0$. Suppose also that X is measured with error:

$$W = X + \eta , \quad (17)$$

where η is independent of X , Y and ϵ and $E(\eta) = 0$. If we regress Y on W rather than upon X , then we are implicitly considering a regression model of the form $Y = \alpha^* + \beta^* W + \epsilon^*$. What is the ‘true’ value of β^* in this model? It seems as though it should be β : substitute $X = W - \eta$ in (16) to obtain $Y = \alpha + \beta W + \delta$, where $\delta = \epsilon - \beta\eta$. However, this argument is incorrect because W and δ are not independent. We need to be a bit more careful.

One way to proceed is to consider what would happen if we had an infinite amount of data and fitted the model $Y = \alpha^* + \beta^* W + \epsilon^*$ by least squares. The resulting parameter estimates would be the “target” values of the least squares estimators for finite samples. It should be clear that these estimates would minimise the expected squared error

$$E \left[(Y - \alpha^* - \beta^* W)^2 \right] .$$

Differentiating with respect to α^* and β^* yields analogues of the normal equations: $E(Y) = \alpha^* + \beta^* E(W)$, and $E(WY) = \alpha^* E(W) + \beta^* E(W^2)$. Writing $E(W) = \mu_W$, $E(Y) = \mu_Y$ and noting that $\mu_Y = \alpha^* + \beta^* \mu_W$ (first equation), we find from the second equation that

$$Cov(W, Y) + \mu_W(\alpha^* + \beta^* \mu_W) = \alpha^* \mu_W + \beta^* (Var(W) + \mu_W^2) .$$

This simplifies to yield $\beta^* = Cov(W, Y)/Var(W)$. This is, in fact, an alternative way to think about regression coefficients in general.

In our case, we have $Cov(W, Y) = Cov(X + \eta, \alpha + \beta X + \epsilon) = \beta Var(X) = \beta \sigma_X^2$, say. Further, we have $Var(W) = \sigma_X^2 + \sigma_\eta^2$, say. Finally, we find that $\beta^* = \beta \sigma_X^2 / (\sigma_X^2 + \sigma_\eta^2) = \gamma \beta$, say. Since $0 \leq \gamma \leq 1$, the effect of measurement error is to underestimate the exposure effect. This phenomenon is known as **regression dilution bias**.

3.3.3 The effect of bias

Let $\bar{d}$ be the difference between two sample means, each based on n observations. The difference between the two population means in general is $\delta^* = \delta + B$, where δ is the true exposure effect and

B a bias term. If a standard significance test for $H_0 : \delta^* = 0$ (ignoring bias) turns out significant, indicating, say, that $\delta^* > 0$, then the conclusion will remain valid provided that $B \ll \bar{d}$, since the observed effect $\bar{d}$ is an unbiased estimator of $\delta + B$ so that $\delta \simeq \bar{d} - B > 0$. Thus we need to assess whether the bias is likely to be smaller than the observed effect $\bar{d}$. *E.g.* in assessing the relationship between smoking and lung cancer, the observed effect is very large and unlikely to be wholly attributable to other confounders.

The effect of bias on confidence statements depends on n and the relative size of B . If there is no bias then a 95% confidence interval for δ is $\bar{d} \pm L$, where $L = 1.96\sigma\sqrt{2/n}$. Thus we need $B \ll L$ for this confidence statement to be approximately correct in the presence of bias. In particular, such statements may be practically useless in large studies where L is very small.

Hence the importance of guarding against bias in the planning and analysis of observational studies, and using judgements about the direction and size of any remaining bias in interpretation of the results of the analysis.

3.4 Planning of observational studies

The main purpose of an observational study is to study associations between a response variable and certain agents, procedures or programmes. Planning should begin with a clear statement of objectives prior to thinking about the following issues.

Measurement of exposure. Should this be detailed? If so, bear in mind possible information bias (§3.3). Alternatively, one may decide to work with just ‘exposed’ and ‘unexposed’ categories. In this case, should try to choose a population with a high level of exposure; *e.g.* to study the relationship between death from lung cancer and number of cigarettes smoked, compare heavy and light smokers. If intermediate levels are used, then need to consider optimum class sizes corresponding to the levels.

Control (comparison) group. In order to avoid confounders, this should be as similar as possible to the exposed group, except that the exposure is absent. Published data on the general population may not be satisfactory if the general population is not representative of the exposed group.

Response variables. These should obviously be relevant to the stated objectives. Care needs to be taken in the measurement of responses:

- When using human raters, a standard precaution against possible bias is for each judge to assess the same proportion of subjects, preferably selected at random from each exposure or control group, and to be ignorant of the group to which a subject belongs (blind rating).
- It may be necessary to decide when the responses should be recorded, depending on whether the effects of exposure are expected to change over time. It is sometimes

helpful to schedule repeated observations, or to use early data to identify a typical pattern of response which may inform the subsequent recording times.

Choice of study design. In an ideal world, design and sample size determination would be based on statistical theory based on power of tests, length of confidence intervals, minimum cost *etc.* (see §2.2.1 and the general theory of §2.14). However, when sampling a population it may be necessary to use more complex sampling schemes for reasons of cost or precision, so more generally we need to consider choice of **sample design**, which is discussed in §4.2. We should also bear in mind the possible effects of bias — there is no point increasing the sample size beyond a level at which bias would start to dominate random sampling error (see above). Some designs may be less sensitive to bias than others.

Pilot studies can provide useful information that helps in the design of the main study.

3.5 Design, analysis and interpretation

We saw earlier that a confounding variable Z may introduce either bias or additional variability into an observational study. It may therefore be worth controlling for Z , either at the planning stage (holding constant or matching) or at the analysis stage (adjustment). The main reason for control, however, is bias removal (contrast this with the situation in experiments, where randomisation is used as a means of removing the effect of unobserved variables and blocking is used to reduce the variation from known confounders)

In practice we need to identify and list, in order of perceived importance in influencing Y , the main confounding variables. Note that a variable Z should not be controlled if its value is influenced by the exposures to be compared ($X \rightarrow Z \rightarrow Y$), otherwise we may be removing part of the overall exposure effect.

3.5.1 Matching

Matching in observational studies is the equivalent of blocking in experiments: we try to find subjects for which the values of confounding variables are as similar as possible, in order to eliminate their effects. The means of achieving this depend on the nature of the confounding variable Z . If Z is categorical (age group, sex, clinic, ...) then matching is easy; with k classified variables, a *match* means belonging to the same cell in the k -way classification. For ordinal variables, however, there are more possibilities:

Within-class matching. Change Z to a classification variable (possibly by grouping) and proceed as above. One might consider choosing the class boundaries so as to achieve maximum reduction of bias under some assumed statistical model for the relationship between Z and Y within each exposure group.

Calliper matching. Here, define a match if two values of Z differ by less than a . This is harder to achieve but tighter than within-class matching. It has the advantage that if a is small, the match is almost perfect and no assumption about the relationship between Y and Z is needed, provided it is of the same functional form in both populations. Suppose, for example, that there are two exposure groups and for an observation in group i we have $E(Y_i|Z) = \mu + \alpha_i + h(Z)$ where $h(\cdot)$ is almost everywhere continuous. Then, in a sample of n **matched pairs** $\{Y_{ij} : i = 1, 2; j = 1, \dots, n\}$, chosen so that $|Z_{1j} - Z_{2j}| < a$, the expected difference in mean response for the two groups is $E(\bar{Y}_1 - \bar{Y}_2) = \alpha_1 - \alpha_2 + \overline{h(Z_1)} - \overline{h(Z_2)}$. But the matching implies that $|\overline{h(Z_1)} - \overline{h(Z_2)}| = O(n^{-1})$ because of the continuity of $h(\cdot)$. Therefore we must have $E(\bar{Y}_1 - \bar{Y}_2) \simeq \alpha_1 - \alpha_2$.

If it is not possible to find a match in the comparison group for all cases within $\pm a$, then ‘nearest available’ pair matching could be used.

Mean matching or balancing does not match individuals, but attempts to match $\bar{z}_1$ and $\bar{z}_2$. This will give unbiased estimates of exposure effects if Y has the same relationship with Z in each population and this relationship is linear (to see this, note that if h is linear then $\overline{h(Z)} = h(\bar{Z})$ and follow the illustration in the previous paragraph).

The biggest problem with matching can be the difficulty of finding matches, especially when there are many Z variables, calliper matching is to be used with small a , large samples are required or the Z distributions differ substantially between exposure groups. In such situations, matching may be feasible only if a large database of potential subjects is available, with matches identified using a computer search.

Statistical analysis of matched samples: Y continuous. If the response variable is continuous, data gathered using within-class or calliper matching can be analysed using the usual ANOVA and linear modelling methods, treating the classes (or individual values) of Z as levels of a factor. Effectively the analysis is the same as for blocked designs, where Z now takes the role of the blocking factor.

Statistical analysis of matched samples: Y binary. Another important case to consider is where the response variable Y is binary. In this case, we could define $p_{ij} = P(Y_{ij} = 1) = 1 - q_{ij}$, where Y_{ij} is the observation for exposure group i in class j . By analogy with the general linear model, it is natural to consider using ideas such as logistic regression to represent the data structure, and to write down a model of the form

$$\log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \log \left(\frac{p_{ij}}{q_{ij}} \right) = \mu + \alpha_i + \beta_j .$$

This is called a **logistic regression model**. An alternative representation of the same model is

$$p_{ij} = \frac{\exp(\mu + \alpha_i + \beta_j)}{1 + \exp(\mu + \alpha_i + \beta_j)} .$$

Notice that in this model, a difference in exposure effects such as $\alpha_1 - \alpha_2$ corresponds to

$$\log \left(\frac{p_{1j}}{q_{1j}} \right) - \log \left(\frac{p_{2j}}{q_{2j}} \right) = \log \left(\frac{p_{1j}/q_{1j}}{p_{2j}/q_{2j}} \right) = \log \psi , \text{ say.}$$

This is called the **log-odds ratio** for comparing the exposure groups.

