BASIC BIOMETRIC TECHNIQUES

IN

EXPERIMENTAL DESIGN AND ANALYSIS

April 2005

Biometrics Unit
International Livestock Research Institute
Nairobi

Biometrics Unit
ILRI Nairobi

Chapter 1 Introduction

Animal research comprises both experiments and surveys conducted either on-station, or on-farm or in the field. Such research has to be properly designed, conducted, statistically analysed and interpreted. Typically, comparable groups of animals are subjected to specified treatments for some predetermined period and the results of these imposed treatments, as measured by one or more variables, are used for estimating different types of responses.

The scope of this course is to introduce some of the basic principles of experimental design and statistical analysis. The purpose of good experimental design is to ensure that the results obtained are free of bias and that interpretations can be made which are uncomplicated by the existence of uncontrollable factors. Statistical methods, from the mere calculation of treatment means and standard errors to the more elaborate techniques of analysis of variance, are useful and necessary tools for the researcher. But only experiments of sound design permit sound statistical analysis and interpretation.

In this course we shall start with experimental design and then go on to discuss the structures of data that we collect in experiments and how we can explore patterns and variation. Then we shall see how a simple statistical model is formed and how it is used to produce an analysis of variance or regression. Worked examples of analyses of variance will illustrate the importance of good experimental design. Genstat will be taught and used in some of the practicals.

Example data sets

a) The following data set in Table 1.1 (data set A), which is artificial, will be one of two sets of data used throughout much of this course.

It describes an experiment carried out to study the effect of supplementation of weaned lambs on their health and growth rate when exposed to helminthiasis. Sixteen Dorper (breed 1) and 16 Red Maasai (breed 2) lambs were treated with an anthelmintic at 3 months of age (following weaning) and assigned at random within 'blocks' of 4 per breed ranked on the basis of 3-month body weight to supplemented and non-supplement groups. Thus, 2 lambs from each block were assigned at random to supplemented and non-supplemented groups. All lambs grazed on pasture for a further 3 months. At night they were housed and lambs in the supplemented group were fed cotton cake and bran meal.

Table 1.1 Data set A (Dorper / Red Maasai supplementation trial) to be used throughout these notes

Record	ID	Breed	Sex	Supp-lement	Block	Weight at 3m (kg)	Weight at 6m (kg)	PCV (%)	FEC (epg)	Weight gain (kg)
1	349	1	2	1	1	8.0	8.9	10	6500	0.9
2	326	1	2	1	1	9.0	10.1	11	2650	1.1
3	393	1	1	1	2	12.0	12.6	22	750	0.6
4	71	1	1	1	2	12.3	14.6	15	5200	2.3
5	271	1	1	1	3	13.0	13.7	19	4800	0.7
6	382	1	2	1	3	15.5	16.8	24	2450	1.3
7	85	1	2	1	4	16.3	18.2	27	200	1.9
8	176	1	2	1	4	15.9	17.7	21	3000	1.8
9	286	1	2	2	1	11.0	13.6	21	1600	2.6
10	183	1	1	2	1	9.9	11.7	21	450	1.8
11	21	1	2	2	2	11.6	13.1	25	2900	1.5
12	122	1	1	2	2	12.5	14.8	25	300	2.3
13	374	1	1	2	3	14.6	17.9	19	2250	3.3
14	32	1	2	2	3	14.2	16.9	22	2800	2.7
15	282	1	2	2	4	16.3	20.2	20	750	3.9
16	94	1	1	2	4	16.7	17.7	13	5600	1.0
17	127	2	2	1	1	7.5	8.1	26	1350	0.6
18	216	2	2	1	1	8.2	9.3	19	1150	1.1
19	133	2	1	1	2	10.1	11.7	30	200	1.6
20	249	2	1	1	2	8.8	10.4	28	0	1.6
21	123	2	2	1	3	1.6	12.6	23	600	1.0
22	222	2	2	1	3	11.3	13.5	24	1500	2.2
23	290	2	2	1	4	12.3	14.3	22	1950	2.0
24	148	2	1	1	4	13.1	14.9	26	500	1.8
25	142	2	2	2	1	8.2	11.5	25	850	3.3
26	154	2	2	2	1	9.5	12.2	35	700	3.7
27	166	2	1	2	2	9.7	12.8	29	400	3.1
28	322	2	1	2	2	8.6	12.0	26	800	3.4
29	156	2	1	2	3	10.2	13.0	28	1550	2.8
30	161	2	2	2	3	11.2	14.6	22	550	3.4
31	321	2	1	2	4	12.1	15.9	25	1250	3.8
32	324	2	1	2	4	13.8	18.1	24	1100	4.3

Data recorded included body weight at 3 months of age and body weight, packed red cell volume (PCV) and faecal egg count (FEC) at 6 months of age.

Some of the questions we will attempt to answer will be:

- Did supplementation improve weight gain?
- Did supplementation affect PCV and FEC?
- Were there any differences in weight gain, PCV or FEC between breeds?
- Is the experiment well designed?
- What size of experiment (number of animals per group) would have been sufficient to detect differences between breeds and diets?

b) The second data set (data set B), shown in Tables 1.2 and 1.3, is from a vaccine immunisation experiment. It involves the evaluation of what is referred to as a'trypanosome recombinant ARF1 protein' for testing its protective capacity incattle against a *Trypanosoma congolense* challenge. (These data are included withthe kind permission of Noel Murphy). Seven Boran cattle aged between 4 and 6 months were immunised subcutaneously with 250 μg of ARF1 and boosted with the same amount of antigen three times at 28-day intervals. Two control groups of 7 animals were also used. The first was immunised with a recombinant p32 from *Theileria parva*, purified in the same manner as ARF1, to ensure that histidine amino acids, to which the proteins ARF1 or p32 are both tagged during the purification process, do not contribute to the protection. This tagging to histidine amino acids is referred to as a "Histag". The second control group was not immunised.

   Fourteen days after the final boost each animal was challenged introvenously. Body weights were recorded weekly up to day 139 post infection. Parasitaemia and PCV were measured daily for the first 21 days and twice a week thereafter. The data set illustrated here shows the form that the data took.

   The questions to be asked in the analysis will be:

   - Did ARF1 immunisation offer protection?
   - Was this protection unrelated to the His tag?
   - How was protection, if it occurred, reflected in terms of weight gain and reduced development of anaemia?
   - Again, was the experiment well designed and were there sufficient numbers of animals used?

Table 1.2 Data set B (part 1) to be used in these notes. The table shows body weights measured at approximately weekly intervals in immunisation experiment.

		Date of		Imm.	Initial	Body weight (kg)												
Record	Animal	birth	Block	group†	Wt. (kg)	0	12	19	.....	49	56	63	69	76	.....	125	132	139
1	4	04/06	1	A	142	180	178	180	.....	188	182	182	184	178	.....	182	184	182
2	6	18/07	2	A	112	148	140	148	.....	162	160	162	162	162	.....	168	170	170
3	8	01/07	2	A	120	150	142	152	.....	148	142	140	134	138	.....	0	0	0
4	11	04/07	2	A	122	156	148	152	.....	162	162	170	170	172	.....	180	180	182
5	16	12/06	1	A	132	160	116	110	.....	102	98	96	98	0	.....	0	0	0
6	19	08/07	2	A	122	160	152	160	.....	160	152	156	150	152	.....	0	0	0
7	20	24/06	1	A	108	147	124	132	.....	132	132	128	124	0	.....	0	0	0
8	1	05/07	2	B	114	139	136	138	.....	140	138	136	132	0	.....	0	0	0
9	5	01/06	1	B	160	198	194	198	.....	200	196	198	196	202	.....	202	200	200
10	9	05/07	2	B	136	172	168	174	.....	176	174	182	176	180	.....	170	168	168
11	12	15/07	2	B	100	118	116	120	.....	132	138	138	138	142	.....	148	152	148
12	14	02/06	1	B	140	200	200	200	.....	212	202	206	200	196	.....	194	196	196
13	17	04/07	2	B	136	210	200	200	.....	210	200	204	206	206	.....	216	214	216
14	18	20/06	1	B	112	130	142	152	.....	148	146	148	152	152	.....	154	154	154
15	2	28/06	2	C	122	152	144	152	.....	144	146	146	138	140	.....	0	0	0
16	3	06/06	1	C	*	144	140	144	.....	146	140	146	134	136	.....	116	116	0
17	7	28/07	2	C	110	132	126	128	.....	142	144	150	150	156	.....	162	162	160
18	10	10/06	1	C	124	120	120	126	.....	118	114	110	112	108	.....	0	0	0
19	13	09/06	1	C	120	148	144	148	.....	166	160	162	160	164	.....	176	178	180
20	15	02/07	2	C	114	138	146	154	.....	136	132	130	134	126	.....	124	124	122
21	21	01/07	2	C	126	164	156	158	.....	160	148	148	0	0	.....	0	0	0

* not recorded

† A = ARF1; B = p32; C = negative control

Table 1.2 Data set (part 2) to be used in the notes. The table shows packed red cell volume (PCV) measured at approximately weekly intervals in immunised experiment.