In principle, it is possible to estimate the parameters in a logistic regression model by maximum likelihood and then to test hypotheses about exposure effects using standard large-sample theory (see courses STAT0028 and STAT0008). However, in the present setting there is a problem with this approach. The logistic regression model is nonlinear, so the justification for likelihood methods is not exact: it is based on arguments that effectively require the sample size to be large relative to the number of parameters estimated (compare this with the general linear model, where the results for effect estimates are exact in all finite samples). However, in a study where, for example, there are 1000 matched pairs (i.e. 2000 observations in total) then there will be 1000 pair effects to estimate in addition to the two parameters of interest α_1 and α_2 . The large number of **nuisance parameters** $\beta_1, \dots, \beta_{1000}$ means that the usual large-sample arguments cannot be used; in fact, in general the maximum likelihood estimators can suffer from severe bias and inconsistency in such situations (this is called the *Neyman-Scott* problem, or sometimes the *incidental parameters* problem).

A neat way to overcome the Neyman-Scott problem is by conditioning on the class totals in forming the likelihood. To illustrate the idea and see why it might work, consider the case of matched pairs $\{(Y_{1j}, Y_{2j}) : j = 1, \dots, n\}$. So, we have n pairs, and two exposure groups. Typically, the two exposure groups are defined by *not exposed* and *exposed*.

For the j th pair, define $T_j = Y_{1j} + Y_{2j}$. T_j takes the values 0, 1 and 2. Notice that, for $i = 1, 2$, $P(Y_{ij} = 1 | T_j = 0) = 0$ and that $P(Y_{ij} = 1 | T_j = 2) = 1$. Furthermore, we have

$$\begin{aligned} P(Y_{1j} = 1 | T_j = 1) &= P(Y_{1j} = 1, T_j = 1) / P(T_j = 1) \\ &= P(Y_{1j} = 1, Y_{2j} = 0) / [P(Y_{1j} = 1, Y_{2j} = 0) + P(Y_{1j} = 0, Y_{2j} = 1)] \\ &= p_{1j}q_{2j} / [p_{1j}q_{2j} + q_{1j}p_{2j}] = \frac{p_{1j}q_{2j}/q_{1j}p_{2j}}{1 + p_{1j}q_{2j}/q_{1j}p_{2j}} = \frac{\psi_j}{1 + \psi_j} , \end{aligned}$$

where ψ_j is the odds ratio for the j th pair. From these results we can write down the joint conditional probability for each pair of observations, given the pair total. This in turn enables us to write down a *conditional likelihood* in terms of the $\{p_{ij}\}$. It is convenient to define $\theta_j = \psi_j / (1 + \psi_j)$.

Then the conditional likelihood function, in terms of the $\{\theta_j\}$, is

$$\begin{aligned} L &= \prod_{t=0}^2 \prod_{T_j=t} P(Y_{1j} = y_{1j}, Y_{2j} = y_{2j} | T_j = t) \\ &= \prod_{T_j=0} 1 \times \prod_{T_j=2} 1 \times \prod_{T_j=1} \theta_j^{y_{1j}} (1 - \theta_j)^{y_{2j}} . \end{aligned}$$

The conditional log-likelihood is therefore

$$\ell = \sum_{T_j=1} [y_{1j} \log \theta_j + y_{2j} \log(1 - \theta_j)] .$$

Now, under the additive logistic model considered earlier, we have $\psi_j = \psi = \exp(\alpha_1 - \alpha_2)$. Since θ_j is a function of ψ_j , the conditional log-likelihood is therefore free of the nuisance parameters $\{\beta_j\}$; and since it represents a *bona fide* log-likelihood function, we can now apply all of the usual likelihood theory. Notice that pairs with identical outcomes contribute nothing to the conditional likelihood. This is basically because the identical outcomes are interpreted as saying more about the pair (i.e. the β) than about the treatment.

Under the additive logistic model, the conditional log-likelihood for the single parameter θ can equivalently be written as

$$\ell = n_{10} \log \theta + n_{01} \log(1 - \theta) ,$$

where n_{10} is the total number of pairs with $y_{1j} = 1, y_{2j} = 0$ and n_{01} is the total number of pairs with $y_{1j} = 0, y_{2j} = 1$. It is straightforward to verify that maximising the conditional log-likelihood with respect to θ leads to the estimate $\hat{\theta} = n_{10} / (n_{10} + n_{01})$. Notice that this is simply the proportion of times that the positive response occurs in exposure group 1, when the responses are different. A binomial confidence interval for θ is therefore easily calculated (assuming $n_{10} + n_{01}$ is large). Since $\psi = \theta / (1 - \theta)$, the conditional likelihood estimate of $\log \psi$ takes the simple form $\log(n_{10}/n_{01})$; confidence intervals for this can be obtained from those for θ .

The above discussion suggests a simple procedure for testing the null hypothesis $H_0 : p_{1j} = p_{2j}$ for all j . In this case, we have $\theta_j = \frac{1}{2}$ for all j . Therefore, given the total number of discordant pairs $n_{10} + n_{01}$, $n_{10} \sim \text{Bin}(n_{10} + n_{01}, \frac{1}{2})$ and a binomial test of H_0 may be performed. This is usually referred to as **McNemar's test** in this context; actually, it is an example of more general class of tests called **sign tests**. For a sufficiently large number of pairs, a normal approximation (with or without continuity correction) may be used, giving rise to a χ_1^2 test statistic (*see exercises*).

In more complex problems, perhaps involving matching on several variables, the conditioning argument above can be extended. The general technique is called **conditional logistic regression** (in R, it is implemented via the `clogit` command in the `survival` library).

3.5.2 Adjustment

Matching may not be feasible or may be difficult (see §3.5.1). Adjustment, which is carried out at the *analysis* stage, is an alternative to matching (which is carried out at the *design* stage).

Suppose that several exposure groups are to be compared and that Y denotes the response and Z the value of potential confounding variables for a given individual. We can distinguish four cases.

Y continuous, Z classified. In this case the data for each group can be arranged into a cross-classified table according to the levels of Z . Suppose that n_{ij} observations fall in cell (i, j) , where $i = 1, \dots, r$ (number of groups), $j = 1, \dots, c$ (number of classes). A linear model for this situation is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad (18)$$

where $k = 1, \dots, n_{ij}$ and $\epsilon_{ijk} \sim N(0, \sigma_{ij}^2)$. Here α_i is the group effect, β_j the effect of Z and γ_{ij} an interaction effect. This is the two-way model (11) of §2.4.5, except that here we have possibly unequal cell numbers (the n_{ij} are not predetermined by the design).

If $\gamma_{ij} \neq 0$ then the effect of Z on Y varies in the populations, so the estimate of exposure effect will differ according to the value of Z . Interpretation is therefore problematic, although sometimes comparison is made with some ‘standard’ distribution for Z . A formal significance test for $\gamma_{ij} = 0$ may be constructed as in §2.4.5.

Y binary, Z classified. The data are as above, where now $\bar{Y}_{ij}$ is the proportion of positive responses in exposure group i and Z -level j . However, since the response data are binary, the additive model (18) is inappropriate. Usually a **logistic regression** model is used to fit such data (*c.f.* Example 14). If the number of Z levels is large compared to the total number of observations, conditional logistic regression should be used instead to avoid the Neyman-Scott problem.

Y continuous, Z continuous. In this case a linear regression model may be appropriate. Suppose that

$$Y_{ij} = \mu + \alpha_i + \beta_i Z_{ij} + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ (in the simplest case). If the β_i are not all equal then the effect of Z on Y varies in the populations and we have the same difficulty of interpretation discussed in the first case. Assume then that $\beta_i = \beta$ for all i (parallel regressions). The analysis of this model is carried out using the analysis of covariance (see §2.12.3). In particular, let b be the least squares estimator of β . Then, since $E(b) = \beta$, we see that

$$\hat{\delta} \equiv \bar{Y}_1 - \bar{Y}_2 - b(\bar{Z}_1 - \bar{Z}_2)$$

is an unbiased estimator of the exposure effect $\delta = \alpha_1 - \alpha_2$ (**Exercise:** verify). The variance of $\hat{\delta}$ may be deduced from general multiple regression formulae (see §2.12.1).

Note that if Z is measured with error and this is not accounted for in the analysis, then $\hat{\delta}$ will be biased because of the regression dilution effect (see §3.3).

Y binary, Z continuous. The data are as above, but since Y_{ij} is binary a linear regression of Y_{ij} on Z_{ij} as in (2.12.3) is inappropriate. Instead a logistic regression model may be used; otherwise, the issues are similar to the above case. Notice that in this case the Neyman-Scott problem does not arise, because there will be few (usually just one) parameters associated with each Z variable.

3.5.3 Instrumental variables

An alternative approach to the elimination of bias due to the presence of confounding variables is to find one or more **instrumental variables**, which are variables correlated with the exposure variable X but uncorrelated with the potential confounders Z and the errors ϵ . Effectively, these variables detect changes in X that are uncorrelated with Z and lead to unbiased estimation of the desired causal effect. This topic is beyond the scope of the present course.

3.5.4 Criteria for judging causality

We said at the beginning of this section that it is often difficult to make statements about causality on the basis of observational studies. Nonetheless, there are many situations where the primary question of interest relates specifically to the presence or absence of causal links (for example, does smoking contribute to the risk of cancer development?). The following criteria have been suggested by Austin Bradford Hill for judging causality in observational studies:

Strength of association: The stronger the association the ‘better’.

Dose-response: in a causal relationship, higher exposure would be expected to give higher incidence.