		Date of		Imm.	Initial	PCV (%)													
Record	Animal	birth	Block	group†	PCV (%)	7	9	10	11	…….	42	46	49	53	56	60	……..	136	139
1	4	04/06	1	A	39.2	38.3	35.6	31.6	34.4	……	20.4	19.8	19.8	17.9	18.2	18.8	…….	23.4	25.5
2	6	18/07	2	A	33.6	34.0	31.6	31.3	31.0	……	20.4	21.0	20.4	18.5	22.5	21.3	…….	21.6	22.8
3	8	01/07	2	A	32.5	34.7	30.1	30.1	28.6	……	16.7	16.1	16.4	14.9	17.0	15.5	……..	0.0	0.0
4	11	04/07	2	A	30.7	33.4	31.6	29.8	28.3	……	15.8	21.9	19.2	20.7	22.8	21.3	…….	20.1	21.9
5	16	12/06	1	A	31.2	31.3	31.6	28.6	28.0	……	14.9	14.6	13.7	12.2	12.2	12.5	…….	0.0	0.0
6	19	08/07	2	A	36.5	35.3	35.3	31.0	29.8	……	16.1	16.7	16.1	15.5	16.1	16.7	…….	0.0	0.0
7	20	24/06	1	A	29.8	28.0	27.1	27.1	26.1	……	12.5	12.2	12.2	11.9	11.9	11.9	……	0.0	0.0
8	1	05/07	2	B	31.8	31.6	31.3	30.4	31.0	……	12.5	13.4	13.1	12.8	14.3	12.8	……	0.0	0.0
9	5	01/06	1	B	36.8	37.4	30.4	35.3	35.3	……	16.7	16.1	17.6	19.2	19.2	18.8	……	21.0	19.5
10	9	05/07	2	B	34.9	33.1	33.4	29.8	30.1	……	19.8	18.8	19.2	20.4	20.4	19.5	……	21.6	22.8
11	12	15/07	2	B	30.0	32.5	30.4	30.4	28.9	……	20.4	21.6	21.3	19.8	21.6	21.9	……	24.6	25.2
12	14	02/06	1	B	36.7	37.4	35.6	33.1	34.7	……	21.0	21.0	18.5	17.6	18.2	19.2	……	17.9	19.5
13	17	04/07	2	B	36.2	35.6	34.7	34.7	33.7	……	19.8	19.5	21.0	18.2	21.0	21.3	……	27.4	28.3
14	18	20/06	1	B	33.7	34.0	31.0	28.9	27.4	……	18.2	18.5	19.5	21.0	20.1	20.7	……	20.4	18.2
15	2	28/06	2	C	30.6	32.5	31.6	30.4	30.4	……	13.7	13.1	14.3	13.4	11.9	13.7	……	0.0	0.0
16	3	06/06	1	C	28.9	31.6	32.5	31.3	31.3	……	21.0	17.0	16.4	14.3	14.9	16.7	……	15.8	0.0
17	7	28/07	2	C	31.9	31.3	32.2	30.7	28.9	……	29.2	28.3	30.7	29.2	28.0	28.0	……	28.3	28.3
18	10	10/06	1	C	25.3	27.7	25.2	24.6	24.0	……	16.7	15.2	14.3	14.0	14.6	14.3	……	0.0	0.0
19	13	09/06	1	C	30.4	31.6	31.0	30.4	28.0	……	30.4	31.6	31.6	31.3	31.3	31.6	……	28.6	27.4
20	15	02/07	2	C	35.6	34.0	33.1	30.1	29.5	……	15.5	13.7	15.2	14.6	15.2	17.3	……	17.9	19.8
21	21	01/07	2	C	33.3	34.4	30.7	30.4	28.0	……	13.4	12.2	12.5	12.2	13.1	12.8	……	0.0	0.0

* not recorded

† A = ARF1; B = p32; C = negative control

Chapter 2 Experimental design

Before attempting to analyse data it is necessary to understand how the experiment, from which the data have been obtained, is designed. If the experiment has not been designed well then it may not be possible to undertake a satisfactory analysis of the data. Before discussing methods of statistical analysis we shall firstly briefly review important aspects of good experimental design.

Objectives

Before designing an experiment it is important that the objectives for the experiment are clear, well-defined, realistic and relevant. This may seem obvious. However, from observations in the ways that IAUCC forms are sometimes completed at ILRI, it is often apparent that the researcher does not always given this sufficient thought. So, objectives should be :

- **Clear**. If the objectives are vague it will be difficult to know how to go about planning an experiment.

- **Well-defined**. If the objectives are not carefully stated then it will not be clear what hypotheses are to be evaluated.

- **Realistic**. The researcher needs to be confident that an experiment can be designed that meets the objectives.

- **Relevant**. The objectives for the experiment need to be relevant to the problem in hand. In other words the researcher will be a step nearer to solving the problem once he/she has the results from the experiment.

All this may seem a little obvious but nevertheless it is very important. Indeed, successful planning of an experiment can lead to a revised set of objectives once it becomes clear that the original objectives are unrealistic in terms of the numbers of animals that may be required.

Exercise 2.1. Here are two examples of vague objectives.

1. *To evaluate the effects of increased concentrate feeding on reproductive performance.*
2. *To evaluate a new method for controlling disease.*

Choose one of these and adapt the objective to an experimental situation that you might be familiar with and write more clear and precise objectives.

Treatments

Most experiments will involve some form of '**treatment'**. This could be a method of immunisation, a form of chemotherapy, an alternative diet or even a different crop variety or animal breed. Many objectives will simply require the comparison of mean

responses to different treatments, and one of these treatments might be a control. In certain cases more complex comparisons are needed. For example, different doses may need to be compared for a particular treatment. In this case the experiment may include three dose levels, and depending on the experimental objective, it may also be necessary to include a control. A control can take various forms. Sometimes it can be a zero treatment in which nothing is done to the animal. Often, however, a placebo treatment will be needed to ensure that the appropriate hypothesis is being tested. Thus, in an on-farm experiment the control could be the farmer's normal practice. In the second Example B two controls were used. The p32-immunised control was used to evaluate the hypothesis that the His tag was not influential in any protection that may be observed. The non-immunised control was to determine the response to infection without any form of immunisation.

Exercise 2.2 Plan a suitable control (or controls) for an experiment that might be planned to answer the objective you set yourself in Exercise 1.

Factorial treatments

A **factorial** treatment structure is a particularly useful concept that can be used to result in efficient experimental design. Such a structure refers to a set of treatments formed from a combination of two or more different treatments, each with two or more levels. When used in this way the different treatments are often referred to as **factors**. Such factorial combinations often arise naturally from the proposed hypothesis, but they can also be used to test other unrelated hypotheses more efficiently in one experiment than in separate ones. For example, two different formulations of an experimental vaccine may require to be compared. At the same time the researcher is interested in evaluating alternative delivery systems for the vaccine. He could either test each hypothesis in turn in consecutive experiments or alternatively test both hypotheses simultaneously by using each delivery system to vaccinate half of the animals receiving each vaccine formulation. The second method is more efficient. Not only may it require fewer animals to be used in total but it also allows the **interaction** between vaccine formulation and delivery system to be studied at the same time. By the term interaction we mean that the type of delivery system influences the efficacy of one vaccine formulation more than the other. If there is no interaction then it means that any difference between delivery systems is the same for each vaccine formulation. Such a design is referred to as a 2 x 2 factorial.

Most experiments feature as part of a long-term research plan. Thus, in designing an experiment one needs to take into account not only the hypotheses that require immediate testing but also others next in line to see whether any can be brought forward to be encompassed within the current one.

Exercise 2.3. Describe the treatment factorial structure in Example data set A. Is there a treatment factorial structure in Example data set B?

Replication or blocking

Estimating the precision of the results obtained from an experiment depends on the level of individual variation. Therefore, it is necessary to be able to calculate this level of variation in the data that one is collecting. In order to do this there must be data from more than one animal in each treatment. Thus the treatments must be replicated.

Very low levels of replication such as two animals per treatment may not be sufficient to give an adequate estimate of the standard deviation (a measure of variation) nor to give reliable estimates of treatment means. The precision of an experiment is increased by increasing the number of times each treatment is replicated. Precision is also increased by making sure that animals are as homogeneous as possible, e.g. in terms of weight, age and breed. If there is some variation in weight, for example, then the principle of '**blocking'** can be applied. By putting animals into blocks of similar body weight and randomly selecting animals from within each block for the different treatments, it is possible that the residual, uncontrollable variation among animals is reduced. Often there are several choices for suitable replicates or blocks, e.g. body weight, age, sex. The use of twins is another useful form of blocking with each twin member receiving a different treatment. Researchers are often tempted to make treatment groups similar in all respects by balancing all attributes across treatments. This is wrong and can introduce systematic errors. This method also gives no scope for randomisation of animals to treatments. In practice it is preferable to choose one or, at the most, two attributes for blocking. The choice will depend on the researcher's knowledge of how the attribute is likely to influence the response variable that he/she is primarily interested in. Thus, sex is likely to influence weight gain, but less likely to influence packed blood cell volume.

It is essential to allocate treatments at random to animals. This is to eliminate any subjective bias which may occur (no matter how honest one tries to be).
Randomisation minimises the risk of systematic errors and helps to ensure that each treatment is represented fairly in the trial. The aim of randomisation is to give each animal an equal chance of being allocated to any of the treatments.

Blocking may also be done in time. It may be that resources or facilities limit the number of animals that can be studied at one time. To achieve the required numbers of animals per treatment the experiment can be repeated on separate occasions. Such replicates in time are used in the analysis in precisely the same way as blocks based on an animal's attribute such as weight. Treatments are randomised to animals within each time replicate. Blocking on weight, for example, can also be undertaken within each batch or time replicate. Thus, we can have two levels of blocking.

Exercise 2.4. Animals in Example B were blocked on age before assigning them to treatment. Considering the range in dates of birth and that the animals were 8 to 9 months of age at the time of challenge, do you think this was a reasonable idea? Are there any other forms of blocking that you might have considered? Think of an experiment that you have been involved in. Can you remember any form of blocking that might have been applied? Do you recall how animals were randomised to treatments?

Numbers of animals

The numbers of animals to use will be based on previous knowledge of the likely variation to be expected. If a researcher does not have sufficient resources he/she will have to consider either simplifying of the objectives, carrying out the experiment in stages or even abandoning the experiment altogether. There is little point in proceeding with an experiment if it is fairly clear that is cannot meet the objectives!

Experiments should be kept as simple as possible in terms of design, execution and analysis. The most efficient designs are often the most simple to analyse. For simplicity of analysis there should, where possible, be equal numbers of animals assigned to each treatment. Such designs are known as **balanced** designs. While this is usually relatively easy to organise for crops, it is not always so easy with trials on livestock, especially on farms. Sometimes the researcher may wish to increase the sample size for one treatment because of anticipated increased variability in responses compared with a control. **Unbalanced** experiments lead to more complicated statistical analyses which are often more difficult to interpret. Sometimes the analysis has to be done that way, either because it was not possible to make the experiment completely balanced in the first place, or because animals may have died during the course of the experiment. Most livestock experimentation leads to some form of imbalance. The art of good experimental design is to try and reduce such imbalance to a minimum.

Practical arrangements for assigning animals should be simple to avoid problems, confusion and errors when implementing an experiment. However, this should not be at the expense of other considerations, such as avoiding systematic errors. For example, with an on-station trial in individual pens, it might be convenient to put all animals on the same diet in adjacent pens. Animals on different diets will then be in different parts of the barn. However, the barn may not be a uniform environment; some areas may be more exposed to, say, wind and cold air. The differences between diets may be **confounded** with differences in environment, making the results impossible to interpret. By ‘confounded’ we mean that it is impossible to distinguish the effect of one factor from another, in this instance between diet and location in the barn. Randomisation should be used to overcome these sorts of problems, even though it may introduce some practical complications.

Chapter 3 Recognising types of data and data structures

Research data consist of observations or measurements recorded on units e.g. animals. Although we talk about **experimental units** in experiments, in surveys we talk about **sampling units** and in observational studies we often talk about **observational units**. To generalise in a way that we can cover each type of study we shall refer to these units generically as **investigational units**. An investigational unit refers to a unit, whether it be a farm, an animal or sample within an animal, in which a measurement is made independent of another. Measurements made on these units such as farm size, body weight, growth, PCV, FEC are called **variables**.

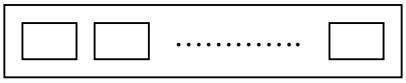
Variables may be of different types and it is important to consider the type of data we are dealing with. In Example A the response variables that have been measured, namely body weight, growth rate, PCV and FEC, are known as continuous variables, i.e. they can take any value within a reasonable range to a given accuracy. Thus, weight gain ranges from 0.6 to 4.3 kg. Some of the measurements used to characterise the lambs prior to the experiment are discrete, i.e. they can only take on only specific not continuous values. Thus, sex is recorded as 1 (male) and 2 (female). These are the only two values that sex can take. Sometime response variables that are measured during the experiment take on the value 0 or 1 only. This could be, for example, mortality or disease. The methods of analysis for continuous and discrete variables are very different and need to be considered carefully when planning a trial. In general, a larger number of investigational units is needed to investigate statistically significant differences between treatments when the outcome is discrete (e.g. death) than when it is continuous (e.g. weight gain).

Next we must consider how the data are structured. Often we find that we have types of units occurring at different layers in a study, e.g. farms, plots within farms; or farms, animals within farms; or animals, repeated samples within animals. We often take measurements at different layers, e.g. at the farm level we may record certain attributes about the household, e.g. numbers of members, size of acreage etc., and at the animal level we may measure milk offtake on different occasions. It is important to understand the structure of a data set compiled during a research study, both in planning the study and in designing the statistical analysis.

Let us diagramatically see how the Dorper/Red Maasai experiment (data set A) is structured. The first step is to decide what are the investigational units and whether these occur at more than one layer. The investigational unit is that unit which can be considered to have been assigned at random to treatments. The only true randomisation that takes place in this experiment is the assignment of lambs to diet. Lambs cannot be assigned at random to breeds as a lamb's breed is part of its genetic make up. The trick though is to recognise that the lambs to be used in the experiment for each breed can be assumed to be a random selection from that breed. Thus, breed features as a special type of 'treatment'.

To summarise, randomisation only takes place at the lamb level – this is both within breed and diet. Thus, the structure of the data comprises just one layer. Later on we shall see an example of a data structure with more than one layer.

We can think of the breed of a lamb and the diet fed to it as being particular attributes for that lamb. Lambs were also ranked within each breed according to their body weight and assigned to one of four blocks. This then is a third attribute. The data can thus be described as follows.

Layer	**Investigational unit**		**Attributes**
1	Lamb		Breed Diet Breed / block

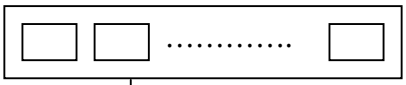
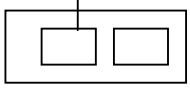
Note the notation 'breed / block'. The slash notation means block within breed. The three attributes, namely breed, block / breed and diet, are known as **fixed effects**. Lamb can be described as a **random effect**. The fixed effects describe the patterns in the data - we are interested in determining the mean values calculated from the data for each level (e.g. Dorper, Red Maasai) of an effect (or attribute). For a random effect we are not so interested in each level, but instead the overall variation in the data expressed by the different lambs. In this case the random effect represents the residual variation among lambs that remains and cannot be explained by the pattern expressed by the fixed effects. This idea can be represented in the form :

data = pattern + residual

To demonstrate how things can become a little more complicated, suppose that this experiment was designed differently. Instead of lambs selected at random, suppose that hyperthetically 8 ewes, each with twins, were selected for each breed. Instead of blocking lambs according to body weight assume that pairs of twins were assigned, respectively, at random to one of the two diets.

There are now two levels at which we can consider randomisation to have taken place – firstly 8 ewes are selected from each breed, and we assume this to be a random selection, and, secondly, twins from each ewe are assigned at random to one of the two diets.

Thus, the complete data structure can be described as :

Layer	**Investigational unit**		**Attributes**
1	Ewe		Breed
2	Lamb		Diet Breed.diet

In this example, therefore, we need to think more carefully about investigational units. Randomisation at the lamb within ewe level ensures that lamb is the investigational unit for examining the effect of supplementation. But what about breed? The 16 lambs used within each breed are not independent. They form 8 pairs of twins. The animals initially selected from each breed were first and foremost the ewes. The investigational unit for comparison of breed is, therefore, the ewe or the pair of litter mates. Thus, there are in this case two sets of investigational units – the ewe (or pair of twins) for comparison of breeds, and the lamb for comparison of diets. This may be a little difficult to understand when none of the measurements are made on the ewe itself. Here, however, we can represent the 'ewe measurement' by the average weight of its two offspring. Breed means are compared in this variation among ewes.

The expression

data = pattern + residual

still applies but the residual has two components : ewe within breed and lamb within ewe. The pattern is still described by fixed effects associated with breed and diet. This type of structure is **hierarchial** or **multilevel** in nature. Note that in the above figure we have also included an **interaction** term breed.diet (we could have also included it in the first example). The interaction term determines whether the effect of supplementation differs according to the breed of lamb. This attribute occurs at the lamb layer.

Exercise 3.1. Draw a diagram to show how the data in Example data set B are structured, taking into account that repeated measurements in time are taken on the individual animals.

We shall not deal with multilevel structures from now on in this course, but it is nevertheless important to be able to recognise the structure in one's data. It is also important to appreciate what type of data structure is likely to be produced when an experiment is being planned. This might have important consequences for sample size and statistical analysis.

Chapter 4. Exploring patterns and variation

Once we understand the structure of the data that we are handling we can start to identify the patterns and variations in the data.

Insufficient attention is usually given to preliminary investigations of data from experiments prior to their formal statistical analysis. All too often the researcher is keen to carry out a statistical analysis and get a P-value, without giving due attention to all the information provided in the data. With sensible preliminary investigations the researcher should be able to identify definite patterns in the data, gain insight into the variability in his/her data and detect any strange observations which need following up. He/she may even come across some unexpected patterns which he/she would wish to investigate further.

Preliminary investigations allow one to look closely at the data collected in one's experiment. As has already been described, data can be described using the expression:

data = **pattern** + **residual**

'Pattern' is the result of factors (or fixed effects) such as breed, diet and other attributes such as sex, which can influence a response, such as weight gain. Identifying the pattern, e.g. components due to breed and diet, is therefore an important part of the analysis. Other parts of the pattern may well be introduced by the way the experiment is set out – differences between blocks, for example.

'Residual' is the remaining variation that exists from animal to animal and that cannot be explained by the way the pattern is defined.

Both pattern and residual should be studied in initial investigations of the data. The results of these investigations will often help to define the way the final analysis is done. Useful tools include descriptive statistics and graphs such as boxplots and scatterplots.

One of the first steps in analysing research data should be to examine the frequencies of different values recorded for the variables measured. This will help to order the data in some way and to summarise their main features, and may also help to spot data errors or extreme values. We shall first use Example Data Set A.

Frequency table

A useful starting point is to examine the range spanned by the data. To do this one finds the lowest and highest value of a variable (e.g. PCV) and divides the range into a reasonable number of intervals. The numbers of data values that occur within each interval are counted and made into a table as shown below:

Table 4.1. Frequency distributions of PCV among the 32 lambs in Example A.

PCV (%) class interval	Frequency of animals	Relative frequency (%)
10 - 12	2	6
13 - 15	2	6
16 - 18	0	0
19 - 21	7	22
22 - 24	8	25
25 - 27	8	25
28 - 30	4	12
31 - 33	0	0
34 - 36	1	3
Total	32	100

Having calculated the number of animals in each class interval, the relative frequency, which is the percentage of values contained in each interval, can then be calculated. This shows how the data are distributed. The table shows that a majority of animals have PCVs between 19 and 30. Four animals have PCVs below and one animal has a PCV above this range.