Time order: cause usually precedes effect.

Coherence: the posited causal relationship is compatible with other knowledge of the system under study.

Consistency: The relationship is found in many studies, typically under different circumstances.

Specificity: Causality is more plausible when the effect is associated with a specific condition rather than a whole host of conditions. Also, ‘better’ if associated with exposure to this condition alone.

Reversibility: if the exposure is removed, the effect disappears.

The last two of these criteria are not universally accepted.

3.6 Epidemiological observational studies

We briefly review two major types of epidemiological observational study: cohort and case-control studies.

3.6.1 Cohort studies

The term ‘cohort’ is used by epidemiologists to mean a group of individuals followed over time. In demography it is sometimes used to distinguish generational as opposed to cross-sectional approaches. A **cohort study** is the epidemiological equivalent of a clinical trial: subjects are compared according to exposure and followed up for outcome. However, unlike a clinical trial the exposure is not assigned by the investigator — it may be self-selected (e.g. smoking) or naturally predetermined (e.g. gender). It is this characteristic that distinguishes cohort studies as being observational.

A cohort may be *closed* or *open*. In the former case, membership is defined at the outset so that numbers can only get smaller as the study progresses due to drop-out, death etc. An open (or *dynamic*) cohort can take on new members as study progresses. This may arise, for example, when studying social trends in a particular geographical region, because the population of that region is constantly changing.

3.6.2 Case-control studies

A **case-control study** may be defined as follows:

“The observational epidemiologic study of persons with the disease (or other outcome variable) of interest and a suitable control (comparison, reference) group of persons without the disease. The relationship of the attribute to the disease is examined by comparing the diseased and nondiseased group with regard to how frequently the attribute is present, or if quantitative, the levels of the attribute in each group. In short the past history of exposure to a suspected risk factor is compared between “cases” and “controls”, persons who resemble the cases in such respects as age and sex but do not have the disease or condition of interest.”

Last, J. M. *A Dictionary of Epidemiology*

Thus the data are sampled by *disease status* (outcome) and *exposure* becomes the random variable. The main objective of a case-control study is to assess the association between disease status and the exposure variable(s) to ascertain potential risk factors for the disease.

One of the most difficult tasks when carrying out a case-control study is the selection of controls. To avoid bias, the control group should represent the population at risk of becoming cases (as an extreme example, there would be little point in including men in the control group of

a study looking into ovarian cancer!). Matching may also be used to select controls (see above). Possible sources for controls include the general population (e.g. using a population register), and other patients in hospitals or clinics.

Example 15: smoking and lung-cancer. In a famous study published in the *British Medical Journal* in 1950, Hill and Doll sampled 1357 cases of lung cancer from four hospitals in the UK along with 1357 hospital-based controls. The two groups were compared as regards smoking history, to yield the following data:

	Cases	Controls	Total
Smoker	1350	1296	2646
Non-smoker	7	61	68
Total	1357	1357	2714

Notice the by definition, the outcome variable in a case-control study will always be binary. The simplest data structure for such a study is a 2×2 table such as that in Example 15:

	Cases	Controls	Total
Exposed	a	b	$a + b$
Unexposed	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Let ψ be the **odds ratio** = (odds of case given exposed)/(odds of case given unexposed). Then it can be shown that

$$\hat{\psi} = \frac{ad}{bc}$$

is an approximately unbiased estimator of ψ . Intuitively, the odds of a case given exposed should be about a/b , the odds of a case given unexposed about c/d . This would lead to the above estimator if the data were *prospective* (i.e. sampled by exposure status). However, the data were not sampled in this way — the sampling was conditional on outcome. To see the effect of this, let E and C denote the events “exposed” and “case” respectively, with E^c and C^c denoting their complements. Then the odds ratio of interest is

$$\psi = \frac{P(C|E)/P(C^c|E)}{P(C|E^c)/P(C^c|E^c)} = \frac{P(E|C)/P(E^c|C)}{P(E|C^c)/P(E^c|C^c)},$$

see also the exercises. This shows that the odds ratio is the same if we condition on outcome as if we condition on exposure — therefore $\hat{\psi}$ is a valid estimator of the odds ratio of interest in a case-control study. Note that we do not need to know the numbers of exposed or unexposed

individuals in the population, nor the total number of cases in the population. Furthermore, it can be shown that

$$\text{Var}(\log \hat{\psi}) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

approximately. Note the implications of this variance formula. The variance cannot be reduced beyond the reciprocal of the entry in a given cell by increasing the frequencies of the other cells.

Analysis of Example 15: The odds ratio for the Doll & Hill data is $\hat{\psi} = (1350 \times 61)/(1296 \times 7) = 9.077$. The log odds is 2.206, with a standard error of $\sqrt{(1350^{-1} + 61^{-1} + 1296^{-1} + 7^{-1})} = 0.401$. If there were no association between smoking and lung cancer we would have $\log \psi = 0$; to test this we can therefore carry out a z -test on $\log \hat{\psi}$. (In a z -test, the distribution of the test statistic under the null hypothesis is approximated by a normal distribution). The z -statistic is $2.206/0.401 = 5.501$, which leads to overwhelming rejection of the null hypothesis.

Another way to test association in such a 2×2 table is using a chi-squared test. Under the null hypothesis of no association, the expected frequencies in the first row of the table are both 1323, and those in the second row are both 34. The test statistic is $27^2/1323 + (-27)^2/1323 + (-27)^2/34 + 27^2/34 = 43.984$, compared with a χ_1^2 distribution, this gives a p -value of around 3×10^{-11} . Yet another way to do the analysis is using a loglinear model (see STAT0028).

There is a strong association between smoking and lung cancer. The odds of cancer given smoking is 9 times the odds of cancer given no smoking. Since lung cancer is relatively rare, the quantities $P(C^c|E)$ and $P(C^c|E^c)$ in the definition of the odds ratio are approximately 1 and the odds ratio is approximately equal to $P(C|E)/P(C|E^c)$. This is called the *relative risk* (ratio of probabilities of lung cancer for a smoker and a non-smoker). Thus a smoker is 9 times more likely to contract lung cancer than a non-smoker.

Design issues for case-control studies: one might consider sample size determination and optimal allocation for inference on the odds-ratio (interval estimation and hypothesis testing). The principles are essentially the same as those in §2.14. It is also necessary to consider other covariates for which adjustment may be required (confounders) if matching was not possible.

Difficulties in the analysis of case-control data: missing data can be a problem — for example, when there are several exposures, not all exposure variables may be available for all subjects. Furthermore, exposure variables are often measured with error (*e.g.* when data are based on subjects' memory). This can lead to problems of bias if not properly accounted for

3.6.3 Comparison of cohort and case-control studies

Cohort	Case-control
Sample from population	Sample cases and controls
Can calculate incidence rates	Can calculate ratios only
Usually expensive	Usually less expensive
Convenient for studying many diseases	Convenient for studying many exposures
Usually prospective	Usually retrospective
- likelihood function obtained from $P(D E)$	- likelihood function obtained from $P(E D)$
Efficient when exposure is rare	Efficient when outcome is rare

4 Surveys and sampling

4.1 Introduction

So far we have been interested in comparisons and with assessing causal or associative relationships between variables. We now look at the case where we are more interested in collecting representative material to *describe* aspects of a population.

Possible applications of sampling methods include: quality control of manufacturing processes (*e.g.* regular on-line sampling); financial audit (*e.g.* sampling of accounts); opinion polls; clinical audit; social surveys; and ecological surveys to determine species abundance.

In a **sample survey**, a sample is drawn from a well-defined population by methods, usually involving randomisation, under the control of the investigator. Conclusions can be drawn with confidence about descriptive properties of the population. However, the interpretation of relationships between variables raises problems similar to those in a pure observational study since, although we may be selecting individuals at random, we are not randomising individuals to ‘treatments’.

The advantages of a sample over a **census** (i.e. collection of data on the entire population) include (i) less cost (money, time, effort), (ii) more careful inspection since there are fewer units to inspect, (iii) feasibility when inspection is destructive (*e.g.* light bulb lifetimes, striking matches), (iv) accessibility (ease of access/use). However, a census is sometimes desirable or necessary, *e.g.* if the population size is very small; in the interests of safety; or if there is a legal requirement (population census).

In this course we consider **probability sampling** in which we have

- a well-defined sampling frame (see below for definition)
- a probabilistic rule for drawing the sample

and in which a **random sample** of n units is selected from the required **target population** of size N . The sample should be selected using random numbers rather than haphazardly, since the latter may introduce unintended bias. A census corresponds to the case when the whole population is used i.e. when $n = N$. The ratio $f = n/N$ is called the **sampling fraction**. Knowledge of the sampling frame and sampling strategy allows probabilistic statements about the population to be made.

4.2 Sample design

The **sample design** specifies the method of collecting the sample. The **sampling units** are the items, or groups of items, to be sampled. The **sampling frame** is a list of the sampling units from which the sample is to be drawn. The sampling frame should cover the target population

or, at least, an accessible but representative subset of it (*e.g.* the phone book does not list all phone users).

The main aim of sample design is to select a representative sample which provides a specified amount of information at a minimum cost.

In any statistical investigation the overall error can be broken down as

Total error = sampling variation + systematic error ('bias')

Sampling variation may be measured by the **standard error** (s.e.) of an estimator (*e.g.* the s.e. of the sample mean is $\sigma/\sqrt{n}$). Sampling variation may be reduced by either increasing sample size or improving the sample design.