As shown in the next table however, the frequency distribution for FEC is very different with the majority of values at the lower end of the range. The distribution is clearly not symmetric and, although this distribution may be partly influenced by the pattern associated with breed, diet etc., it is likely that the residual distribution is also asymmetric.

Table 4.2. Frequency distribution of faecal egg count (FEC) among the 32 lambs in Example A.

FEC (e.p.g.) class interval	Frequency of animals	Relative frequency (%)
0 - 700	12	38
800 - 1500	8	25
1600 - 2300	3	9
2400 - 3100	5	16
3200 - 3900	0	0
4000 - 4700	0	0
4800 - 5500	2	6
5600 - 6300	1	3
6400 - 7100	1	3
Total	32	100

Histogram

The above tables can be presented pictorially in the form of a diagram, known as a histogram. A histogram helps to identify the shape of the distribution. Sometimes the histogram is more or less symmetrical, with the bulk of the data gathered near the centre and the proportions of data on each side roughly balancing each other. Sometimes the histogram is **skewed**. This means that the data are rather bunched up to one side. The figure illustrates the skewed distribution of FEC. As already shown in the above table there are more data values to the left than the right. Biological data often belong to a **normal distribution**; the frequency distribution is `**bell-shaped**'. The distribution of PCV approximates to this shape. The distribution of FEC does not. As we shall see later some of the statistical methods that we shall use require data to be distributed normally. Thus, we shall need to take into account the fact that FEC follows a skewed distribution in our final analysis.

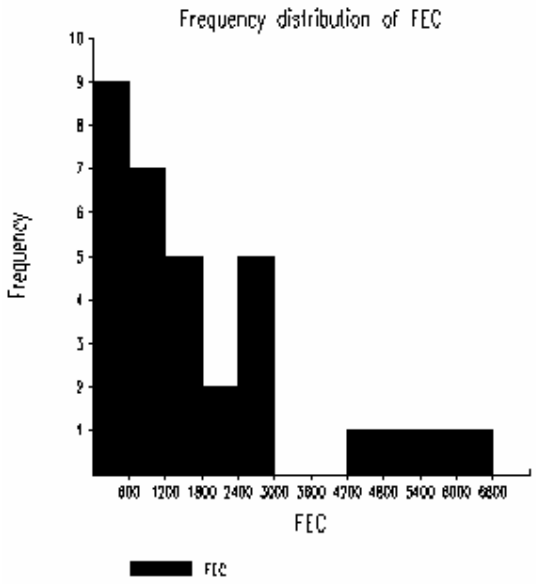


Fig. 4.1 Histogram of frequency distribution of FEC among the 32 lambs in Example A.

Scatter plots

Graphs derived by plotting one variable against another are commonly used to examine relationships between two variables. The next figure shows a scatterplot of PCV and FEC. Such a scatterplot can also indicate outliers (i.e. data values which are rather extreme and appear different from the others). Such outliers may be due to data errors. The figure shows one or two possible outliers. Let us assume that during data entry lamb 94 has had its PCV digits transposed by mistake from 13 (see data in Table 1.1) to 31. With such a high FEC its PCV value is somewhat away from the others towards the top right corner of the diagram. Examination of the input data shows that a mistake has been made in data entry. The value to the bottom of the graph, however, is as recorded.

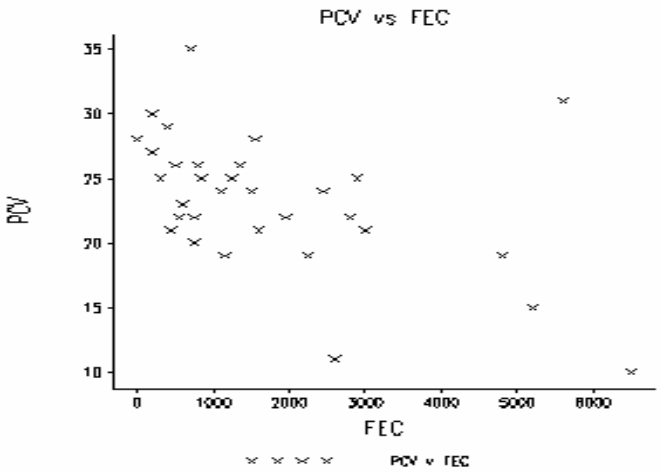


Fig. 4.2 Scatter plot of PCV versus FEC with mistake in PCV value (31 rather than 13%) for lamb 94 in Example A.

Box and whisker plots

Of course when all the data are described together as above variations due to diet and breed are hidden. If possible it is best to summarise the data in each group separately. A nice way to do this is to use what is known as Turkey's box and whisker plot, which displays the range, median and quartiles for each group alongside each other.

The **median** is the middle value. This differs from the **mean** which is the average of all the data values. The **upper** and **lower quartiles** are the points at which one quarter of the data values are above or below, respectively.

The following figure shows the box and whisker plot for FECs for each of the 4 breed x supplementation treatments. The middle horizontal line represents the median (i.e. the middle point). The box contains the middle half of the data (between the quartiles

represented by the horizontal lines at the extremes of the boxes), and the upper and lower vertical lines cover the rest of the data.

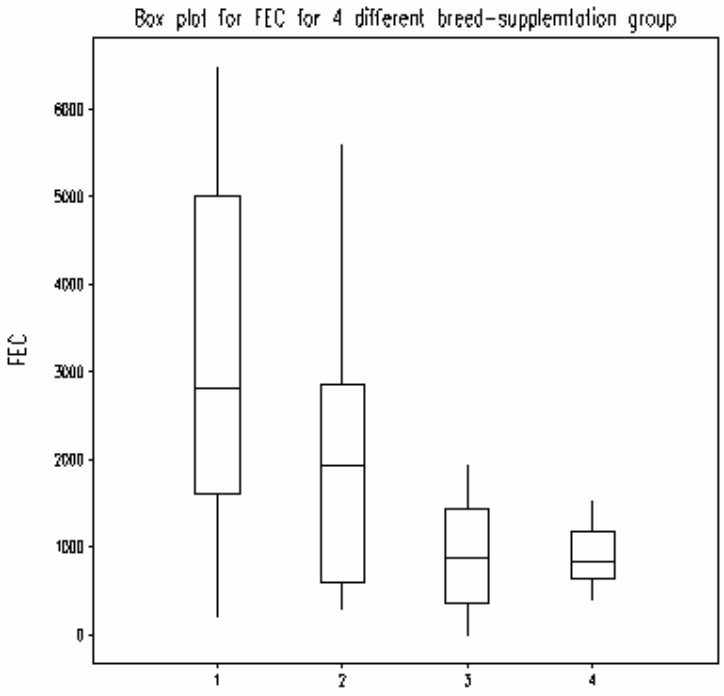


Fig. 4.3 Tukey's box and whisker plot for FEC for each breed x supplementation group in Example A. The codes are : 1 (Dorper / non-supplemented); 2 (Dorper / supplemented); 3 (Red Maasai / non-supplemented); 4 (Red Maasai / supplemented)

The boxplot shows two things. Not only is the mean FEC lower in the Red Maasai than in the Dorper lambs but so is the variation (we may need to take this into account in the way we analyse the data). The whiskers (the upper and lower vertical lines) also illustrate any skewness in the data. Thus, the boxplot is a useful tool for describing both pattern and residual.

We shall concentrate from here on mainly on the analysis of weight gain. The following box plot shows us that there is less to be concerned about with different levels of variation, except that there appears to be little more variation in weight gain amongst animals in the second group.

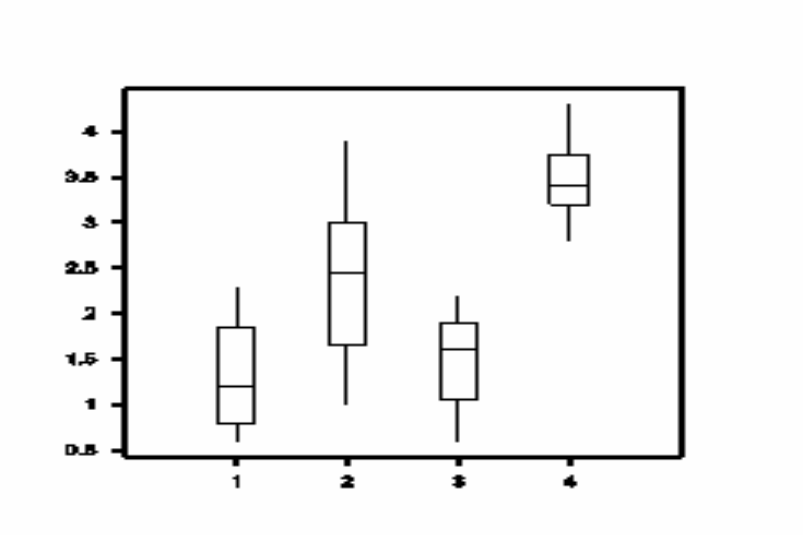


Fig. 4.4 Tukey's box and whisker plot for weight gain for each breed x supplementation group in Example A. (See Fig. 4.3 for description of group codes).

Plotting these data alternatively as a scatter plot enables us to examine the individual values.

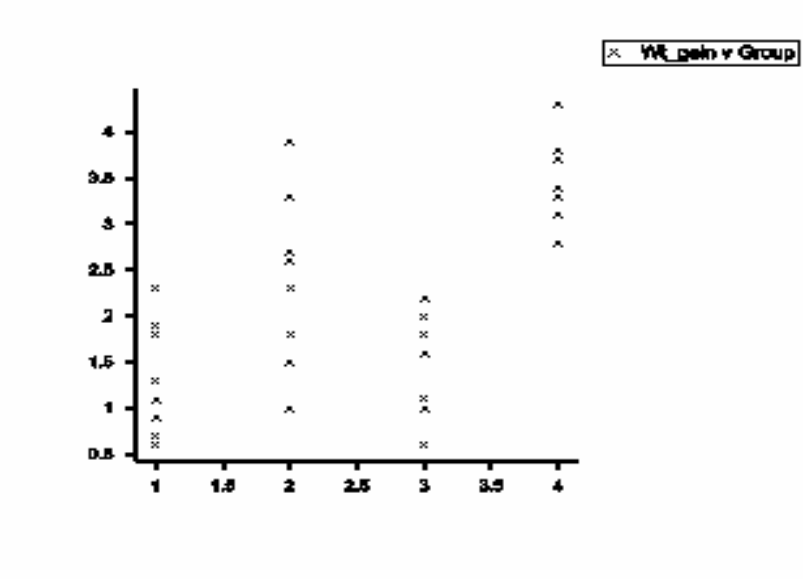


Fig. 4.5 Scatter plot of individual weight gains for each breed x supplementation Group in Example A. (See Fig. 4.3 for description of group codes.)

Blocking on initial body weight was a feature of the experimental design. To see what likely impact this has had in reducing the residual variation in weight gain we can produce a box plot for block.

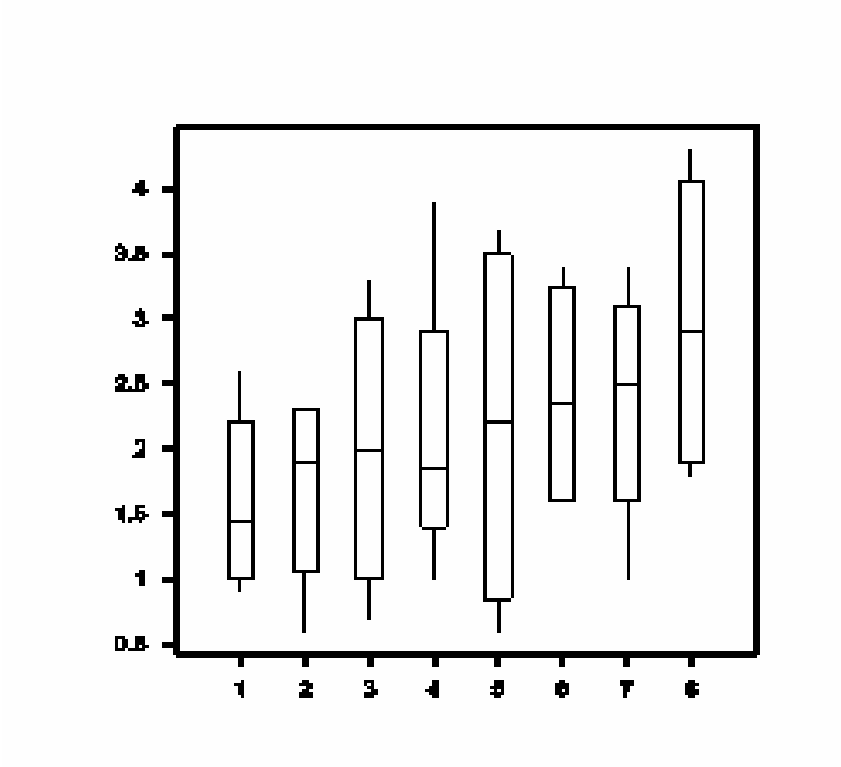


Fig. 4.6 Tukey's box and whisker plot for weight gain for each block defined for assignment of lambs to supplementation groups in Experiment A. Blocks 1-4 refer to Dorper and blocks 5-8 to Red Maasai lambs.

The plot shows that there is little variation among block means within breeds suggesting that in this experiment blocking had little impact. Thus, in retrospect, from a design point of view, completely randomising the lambs to the supplement and non-supplementation groups would have been just as satisfactory.

Sex might feature in the analysis as a factor since it is well known that males grow faster than females. However, the plot below suggests that the difference in weight gain for the two sexes is unlikely to feature as an important contribution to the pattern.

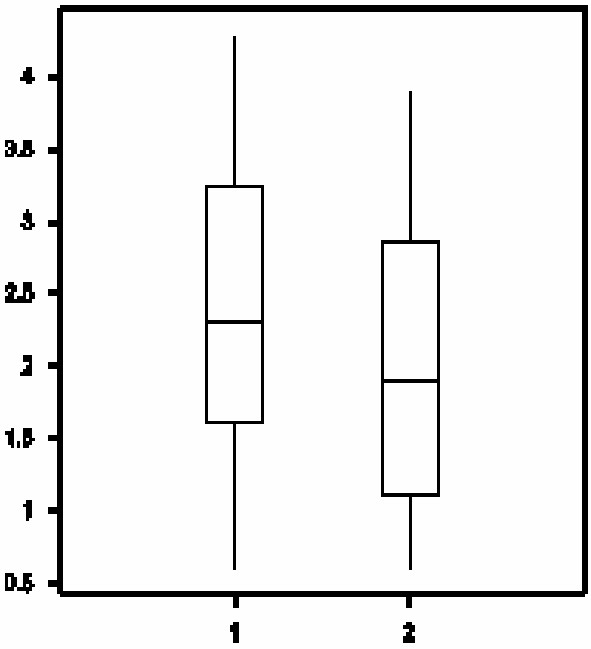


Fig. 4.7 Tukey's box and whisker plot for weight gain for each sex (1: males; 2: females) in Example A.

Table of means

The above methods of presentation allows one to see how the breed x supplementation groups interact. By interaction we mean that the effect of one factor (e.g. supplementation) is different at different levels of another (e.g. breed). It may also be instructive to produce a 2-way table of means to examine further how weight gain varies across both breed and diet. The table below demonstrates a clear effect of supplementation on weight gain and a smaller effect of breed (Red Maasai grew faster than Dorpers). Comparison between the four values within the body of the table suggests that the increase in weight gain due to supplementation might be slightly greater for the Red Maasai than the Dorper. We shall need to check this in the statistical analysis. The tables of means for PCV and FEC, however, show no evidence for an interaction.

Table 4.3 Tables of mean values for breed x supplementation groups for weight gain, PCV and FEC of lambs in Example A.

WTGAIN

	Mean		
SUPP	1	2	Mean
BREED			
1	1.325	2.388	1.856
2	1.488	3.475	2.481
Mean	1.406	2.931	2.169

PCV

	Mean		
SUPP	1	2	Mean
BREED			
1	18.62	23.00	20.81
2	24.75	26.75	25.75
Mean	21.69	24.87	23.28

FEC

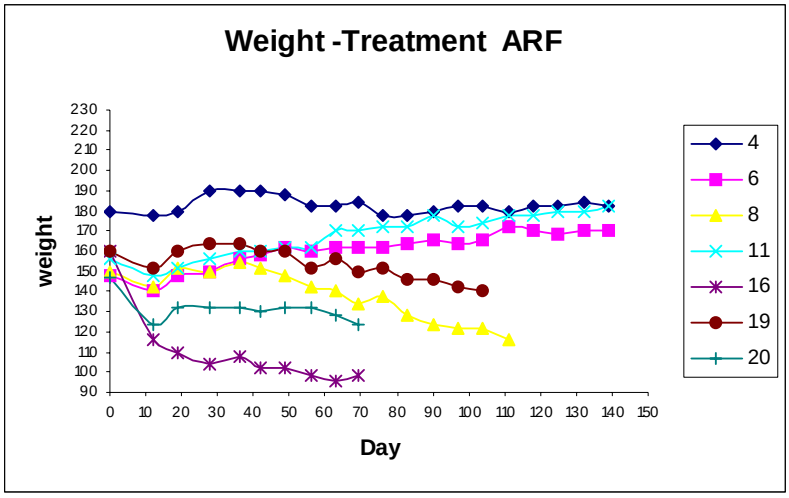
	Mean		
SUPP	1	2	Mean
BREED			
1	3194	2081	2638
2	906	900	903
Mean	2050	1491	1770

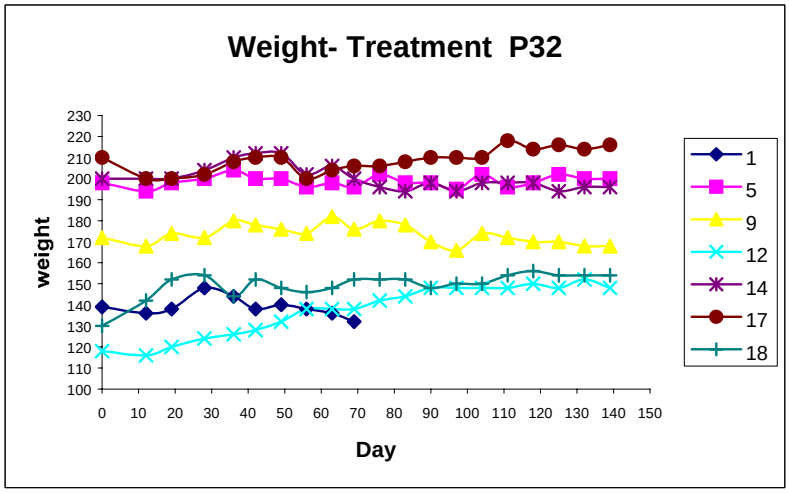
In conclusion, we have been able to carry out a lot of exploratory analysis using very simple techniques. We have seen how we can spot possible outliers in the data that may be due to mistakes in data entry. We have also had a good idea of the effect that input supplementation has had on response variables for each breed. There may or may not be an interaction between breed and supplementation in relation to weight gain.