Bias is more serious. In the present context, it could arise because the sample is not representative of the target population (more generally in statistics, it could arise via the use of biased estimators, but we consider only simple quantities such as means and proportions here, so we do not need to consider the latter possibility). The main sources of bias in sample surveys are:

- *Misspecified sampling frame.* *E.g.* sample of registered voters for an opinion poll — not necessarily the people who will vote on the day (*e.g.* the 1936 US Presidential Election¹); consumer research surveys — working people often under-represented.
- *Nonresponse.* This can be a difficult problem — the easiest way to deal with it is to ignore nonrespondents in a statistical analysis. However, this assumes implicitly that their responses would have been similar to those actually obtained and this is usually not true *e.g.* working people may be less likely to respond and to have different attitudes; only those who are interested respond *etc.*
- *Non-random sampling.* For example, in **quota sampling**, interviewers are given quotas of, for example, numbers of males and females to interview, but are otherwise allowed to choose who to approach. Bias may arise here due to personal selection of the sampling units.

Note that, unlike sampling variation, bias will persist as sample size increases.

The simplest type of sample is the **simple random sample**, in which every possible sample of n units from the population of N is equally likely to be chosen. There are, however, two main reasons for considering sample designs more complex than this:

¹This is a famous example of an opinion poll that went wrong. The *Literary Digest* magazine ran a poll of ten million voters, selected from databases of its own subscribers, a list of telephone users and a list of car owners. On the basis of this poll, it predicted that the outcome of the election would be Landon 370, Roosevelt 161. Similar exercises had correctly predicted the winners of the previous five elections accurately: however, in the event the outcome was Roosevelt 523, Landon 8. The poll failed because in the economic climate of 1936, many less well off people had cancelled their subscriptions to magazines to save money, or could no longer afford a telephone (and certainly not a car!) — hence the sample was biased in favour of wealthy voters who would tend to vote Republican. The fiasco was so embarrassing that the magazine was completely discredited as a serious publication, and it closed within a couple of years.

1. To improve the precision of estimators. For a simple random sample of size n from a (large) population, $Var(\bar{Y}) = \sigma^2/n$, so we can improve **precision** (measured by $1/\text{variance}$) by increasing n . However, this may be costly or impracticable and it will often be better to improve the sample *design* to achieve increased precision. This is done by **restricted randomisation**.
2. For practical or administrative reasons (e.g. use convenient geographical groupings, or just sample a subsection of the population). Such designs usually reduce the cost of sampling, but may also reduce precision.

Some alternatives to simple random sampling are as follows:

Quasi-random sampling: here, one unit from the first k is selected at random, and every k th unit is selected thereafter. This is also known as **systematic sampling**. It is equivalent to simple random sampling when the list is in random order; however, this is never the case in practice. When using quasi-random sampling, some justification must be made that the ordering of the list is unrelated to the subject of the survey.

Stratified sampling: suppose that the population consists of known strata which are *heterogeneous*: *i.e.* there is variation between the strata (*e.g.* different views of males and females). Simple random sampling may result in the under- or over-representation of some strata by chance. One can attempt to eliminate this by sampling from each stratum independently. Thus stratification can be more efficient than simple random sampling (smaller variance) and consequently cost-effective. One can also stratify after interviewing when the sampling frame contains no information on the strata initially (post-stratification). Finally, it may be of interest to obtain separate results for different strata.

Multi-stage sampling: suppose now that the population has a hierarchy of strata (strata within strata), *e.g.* first geographical, then social class, then age. For example, if the sampling frame is that of students in the UK, one could stratify by geographical region, then take a random sample of universities / colleges within regions, then samples of students within these institutions.

Cluster sampling: a cluster sample is a probability sample in which each sampling unit is a collection, or cluster, of elements. At some stage of 'strata within strata' one can take a random sample of groups of individuals and interview whole groups (clusters), *e.g.* interview everyone in a block of flats for convenience. The main advantage is that it is relatively cheap and easy to administer. The disadvantage is that clusters may be homogeneous within themselves but very different from each other so that the sample becomes unrepresentative and estimation less efficient (compare this with stratification).

Quota sampling: The interviewer is only told how many interviews of certain types are required, rather than being given a specified list of interviewees, *e.g.* ten car owners in each neighbourhood. This economises on call-backs and should keep the sample representative,

but allows more scope for interviewer bias. Furthermore, since probability sampling is not used it is difficult to make probabilistic statements about the population.

4.3 Basic sampling theory

4.3.1 Simple random samples

In a **simple random sample** of size $n \geq 1$, every possible collection of n units in the population has the same chance of being the sample. It will be assumed that sampling is **without replacement**.

The typical use of a sample is to say something about a population mean or proportion (for example, to give a point estimate or confidence interval). Occasionally, we are also interested in estimating totals e.g. total weight, total household income.

Let the finite population size be N . To make progress, we must distinguish between the values taken by units in the population, and those actually observed in the sample. We therefore label the population members $1, \dots, N$, and denote the values of interest by $a_1, \dots, a_N$. For example, a_k may be the annual income of the k th worker. The a_k are not necessarily all different. In a simple random sample (SRS) of size n , every possible combination of n elements from the set $\{a_1, \dots, a_N\}$ has the same chance of being the sample. Denote the values observed in the sample by $y_1, \dots, y_n$, which are realisations of random variables $Y_1, \dots, Y_n$. Thus each y_i is equal to some a_k , and two y_i 's cannot be the same a_k since sampling is without replacement (they may have the same value, however). Obviously, in a SRS each Y_i has an equal chance of being any one of the a_k 's, so $P(Y_i = a_k) = 1/N$.

The following algorithm can be used to select a simple random sample:

1. Identify each item of the sampling frame by a unique number in $\{1, \dots, N\}$.
2. 'Generate' a random number between 0 and 1 (using a computer, random number table, or randomising device).
3. Multiply by N and round up.
4. Select the population member indicated.
5. Repeat n times (for sampling without replacement, draw again if any number is chosen twice).

In R, the function `sample` can be used to do this directly.

Estimating a population mean. Define the **population mean** of the a -values as $\mu = N^{-1} \sum_k a_k$ and the **population variance** as $\sigma^2 = (N - 1)^{-1} \sum_k (a_k - \mu)^2$. The divisor $N - 1$ rather than N is mainly for algebraic convenience below. It seems reasonable to estimate μ and σ^2 using the **sample mean** $\bar{y} = n^{-1} \sum_i y_i$ and the **sample variance** $s^2 = (n - 1)^{-1} \sum_i (y_i - \bar{y})^2$.

These are observed values of random variables $\bar{Y} = n^{-1} \sum_i Y_i$ and $S^2 = (n-1)^{-1} \sum_i (Y_i - \bar{Y})^2$. To study the properties of these estimators, we first establish some useful results about the random variables $Y_1, \dots, Y_n$.

Lemma 1:

$$E(Y_i) = \mu \quad \text{Var}(Y_i) = \frac{(N-1)\sigma^2}{N} \quad \text{and} \quad \text{Cov}(Y_i, Y_j) = -\frac{\sigma^2}{N} \text{ for } i \neq j.$$

Note that E , Var and Cov here stand for probability, expectation, variance and covariance over repeated samples. Imagine that a list is made of all the possible samples of size n from the population. Then the i th member y_i from each sample is picked out. For a SRS the simple unweighted average of these y_i 's is $E(Y_i)$.

Proof of Lemma 1. We have $E(Y_i) = \sum_{k=1}^N a_k P(Y_i = a_k) = N^{-1} \sum_{k=1}^N a_k = \mu$ as required. For the variance, we have $E(Y_i - \mu)^2 = \sum_{k=1}^N (a_k - \mu)^2 P(Y_i = a_k) = N^{-1} \sum_{k=1}^N (a_k - \mu)^2 = (N-1)\sigma^2/N$. The covariance result is a bit more difficult:

$$\text{Cov}(Y_i, Y_j) = E[(Y_i - \mu)(Y_j - \mu)] = \sum_{k=1}^N \sum_{\ell=1}^N (a_k - \mu)(a_\ell - \mu) P(Y_i = a_k, Y_j = a_\ell).$$

Now, for $i \neq j$, we have

$$P(Y_i = a_k, Y_j = a_\ell) = \begin{cases} 0 & k = \ell \\ 1/[N(N-1)] & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= \frac{1}{N(N-1)} \left\{ \sum_{k=1}^N \sum_{\ell=1}^N (a_k - \mu)(a_\ell - \mu) - \sum_{k=1}^N (a_k - \mu)^2 \right\} \\ &\quad \text{(the second term removes the unwanted contributions where } k = \ell) \\ &= \frac{1}{N(N-1)} \left\{ \sum_{k=1}^N (a_k - \mu) \sum_{\ell=1}^N (a_\ell - \mu) - (N-1)\sigma^2 \right\}. \end{aligned}$$

The first term is zero, and the required result follows. □

We are now in a position to show the following:

Theorem 1:

$$E(\bar{Y}) = \mu \quad \text{Var}(\bar{Y}) = \frac{(1-f)\sigma^2}{n} \quad \text{and} \quad E(S^2) = \sigma^2,$$

where $f = n/N$ is the **sampling fraction**. Thus $\bar{Y}$ and S^2 are unbiased estimators of μ and σ^2 respectively. Also, the usual (infinite population) variance formula, σ^2/n , for $\bar{Y}$ is modified to

$(1-f)\sigma^2/n$. The factor $1-f$ is called the **finite population correction**. In practice, when σ^2 is unknown $Var(\bar{Y})$ may be estimated unbiasedly by $(1-f)S^2/n$ (you might now like to refer back to Example 3 in §1!).