We were able to examine the residual variation among observations, i.e. the variability which was unexplained by the diet x breed pattern, using scatterplots. We suspect that some of this residual variation is unlikely to be explained by differences between the blocks in the experiment, and also that there is unlikely to be a major sex effect. We have therefore been able to deduce quite a lot about the likely effects of our factors from this preliminary analysis.

What we have not been able to do is look at all the above attributes together. We have only looked at “slices” of the data - i.e. we separately looked at patterns due to breed and supplementation and the residual variability ignoring any effects of blocks, and then looked at the patterns due to blocks and to sex. It would have been more complicated to look at pattern due to blocks, sex, breed and supplementation together, and then inspect the resulting residual variability. To do this we need to take the analysis further and use methods such as analysis of variance. This is described in the next chapter.

Example B provides a different form of example. Data are not collected at two time points at the beginning and end of the experiment as in Example A but at frequent time periods during the course of the experiment. Exploratory analysis is particularly needed in this case to determine how to go about the statistical analysis. The most useful thing to do is to first plot the data as shown in Fig. 4.8. This figure shows body weight changes for each individual animal in the three treatment groups.





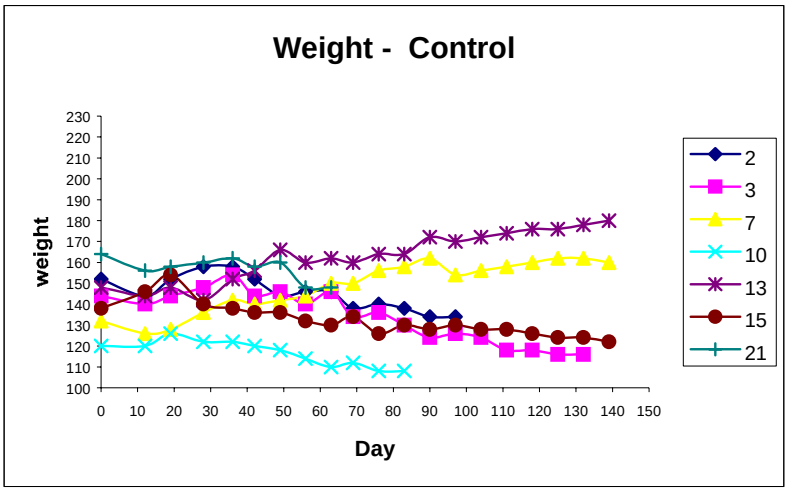
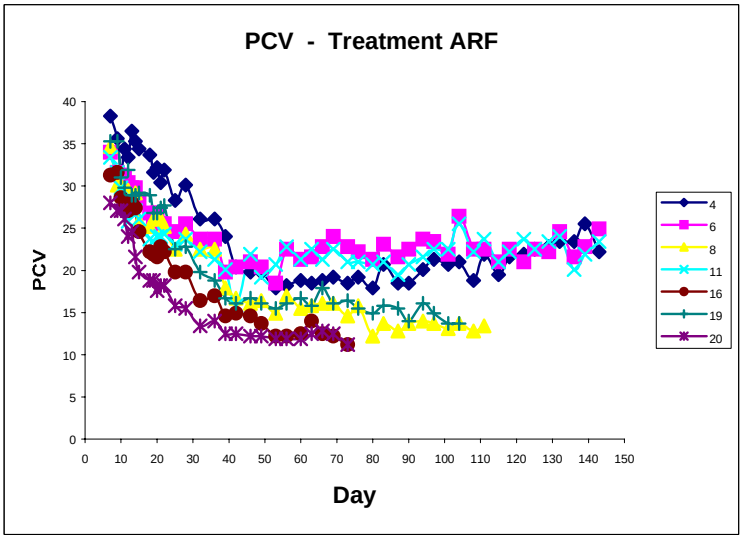
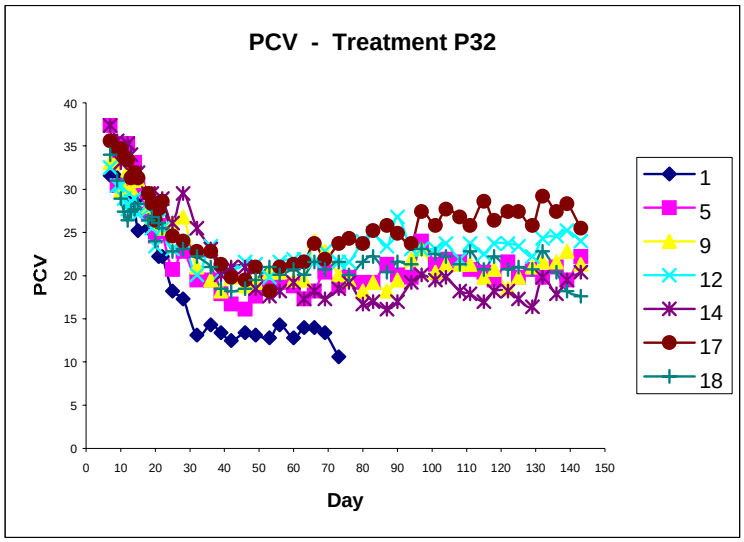


Fig. 4.8 Body weight changes for individual animals in the three immunised groups in Example B.

Exercise 4.1 Consider the graphs in the above figure. Comment on the overall trends both within and between the 3 treatment groups. Are there any questions that you would like to pose to the researcher in relation to any odd patterns in gains or losses in body weights?

The following figure shows similar graphs for PCV. The separation of PCVs among the control animals illustrates how the two animals 7 and 13 failed to be infected. There are one or two cases where a sudden rise or fall in PCV from one time point to the next for an individual animal may indicate a data error. These could be checked against the original recording sheet.





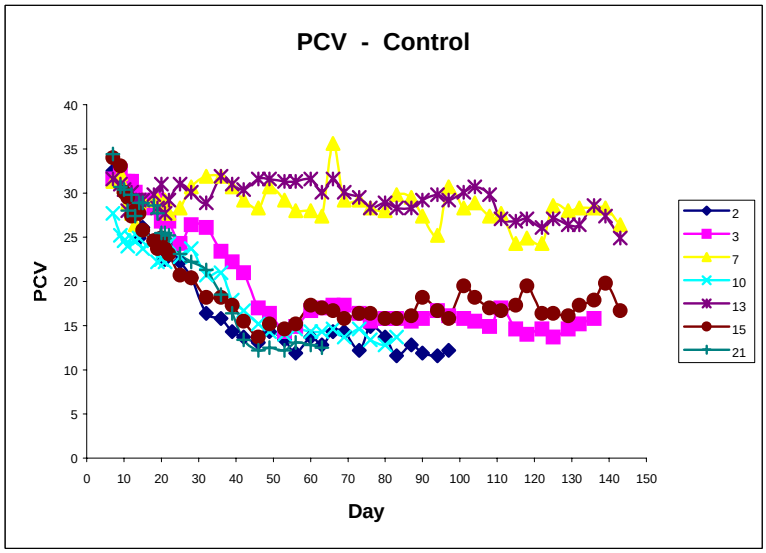


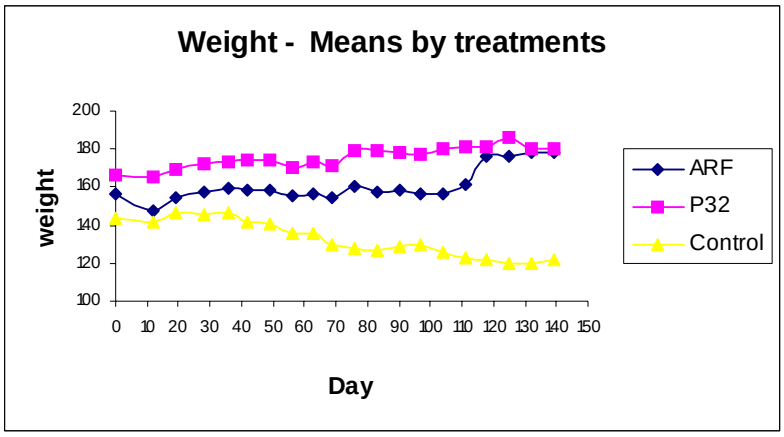
Fig. 4.9 Changes in PCV for individual animals in the three immunised groups in Example B.

There are different forms of **outliers** in this experiment, some that can justifiably be excluded in the analysis and others not so. An outlier is an observation, or group of observations, that does not conform to the general pattern determined by the other points. Thus, it is clear that the two non-infected control animals should be excluded because they did not conform to the experimental description of animals in that group. Animal 16 in the ARF1 group was also excluded because it rapidly lost weight for reasons that veterinary opinion believed were not due to trypanosomosis. Exclusion of this animal, however, was not so clear cut as the two controls since no definite diagnosis of the clinical condition was possible. Other possible outliers are the

individual PCV values mentioned above that show large deviations from their adjacent values.

As a general rule an outlier should not be excluded if there is no clear reason for doing so. In the final analysis, if a particular observation appears to have a major influence on the overall pattern, then the analysis can be repeated without this observation to see whether it results in any change in the pattern. If any outliers are excluded then this exclusion must be mentioned in the final report. The criteria for rejection of observations should be clearly stated in order to convince the reader that they are not biased in favour of the hypothesis being tested.

The following figure shows average trends for weight gain and changes in PCV, excluding the 2 control and one ARF1 animal described above. It is important to appreciate that once an animal leaves the experiment it no longer features in the remaining part of the curve. In particular by day 139 only one control and 3 ARF1 animals remained. Therefore, caution is needed in the interpretation of such curves. The apparent increase in body weight from about day 115 in the ARF1 immunised animals, for instance, was due to the loss of two animals with lower than average body weight shortly before this day (see Fig. 4.8). Nevertheless, averaging the data in this way is a useful approach for a first examination of the different trends.



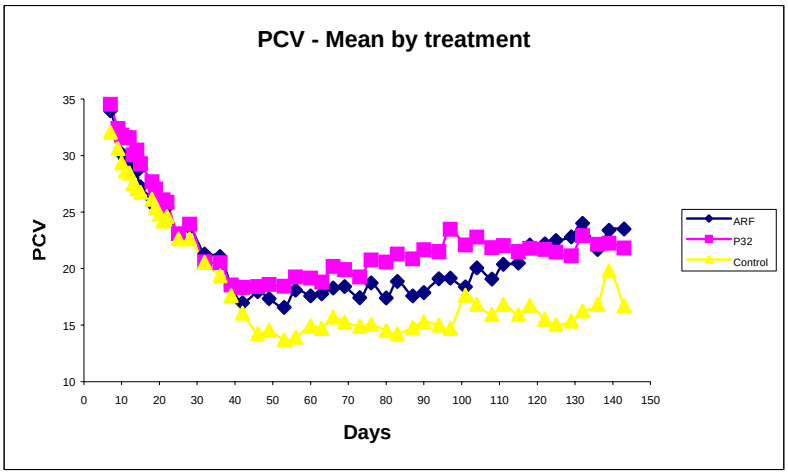


Fig. 4.10 Changes in mean body weight and PCV in the 3 treatment groups excluding 2 control animals and one animal from the ARF1 group of Example B.

Exercise 4.2 Study the curves for body weight and PCV in Figs. 4.7 – 4.9. Define one variable that can be derived from the weight curves and one variable from the PCV curves that you feel best define differences in pattern between groups and that might be used for further statistical analysis.

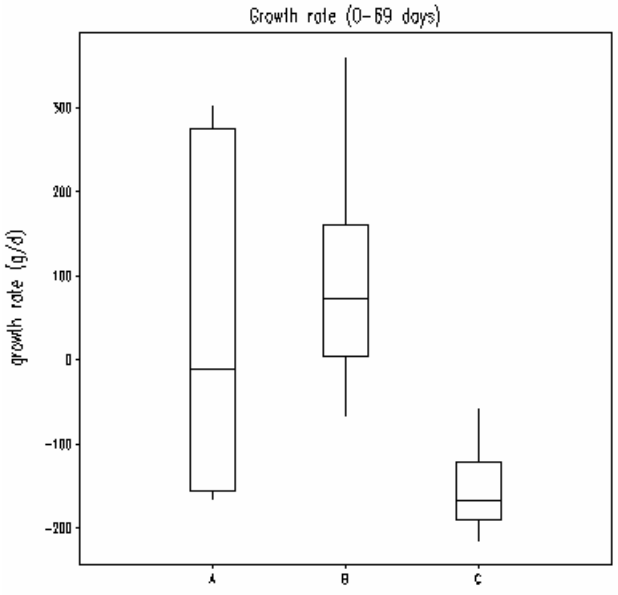
Analysis of longitudinal data such as in Example B can be complex. It is often a good idea to see how one can simplify the patterns. Thus, one can represent trends by a slope parameter, describe extreme values (maxima or minima) reached, the time when this occur, or summarise the patterns by calculating means over particular time periods of interest. Such approaches lend themselves to much simpler forms of statistical analysis which are often easier to interpret. Indeed by doing so we reduce the data structure from two layers representing animal and day within animal to just one representing animal only. This is because we have derived variables that represent summaries of data calculated over periods of time.

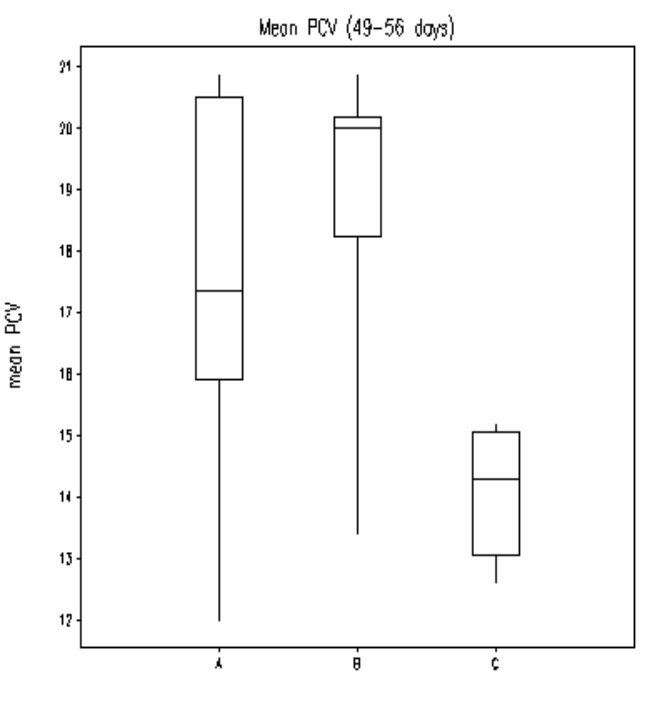
It was decided to represent the data in Experiment B by slopes representing the average growth rates to 69 and 139 days post infection and the mean PCV between 49 and 56 days when minimum values in PCV were generally reached. The data file used to analyse these data and containing these variables is shown in Table 4.4. Sixty nine days were chosen because the majority of cattle were still retained in the experiment on that day (see Table 1.2). Growth rates to 139 days used all available data for each animal up to day 139, or earlier if treated. Growth rates were calculated by linear regression – this method is described in the next chapter.

Table 4.4. Summary variables for growth rates to days 69 and 139 and mean PCV between days 49 and 56 derived from data for Example B shown in Tables 1.2 and 1.3

Record	Imm. group	Sex	Date of birth	Block	Initial wt. (kg)	Weight day 0 (kg)	PCV day 0 (kg)	Growth rate to day 69 (g/d)	Growth rate to day 69 (g/d)	Mean PCV day 49-56 (%)
1	A	F	04/06	1	142	180	39.2	64	-18	18.6
2	A	M	18/07	2	112	148	33.6	304	191	20.5
3	A	M	01/07	2	120	150	32.5	-165	-331	16.1
4	A	F	04/07	2	122	156	30.7	275	236	20.9
5	A	M	08/07	2	132	160	36.5	-88	-196	15.9
7	A	F	24/06	1	108	147	29.8	-156	-156	12.0
8	B	F	05/07	2	114	139	31.8	-66	-66	13.4
9	B	F	01/06	1	160	198	36.8	1	10	18.7
10	B	M	05/07	2	136	172	34.9	118	-39	20.0
11	B	F	15/07	2	100	118	30.0	361	271	20.9
12	B	F	02/06	1	140	200	36.7	72	-78	18.1
13	B	F	04/07	2	136	210	36.2	11	97	20.1
14	B	M	20/06	1	112	130	33.7	175	92	20.2
15	C	M	28/06	2	122	152	30.6	-143	-212	13.2
16	C	F	06/06	1	*	144	28.9	-58	-275	15.2
18	C	M	10/06	1	124	120	25.3	-168	-198	14.3
20	C	F	02/07	2	114	138	35.6	-214	-166	15.0
21	C	M	01/07	2	126	164	33.3	-183	-183	12.6

Exercise 4.3 The following two diagrams give the box plots for the summary variables growth rate to 69 days and mean PCV between 49 and 56 days. Comment on what they show.





Variance and standard deviation

Before proceeding to the next chapter let us see how we define variation in a statistical way. The most common measure of variability is the **variance** or **standard deviation.** The variance of a group of observations is calculated as the sum of squares of deviations from the mean divided by one fewer than the number of observations.

Thus, the variance of the 6-month body weights for Dorper (breed 1) without supplementation (code 1) in Example data set A is calculated as:

$$[(8.9 - 14.075)^2 + (10.1 - 14.075)^2 + ... + (17.7-14.075)^2] /7$$

$$= [(-5.175)^2 + (-3.975)^2 + ... + (3.635)^2]/7 = 82.755/7 = 11.8221 \text{ kg}^2$$

An alternative formula, which is easier to use with a large number of observations, is to calculate the sum of each value squared, subtract the total squared divided by the number of observations, and divide the answer by the number of observations less one.

Using this method the variance becomes

$$\{8.9^2 + 10.1^2 + ... + 17.7^2 - (112.6^2/8)\}/7$$

$$= (1667.6 - 1584.845)/7 = 82.755/7 = 11.8221 \text{ kg}^2 \text{ and the answer is the same.}$$

While it is often convenient to use the variance based on squared deviations as a measure of dispersion or variation of the data, it is often usual also to think of this variation in terms of the original units. By taking the square root of the variance one returns to the original scale of measurement. The square root of the variance is known as the **standard deviation (SD).**

Thus, the standard deviation = $\sqrt{11.8221}$ = 3.438 kg

So what does this mean? The standard deviation gives some measure of the spread of the data. Indeed, if we calculate mean ± 2 x standard deviations we obtain what is known as a confidence interval which includes approximately 95% of the data.