Proof of Theorem 1: we have $E(\bar{Y}) = n^{-1} \sum_{i=1}^n E(Y_i) = \mu$. For $Var(\bar{Y})$ we use

$$Var(\bar{Y}) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n Cov(Y_i, Y_j) = \frac{1}{n^2} \left[\sum_{i=1}^n Var(Y_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n Cov(Y_i, Y_j) \right],$$

which simplifies to the required result on substituting the previous results for the variances and covariances. Finally, the expectation of S^2 is obtained by writing

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n [Y_i - \mu - (\bar{Y} - \mu)]^2,$$

which simplifies to $(n-1)^{-1} \left[\sum_{i=1}^n (Y_i - \mu)^2 - n(\bar{Y} - \mu)^2 \right]$. The expectation of each term in the first sum is equal to the variance of one of the Y s; and the expectation of $(\bar{Y} - \mu)^2$ is $Var(\bar{Y})$. These quantities are already known, whence the result can be derived. $\square$

Note the following three equivalent forms of the variance formula:

$$Var(\bar{Y}) = \frac{(1-f)\sigma^2}{n} = \frac{(N-n)\sigma^2}{nN} = \left(\frac{1}{n} - \frac{1}{N} \right) \sigma^2.$$

4.3.2 Binary values

Suppose that the a_k are binary, 1 or 0, indicating the presence or absence of some characteristic in the k th individual. Suppose there are N_1 individuals with the characteristic (i.e. with $a_k = 1$), and let Σ_1 denote summation over just these individuals. Then

$$\mu = N^{-1} \sum_{k=1}^N a_k = N^{-1} \Sigma_1(1) = N_1/N,$$

is the proportion of individuals with the characteristic, and we can write $N_1 = N\mu$. Notice that the number of individuals without the characteristic is $N - N_1 = N(1 - \mu)$. Therefore the population variance is

$$\begin{aligned} \sigma^2 &= \frac{\Sigma_1(1 - \mu)^2 + \Sigma_0(0 - \mu)^2}{N - 1} \\ &= \frac{N\mu(1 - \mu)^2 + N(1 - \mu)\mu^2}{N - 1} = \frac{N\mu(1 - \mu)}{N - 1}. \end{aligned}$$

In the sample $\bar{y}$ is the proportion of individuals, of the n selected, with the characteristic and, in a similar way, $s^2 = n\bar{y}(1 - \bar{y})/(n - 1)$ (*exercise: verify*). The results of Theorem 1 all hold; for example,

$$\text{Var}(\bar{Y}) = \frac{(1 - f)\sigma^2}{n} = \frac{(1 - f)\mu(1 - \mu)}{(N - 1)f}.$$

In practice $\text{Var}(\bar{Y})$ may be estimated unbiasedly by $(1 - f)\bar{Y}(1 - \bar{Y})/(n - 1)$.

More complete results than just the mean and variance of $\bar{Y}$ can be obtained in the binary case. In fact $n\bar{Y}$ has the **hypergeometric distribution** with

$$\text{pr}(n\bar{Y} = r) = \frac{\binom{N_1}{r} \binom{N_0}{n-r}}{\binom{N}{n}},$$

for $r = 0, 1, \dots, \min(N_1, n)$, where $N_1 = N\mu$ is the number of ones among the population a_k 's and $N_0 = N - N_1$ is the number of zeroes.

4.3.3 Stratified samples

A stratified random sample is one obtained by separating the population elements into nonoverlapping groups, called strata, and then selecting a simple random sample from each stratum.

Suppose then that the population is divided into s strata (levels, or groups) of sizes $N_1, \dots, N_s$, so that $\sum_l N_l = N$. Let the marked values within the l th stratum be $a_{l1}, \dots, a_{lN_l}$ and let $\mu_l = N_l^{-1} \sum_k a_{lk}$ be the mean and $\sigma_l^2 = (N_l - 1)^{-1} \sum_k (a_{lk} - \mu_l)^2$ be the variance in the l th stratum. In a **stratified sample** independent SRS's are taken from each stratum. Denote the strata sample sizes by n_l , with $\sum_l n_l = n$, the sample means by $\bar{y}_l$ and the sample variances by s_l^2 . Also define

$$\bar{y}_{\text{ST}} = \frac{\sum_{l=1}^s N_l \bar{y}_l}{N}$$

as the weighted or stratified sample mean. Note that the strata means are weighted by the *stratum* sizes, not the sample sizes. The sample values $\bar{y}_l$, $\bar{y}_{\text{ST}}$ and s_l^2 are observed values of the obvious random variables $\bar{Y}_l$, $\bar{Y}_{\text{ST}}$ and S_l^2 . Let $f_l = n_l/N_l$ be the sampling fraction in the l th stratum.

Theorem 2:

$$E(\bar{Y}_{\text{ST}}) = \mu \quad \text{and} \quad \text{Var}(\bar{Y}_{\text{ST}}) = \frac{\sum_{l=1}^s N_l \sigma_l^2 (1 - f_l) / f_l}{N^2}.$$

An unbiased estimator of $\text{Var}(\bar{Y}_{\text{ST}})$ is given by

$$\frac{\sum_{l=1}^s N_l S_l^2 (1 - f_l) / f_l}{N^2}.$$

These results can be proved by applying the results of Theorem 1 to the individual stratum means $\{\bar{Y}_l\}$, which are independent. $\square$

If the f_l are taken to be equal, say to f , then $n_l = fN_l$ for all l and this is called sampling with **proportional allocation** (PA). In this case $n/N = f$ and $Var(\bar{Y}_{ST}) = (1 - f) \sum_l N_l \sigma_l^2 / (nN)$. This variance may be compared with that for SRS. It can be shown that $V_{SRS} > V_{PA}$ when the N_l and the $\frac{|\mu_l - \mu|}{\sigma_l}$ are each not too small.

4.3.4 Confidence intervals

Simple random sampling. If n is large (> 50) and f is not near 1, then an approximate $100(1 - \alpha)\%$ confidence interval for μ is

$$\bar{Y} \pm z_{\alpha/2} S \sqrt{\frac{1-f}{n}}$$

where z_α is the upper $100\alpha\%$ point of the $N(0, 1)$ distribution. When n is small and the population is thought to be approximately normal, $z_{\alpha/2}$ should be replaced by $t_{\alpha/2}(n - 1)$.

Stratified sampling. If the n_l are large and the f_l are not near 1, then an approximate $100(1 - \alpha)\%$ confidence interval for μ is

$$\bar{Y}_{ST} \pm z_{\alpha/2} \hat{se}_{ST}$$

where

$$\hat{se}_{ST} = \hat{se}(\bar{Y}_{ST}) = \sqrt{\frac{\sum_{l=1}^s N_l s_l^2 (1 - f_l) / f_l}{N^2}}.$$

When the n_l are not all large but the population is approximately normal, $z_{\alpha/2}$ should be replaced by $t_{\alpha/2}(d)$, for appropriate degrees of freedom d . An approximation that is sometimes used is **Satterthwaite's approximation**

$$d = \frac{(\sum_{l=1}^s g_l s_l^2)^2}{\sum_{l=1}^s g_l^2 s_l^4 / (n_l - 1)},$$

where $g_l = N_l(1 - f_l)/f_l$. This will always yield a value of d between $\min(n_l - 1)$ and $\sum_l (n_l - 1)$.

4.3.5 Optimal sample sizes

Simple random sampling. Suppose that the cost of a sample of size n is $c_0 + c_1 n$, where c_0 is the set-up cost and c_1 is the running cost per individual. The 'benefit' will be taken as

an increasing function of the precision of the estimator or, equivalently, a decreasing function of its variance which, in the case of μ , is $Var(\bar{Y})$. The usual goal is to maximise the benefit and minimise the cost. Since the former can be achieved with $n = N$, and the latter with $n = 0$, a compromise is necessary.

Example. Take the loss function to be $\text{Cost} + a(\bar{y} - \mu)^2$, where a is a positive constant. Then the expected loss is $L_n = \text{Cost} + a Var(\bar{Y})$ and we choose n to minimise L_n . Now, $Var(\bar{Y}) = \sigma^2(n^{-1} - N^{-1})$ so $\partial L_n / \partial n = c_1 - a\sigma^2/n^2$. Equating this to zero yields $n^* = \sigma\sqrt{a/c_1}$; if $n^* > N$ take $n^* = N$. Also, $\partial^2 L_n / \partial n^2 = 2a\sigma^2/n^3 > 0$ for all $n \geq 0$. Thus the solution n^* for n yields a minimum of L_n . Note that n^* depends on σ^2 so, if this is unknown, a preliminary estimate must be inserted.

Stratified sampling. Suppose now that it is required to minimise $Var(\bar{Y}_{ST})$ subject to given n (fixed cost) by suitable choice of the n_l . Now, $Var(\bar{Y}_{ST}) = N^{-2} \sum_l N_l^2 \sigma_l^2 (n_l^{-1} - N_l^{-1})$ and this is minimised with respect to the n_l , subject to the constraint $\sum_l n_l = n$, when

$$\frac{n_l}{n} = \frac{N_l \sigma_l}{\sum_{l'} N_{l'} \sigma_{l'}}.$$

Note that this condition is equivalent to

$$\boxed{\frac{n_l}{N_l} \propto \sigma_l}$$

That is, the sampling fractions are proportional to the stratum standard deviations. With this choice, $Var(\bar{Y}_{ST}) = N^{-2} \{n^{-1} (\sum_l N_l \sigma_l)^2 - \sum_l N_l \sigma_l^2\}$. Note that in order to calculate this optimum choice of n_l , the N_l and σ_l must be specified.

When the within-strata standard deviation is constant, so that $\sigma_l = \sigma$, then the results above yield $\frac{n_l}{n} = \frac{N_l}{N}$, which is equivalent to keeping $\frac{n_l}{N_l}$ constant. In other words, a **uniform sampling fraction** (proportional allocation) is optimal. When the within-strata standard deviations vary, we use a **variable sampling fraction**.

4.3.6 Cluster and multi-stage sampling

A population is generally made up of a hierarchy of units of different sizes and types. Consider the following sampling scheme: at stage I select at random a number of **clusters**; at stage II select all units within the chosen clusters. This scheme is known as **cluster sampling**. For example, if we wanted to obtain an impression of reading skills amongst year 8 children in the UK, we could select a simple random sample of schools and test the reading skills of each year 8 child in the chosen schools.

Alternatively, if we take a SRS at stage II then the scheme is known as **two-stage sampling**. If there are more than two stages then we have **multi-stage sampling**. These designs are used mainly for cost and practical reasons but are generally less precise than SRS as the samples are likely to be less representative. To ensure good representation it is best to use clusters with units as *heterogeneous* as possible within clusters (compare this with stratification, where the strata should be as homogeneous as possible).

For theoretical purposes, it is easiest to proceed as though the sampling of clusters is carried out *with replacement* (this will make little difference in practice providing the population contains a large number of clusters). If the cluster sizes are approximately equal then SRS of clusters is adequate. Otherwise it is best to sample clusters with “probability proportional to size” (**PPS sampling**).

Suppose then that there are M distinct clusters of sizes $N_1, \dots, N_M$, so that $\sum_{\ell} N_{\ell} = N$. Let the marked values within the ℓ th cluster be $a_{\ell 1}, \dots, a_{\ell N_{\ell}}$ and let $\mu_{\ell} = N_{\ell}^{-1} \sum_k a_{\ell k}$ be the mean in the ℓ th cluster. At the first stage we select a PPS sample with replacement of size m from the M clusters; that is, the probability p_{ℓ} that cluster ℓ is selected at any drawing is N_{ℓ}/N . At the second stage we select all units within the chosen clusters. Define

$$\bar{y}_{\text{CL}} = \frac{\sum_{i=1}^m \bar{Y}_i}{m}$$

where $\bar{Y}_i$ is the mean of the i th selected cluster (which is equal to the corresponding μ_{ℓ}). Note that this is an *unweighted* average of the sampled cluster means. This is a consequence of the PPS sampling. Let $\bar{Y}_{\text{CL}}$ denote the corresponding random variable. The properties of $\bar{Y}_{\text{CL}}$ are given by

Theorem 3:

$$E(\bar{Y}_{\text{CL}}) = \mu \quad \text{and} \quad \text{Var}(\bar{Y}_{\text{CL}}) = \frac{\sum_{\ell=1}^M N_{\ell}(\mu_{\ell} - \mu)^2}{Nm}.$$

An unbiased estimator of $\text{Var}(\bar{Y}_{\text{CL}})$ is given by

$$\frac{\sum_{i=1}^m (\bar{Y}_i - \bar{Y}_{\text{CL}})^2}{m(m-1)}.$$

Proof of Theorem 3: we start by establishing the mean and variance of $\bar{Y}_i$. Because of the PPS sampling, we have $P(\bar{Y}_i = \mu_\ell) = N_\ell/N$. We therefore have

$$E(\bar{Y}_i) = \sum_{\ell=1}^M \mu_\ell P(\bar{Y}_i = \mu_\ell) = \sum_{\ell=1}^M \mu_\ell \frac{N_\ell}{N} = \frac{1}{N} \sum_{\ell=1}^M \sum_{k=1}^{N_\ell} a_{\ell k} = \mu,$$

and

$$Var(\bar{Y}_i) = E(\bar{Y}_i - \mu)^2 = \sum_{\ell=1}^M (\mu_\ell - \mu)^2 P(\bar{Y}_i = \mu_\ell) = \frac{1}{N} \sum_{\ell=1}^M N_\ell (\mu_\ell - \mu)^2.$$

It is now trivial to show that $E(\bar{Y}_{CL}) = \mu$. For the variance, since clusters are sampled with replacement we have

$$Var(\bar{Y}_{CL}) = \frac{1}{m^2} \sum_{\ell=1}^m Var(\bar{Y}_i) = \frac{1}{m^2} \left[m \times \frac{1}{N} \sum_{\ell=1}^M N_\ell (\mu_\ell - \mu)^2 \right] = \frac{1}{mN} \sum_{\ell=1}^M N_\ell (\mu_\ell - \mu)^2.$$

It remains to show that $\sum_{i=1}^m (\bar{Y}_i - \bar{Y}_{CL})^2 / m(m-1)$ is an unbiased estimator of $Var(\bar{Y}_{CL})$. As usual, we proceed by writing

$$E \left[\sum_{i=1}^m (\bar{Y}_i - \bar{Y}_{CL})^2 \right] = E \left[\sum_{i=1}^m (\bar{Y}_i - \mu - (\bar{Y}_{CL} - \mu))^2 \right] = \sum_{i=1}^m Var(\bar{Y}_i) - m Var(\bar{Y}_{CL}).$$

The result now follows on simplification. □

Two-stage sampling. Suppose now that instead of sampling all units in the chosen clusters, we draw a SRS at the second stage: therefore we have a two-stage sampling scheme. Define

$$\bar{y}_{ts} = \frac{\sum_{i=1}^m \bar{y}_i}{m},$$

where now $\bar{y}_i$ is the sample mean from the i th selected cluster. Exactly as in Theorem 3, it can be shown that $E(\bar{y}_{ts}) = \mu$, *i.e.* $\bar{y}_{ts}$ is unbiased. The variance formula is slightly more complex.

Detailed analysis of properties of alternative estimators and optimal choice of sample sizes is required for various multi-stage sampling schemes and we do not explore these matters further in this course.

4.3.7 Examples

Example 16. In the analysis of the cash position of a departmental store, an accounting firm selects a SRS without replacement of 15 monthly retail accounts receivable from a total of $N = 200$ current accounts. The sampled account balances (in dollars) are as follows:

14.50	30.20	17.80	10.00	8.50	23.40	15.50	27.50
6.90	19.50	42.00	13.30	23.70	18.40	12.10	

1. Estimate the mean balance μ for all 200 accounts along with its standard error and obtain an approximate 95% confidence interval for μ .

Here $n = 15$, $\sum y = 283.3$, $\sum y^2 = 6570.85$. An unbiased estimate of μ is therefore $\bar{y} = \$18.8867$. Also $s^2 = 87.16124$ giving

$$s.e.(\bar{Y}) = \sqrt{\left(\frac{200 - 15}{200}\right) \left(\frac{87.16124}{15}\right)} = 2.3184.$$

Assuming approximate normality and $t_{\alpha/2}(14) = 2.145$, an approximate 95% confidence interval for μ is $18.8867 \pm 2.145 \times 2.3184 = (\$13.91, \$23.86)$.

2. Estimate the total τ of all outstanding accounts and give an approximate 95% confidence interval for τ .

$\tau = N\mu = 200\mu$ so an unbiased estimate of τ is $\hat{\tau} = 200 \times 18.887 = \$3777.7$. A 95% confidence interval is $(\$2782.8, \$4771.8)$.

Example 17. A wholesale distributor of frozen foods is considering discounting the price of a frozen food for a bulk purchase. A random sample of $n = 50$ of the firm's $N = 430$ buyers was contacted and 15 indicated that they would accept the offer.

1. Estimate the proportion π of the 430 buyers who would take advantage of the offer and give 95% confidence limits.

Here $p = 15/50 = 0.3$ is an unbiased estimate of π and

$$s.e.(p) = \sqrt{\left(\frac{430 - 50}{430 - 1}\right) \left(\frac{0.3 \times 0.7}{50}\right)} = 0.061.$$

Approximate 95% confidence interval for π is $0.3 \pm 1.96 \times 0.061 = (0.181, 0.420)$.

2. Give 95% confidence limits on the number of the firm's buyers who would accept the offer.

Total = $N\pi$, estimated by $Np = 430 \times 0.3 = 129$. Confidence limits are 77 to 181 customers.

Example 18. In order to estimate the mean number of children per family, μ , in a district with 14,848 couples, the area was stratified into four regions, roughly reflecting social differences in the district. Simple random samples were taken in each of the four regions with the following results.

Region (Stratum)	1	2	3	4
Stratum size N_ℓ	2560	5211	3718	3359
Sample size n_ℓ	200	500	300	200
Sample mean $\bar{y}_\ell$	2.3	3.1	1.7	2.4
Sample s.d. s_ℓ	1.1	1.3	0.6	1.0

1. Obtain an unbiased estimate of μ .

An unbiased estimate of μ is $\bar{y}_{ST} = 2.453$ (verify).

2. Estimate the s.e. of this estimator and write down an approximate 95% confidence interval for μ .

$$s.e.(\bar{Y}_{ST}) = \frac{\sqrt{\sum_{\ell=1}^s N_\ell s_\ell^2 (1 - f_\ell) / f_\ell}}{N} = 0.0290,$$

(verify). Approximate 95% confidence interval is $2.453 \pm 1.96 \times 0.0292 = (2.40, 2.51)$.

3. Assuming the population variances in the four regions are similar, obtain the optimal allocation if a sample size of 100 is required.

Use a uniform sampling fraction, so that

$$\frac{n_l}{N_l} = \frac{n}{N} = \frac{100}{14848} = 0.006735,$$

for all strata, giving $n_1 = 0.006735 \times 2560 = 17.2$ etc. Sample sizes after rounding are 17, 35, 25, 23.

4. Obtain the optimal allocation if a sample size of 100 is required when the population sd's within the four regions are known to be approximately 0.8, 2.1, 1.2 and 1.3 respectively.

Take $n_l/N_l \propto \sigma_l$, so that

$$\frac{n_l}{N_l \sigma_l} = \frac{n}{\sum_l N_l \sigma_l} = \frac{100}{21820} = 0.004583$$

for all strata, giving $n_1 = 0.004583 \times 2560 \times 0.8 = 9.39$ etc. Sample sizes after rounding are 9, 50, 21, 20 (taking into account that the sum should be 100). Note that more observations are needed when there is greater variability within a stratum.

4.4 Ratio and regression estimators

4.4.1 Bivariate sampling

Until now we have considered recording the value of a single variable for each individual in a sample. Suppose now that we record two marked values, a_k and b_k , *e.g.* the height and weight of the k th fireman. The sample will then comprise pairs (x_i, y_i) , $i = 1, \dots, n$; and relationships between the observed x 's and y 's should reflect those between the a 's and b 's in the population. Population moments will be defined as $\mu_a = N^{-1} \sum_k a_k$ and $\mu_b = N^{-1} \sum_k b_k$ (means), $\sigma_a^2 = (N-1)^{-1} \sum_k (a_k - \mu_a)^2$ and $\sigma_b^2 = (N-1)^{-1} \sum_k (b_k - \mu_b)^2$ (variances) and $\sigma_{ab} = (N-1)^{-1} \sum_k (a_k - \mu_a)(b_k - \mu_b)$ (covariance).

Assume that sampling is SRS. Again, x_i, y_i are outcomes of random variables X_i, Y_i .

Lemma 2:

$$\begin{aligned} E(X_i) &= \mu_a & \text{and} & & E(Y_i) &= \mu_b \\ \text{Var}(X_i) &= \frac{(N-1)\sigma_a^2}{N} & \text{and} & & \text{Var}(Y_i) &= \frac{(N-1)\sigma_b^2}{N} \\ \text{Cov}(X_i, X_j) &= -\frac{\sigma_a^2}{N} & \text{and} & & \text{Cov}(Y_i, Y_j) &= -\frac{\sigma_b^2}{N} \text{ for } i \neq j \\ \text{Cov}(X_i, Y_i) &= \frac{(N-1)\sigma_{ab}}{N} & \text{and} & & \text{Cov}(X_i, Y_j) &= -\frac{\sigma_{ab}}{N} \text{ for } i \neq j. \end{aligned}$$

Most of these results have been established already in Lemma 1. The only new ones are those for $\text{Cov}(X_i, Y_i)$ and $\text{Cov}(X_i, Y_j)$; the derivation follows exactly the same lines as in Lemma 1; see the optional exercises. $\square$

Let $(\bar{x}, \bar{y})$, (s_x^2, s_y^2) and $s_{xy} = (n-1)^{-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$ denote the sample means, variances and covariance respectively and $(\bar{X}, \bar{Y})$, (S_X^2, S_Y^2) and S_{XY} denote the corresponding random variables. The properties of these quantities are summarised in

Theorem 4:

$$\begin{aligned} E(\bar{X}) &= \mu_a & \text{and} & & E(\bar{Y}) &= \mu_b \\ \text{Var}(\bar{X}) &= \frac{(1-f)\sigma_a^2}{n} & \text{and} & & \text{Var}(\bar{Y}) &= \frac{(1-f)\sigma_b^2}{n} \\ E(S_X^2) &= \sigma_a^2 & \text{and} & & E(S_Y^2) &= \sigma_b^2 \\ \text{Cov}(\bar{X}, \bar{Y}) &= & & & \frac{(1-f)\sigma_{ab}}{n} \\ E(S_{XY}) &= & & & \sigma_{ab} \end{aligned}$$

The proof follows the same lines as that for Theorem 1; see the optional exercises. $\square$

4.4.2 Estimation of a ratio

Suppose we wish to estimate the ratio μ_b/μ_a . It is natural to consider the estimator $\bar{Y}/\bar{X}$. Note, however, that $E(\bar{Y}/\bar{X}) \neq E(\bar{Y})/E(\bar{X})$ in general, so $\bar{Y}/\bar{X}$ will not automatically be an unbiased estimator of μ_b/μ_a . To calculate the (approximate) mean and variance of $\bar{Y}/\bar{X}$ we use a generalisation of a result that we have already seen in §2.13:

Lemma 3: Let $\mathbf{Z}$ be a random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and let $g(\mathbf{z})$ be a differentiable scalar-valued function. Then

$$E\{g(\mathbf{Z})\} \simeq g(\boldsymbol{\mu}) \quad \text{and} \quad \text{Var}\{g(\mathbf{Z})\} \simeq [g'(\boldsymbol{\mu})]^T \boldsymbol{\Sigma} g'(\boldsymbol{\mu}),$$

where $g'(\boldsymbol{\mu})$ is the vector of partial derivatives of g with respect to each element of $\boldsymbol{\mu}$.

Proof: A Taylor series expansion of $g(\mathbf{Z})$ about $\boldsymbol{\mu}$ yields $g(\mathbf{Z}) \simeq g(\boldsymbol{\mu}) + (\mathbf{Z} - \boldsymbol{\mu})^T g'(\boldsymbol{\mu})$. Taking expectations, we find $E[g(\mathbf{Z})] \simeq g(\boldsymbol{\mu})$ as required. We therefore have $g(\mathbf{Z}) - E[g(\mathbf{Z})] \simeq (\mathbf{Z} - \boldsymbol{\mu})^T g'(\boldsymbol{\mu})$, so that

$$\begin{aligned} \text{Var}[g(\mathbf{Z})] &= E\left\{[g(\mathbf{Z}) - E(g(\mathbf{Z}))]^T [g(\mathbf{Z}) - E(g(\mathbf{Z}))]\right\} \\ &\simeq E\left\{[g'(\boldsymbol{\mu})]^T (\mathbf{Z} - \boldsymbol{\mu})(\mathbf{Z} - \boldsymbol{\mu})^T g'(\boldsymbol{\mu})\right\} = [g'(\boldsymbol{\mu})]^T \boldsymbol{\Sigma} g'(\boldsymbol{\mu}), \end{aligned}$$

as required. $\square$

The method of finding approximate moments as given in Lemma 3 is known as the **delta method** or the **method of differentials**. The terminology stems from the older approach of writing $\delta f(z) = f(z + \delta z) - f(z) = \delta z f'(z) + \dots$, *etc.* Application of the delta method to a ratio $g(z_1, z_2) = z_1/z_2$ yields

$$\begin{aligned} E\left(\frac{Z_1}{Z_2}\right) &\simeq \frac{\mu_1}{\mu_2} \\ \text{Var}\left(\frac{Z_1}{Z_2}\right) &\simeq \frac{\sigma_1^2}{\mu_2^2} - \frac{2\sigma_{12}\mu_1}{\mu_2^3} + \frac{\sigma_2^2\mu_1^2}{\mu_2^4}, \end{aligned}$$

where $\sigma_1^2 = \sigma_{11}$ and $\sigma_2^2 = \sigma_{22}$. The approximate variance may be written as

$$\text{Var}\left(\frac{Z_1}{Z_2}\right) \simeq \left(\frac{\mu_1}{\mu_2}\right)^2 (c_1^2 - 2\rho_{12}c_1c_2 + c_2^2),$$

where $c_j = \sigma_j/\mu_j$ is the coefficient of variation of Z_j , $j = 1, 2$, and ρ_{12} is the correlation between Z_1 and Z_2 .

Returning now to the problem of estimating μ_b/μ_a , the delta method and the results of section

4.4.1 together yield

$$\begin{aligned} E\left(\frac{\bar{Y}}{\bar{X}}\right) &\simeq \frac{\mu_b}{\mu_a} \\ \text{Var}\left(\frac{\bar{Y}}{\bar{X}}\right) &\simeq \left(\frac{1-f}{n}\right) \left(\frac{\mu_b}{\mu_a}\right)^2 (c_a^2 - 2\rho_{ab}c_ac_b + c_b^2), \end{aligned}$$

where $c_a = \sigma_a/\mu_a$, $c_b = \sigma_b/\mu_b$ and $\rho_{ab} = \sigma_{ab}/(\sigma_a\sigma_b)$. The estimator $\bar{Y}/\bar{X}$ of the ratio μ_b/μ_a is therefore approximately unbiased and its variance can be estimated by inserting sample estimates into the approximate formula, *i.e.* $\bar{y}$ for μ_b , $\bar{x}$ for μ_a , s_y^2 for σ_b^2 , s_x^2 for σ_a^2 and s_{xy} for σ_{ab} .

4.4.3 Ratio estimators

Suppose that we wish to estimate μ_b from a SRS. We could just record the b -value for each sampled individual, yielding a sample $y_1, \dots, y_n$, and then use $\bar{Y}$ as an unbiased estimator of μ_b . However, suppose that it is easy to record a as well as for b for each individual (thus yielding a sample of (x, y) pairs), that a and b are known to be roughly proportional in the population (that is, $b_k/a_k \simeq \mu_b/\mu_a$ for each k) and that μ_a is known. Then one can apply a “correction factor” to the raw estimate $\bar{y}$ as follows. If $\bar{x}$ is too low (*i.e.* smaller than μ_a) then $\bar{y}$ is also likely to be too low (since $\bar{y}/\bar{x} \simeq \mu_b/\mu_a$) and the corrected version $(\mu_a/\bar{x})\bar{y}$ is suggested. The same obtains if $\bar{x}$ is too high. Formally, the **ratio estimator** of μ_b is defined as

$$\hat{\mu}_b = \mu_a(\bar{Y}/\bar{X})$$

Another way of looking at this is that one starts with the known μ_a and then multiplies by the sample estimator of μ_b/μ_a .

The results of Section 4.4.2 apply directly to $\hat{\mu}_b$:

$$\begin{aligned} E(\hat{\mu}_b) &\simeq \mu_b \\ \text{Var}(\hat{\mu}_b) &\simeq \left(\frac{1-f}{n}\right) \mu_b^2 (c_a^2 - 2\rho_{ab}c_ac_b + c_b^2). \end{aligned}$$

From Section 4.4.1, the variance of the raw estimator $\bar{Y}$ for μ_b is $\text{Var}(\bar{Y}) = (1-f)\sigma_b^2/n$. It follows from this that $\bar{Y}$ is less efficient than $\hat{\mu}_b$, in the sense of having larger variance, when

$$\rho_{ab} \geq \frac{c_a}{2c_b},$$

assuming that $c_a/c_b > 0$, *i.e.* $\mu_a/\mu_b > 0$. Therefore ratio estimation is efficient when the two variables are highly correlated.

4.4.4 Regression estimators (Optional)

The ratio estimator of the previous section was based on the approximate proportionality $b_k \simeq \beta a_k$ with $\beta = \mu_b/\mu_a$ for all k . This is now generalised to a linear relationship $b_k = \alpha + \beta a_k + e_k$ with $E(e_k) = 0$ over the population (*i.e.* $\sum_k e_k = 0$). Note that averaging the relationship over k yields $\mu_b = \alpha + \beta\mu_a$, so it may be written as $b_k - \mu_b = \beta(a_k - \mu_a) + e_k$. Also note that such a linear relationship holds for *any* given value of β since the e_k thus defined automatically sum to zero. What matters is whether useful results can be obtained for some β ; see below.

The straightforward application of this is to “correct” $\bar{y}$ (the raw estimate of μ_b) downwards by the amount $\beta(\bar{x} - \mu_a)$. This can be seen by rewriting the linear relationship as $\mu_b = b_k - \beta(a_k - \mu_a) - e_k$. Formally, the **regression estimator** of μ_b is defined as

$$\tilde{\mu}_b = \bar{Y} - \beta(\bar{X} - \mu_a)$$

Note that its use depends on knowing μ_a and specifying a value for β .

Example 19. Let a_k and b_k be the weight and area of the k th leaf on a plant. It is required to estimate the total leaf area $\sum_k b_k = N\mu_b$. One could therefore measure every b_k . More easily, one could weigh all the leaves together, giving $\sum_k a_k = N\mu_a$ and then record the (a_k, b_k) ’s just for a sample of leaves.

Now a_k comprises the weight of the leaf stem (of comparatively small area) and the weight of the main area of leaf, so $a_k \simeq (\text{sw}) + db_k$, where sw = stem weight and d = weight/unit area of the main leaf. Hence $b_k \simeq \alpha + \beta a_k$ with $\beta = d^{-1}$ and $\alpha = -d^{-1}(\text{sw})$. Hence the regression estimator for total leaf area is $N\tilde{\mu}_b = N\bar{Y} - N\beta(\bar{X} - \mu_a)$, with $\beta = 1/d$.

Theorem 5: The regression estimator $\tilde{\mu}_b$ is unbiased for μ_b and has variance

$$\text{Var}(\tilde{\mu}_b) = \left(\frac{1-f}{n} \right) (\sigma_b^2 - 2\beta\sigma_{ab} + \beta^2\sigma_a^2).$$

Proof is omitted. $\square$

The decrease in variance gained by using $\tilde{\mu}_b$ in place of $\bar{Y}$ as an estimator of μ_b is

$$\text{Var}(\bar{Y}) - \text{Var}(\tilde{\mu}_b) = \left(\frac{1-f}{n} \right) (2\beta\sigma_{ab} - \beta^2\sigma_a^2).$$

This is non-negative provided

$$\beta\rho_{ab} \geq \beta^2 \frac{\sigma_a}{2\sigma_b},$$

so β must have the same sign as ρ_{ab} . This is analogous to the result for the ratio estimator given above. In practice one can use a pre-assigned value of β , as in the above example, or estimate β from the sample (*e.g.* by least-squares).

Note that $Var(\tilde{\mu}_b)$ is a quadratic function of β . In particular, the minimum value of this variance is

$$V_{min} = \left(\frac{1-f}{n} \right) \sigma_b^2 (1 - \rho_{ab}^2),$$

which is attained at $\beta^* = \sigma_{ab}/\sigma_a^2$. The increase in $Var(\tilde{\mu}_b)$ due to using β in place of β^* is

$$\left(\frac{1-f}{n} \right) \sigma_a^2 \left(\beta - \frac{\sigma_{ab}}{\sigma_a^2} \right)^2 = \left(\frac{1-f}{n} \right) \left(1 - \frac{\beta}{\beta^*} \right)^2 \sigma_b^2 \rho_{ab}^2,$$

which, expressed as a proportion of V_{min} , is

$$\left(1 - \frac{\beta}{\beta^*} \right)^2 \frac{\rho_{ab}^2}{1 - \rho_{ab}^2}.$$

Thus, the loss of efficiency increases with ρ_{ab}^2 .

Finally, suppose that the least-squares estimator $\hat{\beta}^* = S_{XY}/S_X^2$ is used for β , yielding the estimator $\tilde{\mu}_b^* = \bar{Y} - \hat{\beta}^*(\bar{X} - \mu_a)$. It can be shown that $E(\tilde{\mu}_b^*) \simeq \mu_b$, and that $Var(\tilde{\mu}_b^*) \simeq V_{min}$. This result is not unexpected, since $\hat{\beta}^*$ estimates β^* .

4.5 Data collection (*Optional*)

Having decided on a suitable sampling scheme we need to consider exactly how we will collect the required data. There are many possibilities, their appropriateness depending on the subject matter of the survey.

- Use recorded information that is already available — *e.g.* compiled for administrative or governmental purposes, information in medical records, police files *etc.*
- Direct observation *e.g.* to assess the success of a new product package design, observe points of sale (rather than asking people).
- Face-to-face interviews, usually with an appropriate questionnaire. Careful training is needed to avoid interviewer bias (for example, to prevent the interviewer ‘leading’ the interviewee, intentionally or otherwise).
- Postal surveys, with questionnaires sent by post. This is a frequently used method, but often suffers from *nonresponse*.
- Telephone surveys. Here, there is some concern about coverage — an obvious point is that a telephone survey will only be valid in countries where most people own phones.

Introduction to questionnaire design

Many aspects of questionnaire design are basically matters of common sense:

- The aims and objectives of the survey should be clear.
- The questionnaire should be well presented with clear instructions.
- The questionnaire should not be so long as to discourage respondents from participating.
- A questionnaire should be trialled initially to weed out unsatisfactory questions or to pinpoint any important gaps.
- Do not frighten the respondent with tough questions immediately.
- Remember that the responses will have to be analysed.

You may be able to think of other general points. The **wording** of questions is also important. Again some aspects are “obvious”. For example:

- Use simple language.
- It is usually better to ask relatively short questions than longer ones, so that the meaning is clear.
- Do not ask two or more questions in one (*e.g.* ‘do you think women and children should be given the first available flu jabs?’).
- Hypothetical questions (‘if ... what would ...’) tend to get hypothetical answers.
- Avoid subtle constructions that involve multiple negatives *e.g.* ‘Teachers should not be required to supervise pupils during meal times — agree or disagree?’ (the interviewee thinks ‘teachers should not not ...’).

Again, you can probably think of other common sense rules for wording questions. However, wording can influence responses considerably in subtle ways. Some examples are given below.

Open vs. Closed Questions. An open question is one where the subject has complete freedom in how to respond. A closed question forces them down (one of) some predetermined directions. Examples:

- People look for different things in a job. What do you most prefer in a job?
People look for different things in a job. Which of the following five things would you most prefer in a job - work that pays well; work that gives a feeling of accomplishment; work where there is not too much supervision and you make most of the decisions yourself; work that is pleasant and where the people are nice to work with; or work that is steady with little chance of being laid off?

- How often do you drink wine?

How often do you drink wine - never; once a year or less; every few months; once a month; every few weeks; once a week; several times a week; or daily?

Do not Know/No Opinion Option. It is often useful to allow interviewees explicitly to state that they do not know, they do not have an opinion or that none of the available options apply to them. For example:

Here is a statement about another country. Not everyone has an opinion of this. If you do not have an opinion just say so. Here's the statement: The Russian leaders are basically trying to get along with America. Do you have an opinion on that? (If YES:) Do you agree or disagree?

Here is a statement about another country. The Russian leaders are basically trying to get along with America. Do you agree, disagree or do you not have an opinion on that?

Do you include a Middle Category? For example:

Should divorce be easier or more difficult to obtain than it is now?

Should divorce be easier to obtain, more difficult to obtain, or stay as it is now?

Agree/Disagree or Forced Choice? For example:

- Please say whether you agree or disagree with this statement: individuals are more to blame than social conditions for crime and lawlessness in this country.

Please say whether you agree or disagree with this statement: social conditions are more to blame than individuals for crime and lawlessness in this country.

Which in your opinion are more to blame for crime and lawlessness in this country: individuals or social conditions?

- Would you favour a law that would require a person to obtain a police permit before he/she could buy a gun?

Would you favour or oppose a law that would require a person to obtain a police permit before he/she could buy a gun?

Would you favour a law that would require a person to obtain a police permit before he/she could buy a gun, or do you think such a law would interfere too much with the right of citizens to own guns?

- Would you say that most men are better suited emotionally for politics than are most women, that men and women are equally suited, or that women are better suited than men in this area?

Do you agree or disagree with this statement: Most men are better suited emotionally for politics than most women.

Stages in questionnaire design. In constructing a questionnaire, it may be helpful to proceed in the following way:

1. Clarify the *research problem*. Identify *concepts* that are to be *measured*.
2. Clarify the concepts by constructing *working definitions*. For each concept, decompose its definition into components and search for *items* that are (i) *measurable* and (ii) have a *close association* with the concept. At this point it is a good idea to conduct a small-scale qualitative study to gain information on the concepts.
3. Analysis stage. Only select questions that yield data that are amenable to *analysis* (e.g. ‘open’ questions may cause problems); decide on coding methods and computer software.
4. Finally, before finalising a questionnaire, a *pilot study* should be conducted to test its ability to do the job for which it is designed.