Degrees of freedom

The denominator in the formula for the variance (the number of observations minus 1, or n-1) is known as the **degrees of freedom**. The degrees of freedom represent the independent freedom with which observations can be used in the formula for the variance. The observations (which we shall write y_i (i = 1, ….., n)) are first used to calculate the mean, m. The mean is then used in the calculation of the sum of squares of deviations from the mean, namely sum $(y_i - m)^2$. Observations can be used independently to provide the first n-1 deviations squared, but the n^{th} observation does not provide new independent information since it can be derived from a knowledge of y_1, y_2, ….., y_{n-1} and m = $(y_1 + y_2 + ….. + y_n / n)$. This results in the loss of one degree of freedom. In other words a variance or standard deviation has n-1 degrees of freedom.

Coefficient of variation

We have shown that the variation among a set of observations can be measured by the variance or the standard deviation. There are occasions when we wish to compare the relative amounts of variation between two variables having different means, e.g. birth weights of calves and adult weights of cows. Since the adult weights would be expected to have a larger standard deviation because of the larger values involved, a comparison of the two standard deviations may not very helpful. We might also be interested in comparing the relative amounts of variation between different characteristics or traits of the same animals, e.g. weaning weights and packed cell volumes of lambs. The **coefficient of variation** (**CV**), which is calculated as the standard deviation expressed as a percentage of the mean, is useful for such comparison.

Thus, for unsupplement Dorpers the coefficient of variation for 6-month body weight

= (3.438/14.075) x 100 = 24.4%

The CV is independent of the unit of measurement. For body weight or packed cell volume, for example, coefficients of variation of the order of 10 - 15% depending on the type of experiment / field study might often be expected. Growth rates, however, often have higher coefficients of variation, also variables such as FEC that are skew. This example shows that the spread of weights and PCVs in this group of lambs is wider than one would normally expect. But remember that this ignores the effects of breed,

supplementation, etc. Once these are taken into account in the calculation of the residual variation, the coefficient of variation may be reduced.

Table 4.2 summarises the descriptive statistics for the 32 animals in the Example A data set. Note the high CVs for all variables, particularly for weight gain and FEC. In a real situation the comparatively large CVs for body weight and PCV may have been associated with exposure of these lambs to high levels of helminth infestation. However, these data have been artificially created and have probably been made more variable than would be the case in real life. The medians represent the mid points of the distributions, i.e. the value in the middle.

Table 4.2 Descriptive statistics of variables measured in Example A (number of animals = 32)

Variable	Mean	Median	Standard deviation	Coefficient of variation
3-month weight (kg)	11.69	11.60	2. 71	23.2
6-month weight (kg)	13.61	13.15	2.92	21.49
Weight gain (kg)	2.17	1.95	1.08	49.7
PCV (%)	22.72	23.50	5.35	23.6
FEC (epg)	1770	1200	1686	95.3

Exercise 4.4. Calculate the mean, median, standard deviation and coefficient of variation of the PCVs of the 8 lambs in breed 1 that were not supplemented.

Standard error

Suppose the experiment were repeated several times. Each time we would calculate a mean, say for breed 1, without supplementation. Each mean would be different. Thus there would be variability among means just as there is variability among individual observations in each sample. However, since each mean is averaged over 8 observations, the variation among means would be smaller than the variation among individual observations. The variance of a mean can be calculated as the average of the variances calculated each time the experiment is conducted, divided by the number of observations used for calculating each mean. The standard deviation of the mean is the square root of the variance of the mean. The standard deviation of the mean is known as the **standard error.**

The standard deviation is a useful measure of the variation of an individual observation. The standard error is a useful measure of the variation of a mean.

Normally we would not repeat an experiment to calculate the average variation of the mean. Instead, we pretend that the variance of individual observations will remain constant from experiment to experiment (in practice it won't vary much) and use the

value already calculated as an estimate of this average. Thus, we can then calculate, for the 6-month body weights of non-supplemented Dorper lambs:

variance of mean = 11.8221/8 = 1.4778 kg^2

standard error = $\sqrt{1.4778}$ = 1.22 kg

The standard error or variance of a mean decreases as the number of observations increases. For example, if 24 lambs were used for each breed then, for breed 1, non-supplemented,

variance of mean = 11.8212/12 = 0.9926 kg^2
standard error = 1.00 kg

Note that this standard error is lower than the value calculated above with 8 sheep.

We can calculate the standard error for other statistics such as the slope of the relationship between two variables. The formula is given in the next section, but the principle is the same – the standard error represents the standard deviation of a mean, of a slope, or of whatever other statistic is used.

Confidence interval

The standard error can be used to develop what is known as a **confidence interval.** Thus, for the mean**,** a confidence interval is a range between upper and lower limits, which is expected to include at a given level of probability the true (or population) mean value. This is the value for which the sample in the experiment is providing an unbiased estimate.

Usually we talk about the 95% confidence interval. This is the interval in which the true mean should lie with a 95%, or 19 times in 20, chance of being correct. Similarly, the 99% confidence interval gives the range within which we expect the true mean to lie with a 99%, or 99 times in 100, chance of being correct.

The approximate 95% confidence interval can be calculated as the sample mean plus or minus twice the standard error. Similarly, the 99% confidence interval is the sample mean plus or minus approximately 2.6 times the standard error.

Thus the approximate 95% confidence interval for the mean 6-month body weight of non-supplemented Dorper lambs is:

$14.08 \pm 2 \times 1.22$

= (11.64 to 16.52) kg

Similarly, the approximate 99% confidence interval is

$14.08 \pm 2.6 \times 1.22$

= (10.91 to 17.25) kg

As one might expect, the 99% confidence interval is wider than the 95% one. Another way of thinking about this is to say that if the experiment were repeated 100 times then we would expect that on 99 occasions the sample mean would fall within the range 10.91 to 17.25 kg, and on the other occasion it would fall outside.

In summary, therefore, given that we know how to calculate the standard error of a statistic (e.g. mean, or slope) determined from the experiment, we are able to determine a likely range within which the true (or population) value for the statistic lies.

Exercise 4.5 Using the results of Exercise 4.4, calculate the standard error and 95% confidence interval for mean PCV of non-supplemented lambs of breed 1.

Binomial and Poisson variables

Sometimes response variables take on the values 0 or 1. These could be, for example, mortality or disease. The variables are sometimes described as **discrete** variables, i.e. they can only take on discrete and not continuous values. Variables such as body weight are known as **continuous** variables, i.e. they can take on any value within a reasonable range to a given degree of accuracy. Discrete variables such as mortality and disease often belong to distributions that are not normal distributions, but **binomial** or **Poisson** distributions. Discrete variables are also sometimes called **binary** variables. For example, suppose in a sample of 20 lambs, 4 die before weaning. We can record a death as 1, a lamb that survived as 0. The proportion of lambs that died is $p = 4/20 = 0.2$. This is simply the mean of the 0s and 1s, i.e. $(1+1+1+1+0+.... +0)/20 = 4/20 = 0.2$. Mortality is typical of a variable that is associated with a probability of occurrence p. Such variables typically belong to a binomial distribution.

We have seen that the mean = p, and that this can be calculated in an analogous way to that of a normally distributed variable. The binomial distribution, however, has a special expression for the variance, namely p(1-p). Thus if $p = 0.2$, then variance $= 0.2 \times 0.8 = 0.16$.

The standard deviation and standard error are similarly calculated as for a continuous variable, i.e.

$$\text{standard deviation} = \sqrt{0.16} = 0.4$$
$$\text{standard error} = \sqrt{0.16/20} = 0.089$$

The distribution of certain binary data is often closer to that of a Poisson distribution, a distribution sometimes associated with rare events. Incidence of trypanosomosis, for example, can often be considered as being associated with a Poisson distribution, particularly if the prevalence is very low (e.g. 5%). The mean is the same as that for the binomial, namely p. The variance, however, is different from that of the binomial distribution and is also p. Thus, if the prevalence of trypanosome infections is 0.06 in a sample of 100 cattle, (i.e. 6/100 are detected parasitaemic), and a Poisson distribution is assumed then the variance is also 0.06 and

$$\text{standard deviation} = \sqrt{0.06} = 0.24$$

standard error $= \sqrt{0.6/100} = 0.024$

In practice when p is small (<0.1), it does not matter whether one assumes a binomial or Poisson distribution.

A Poisson distribution is often used for analysing count data, e.g. ticks on an animal. More animals will be found with few ticks than with many ticks. Thus, the frequency distribution will be skew. The distribution often follows a Poisson distribution.

Exercise 4.6 In a sample of 80 sheep 5 are detected with trypanosomes. Calculate the standard error of the prevalence assuming first a Poisson distribution and second a binomial distribution. Calculate the 95% confidence interval using the formula given for the normal distribution.

Chapter 5 Simple statistical models and analysis of variance

So far in this course we have seen exploratory data analysis to be a useful method for obtaining a preliminary idea of the primary form of the patterns in the data and how we might go about analysing the data. With exploratory investigations, however, we are not always able to look at the effects of all different sources of variation simultaneously. This is where statistical modelling can play a role. The underlying philosophy is still one of separating data into

pattern + residual

and quantifying and describing both.

We shall first use two simple examples :

a) a simple linear regression
b) a balanced factorial analysis of variance

Simple linear regression

Let us consider the association between PCV and FEC in Example A and ignore, for the time being, that lambs were of different breeds and were fed on different diets. From the scatter plot given in the previous chapter we can see that there is a strong association between PCV and FEC. In terms of the idea that

data = pattern + residual

our pattern can be described by a linear equation i.e. an equation of the form

PCV = a + b * FEC

where a is the intercept on the PCV (y) axis and b is the slope or gradient of the line. We often call this slope the regression coefficient of PCV (y) on FEC (x). Our model for the data is therefore :

PCV = a + b * FEC + residual

PCV (y) is often referred to as the dependent or response variable and FEC (x) as the independent or explanatory variable.

The following figure shows the regression line superimposed on the scatter of points.

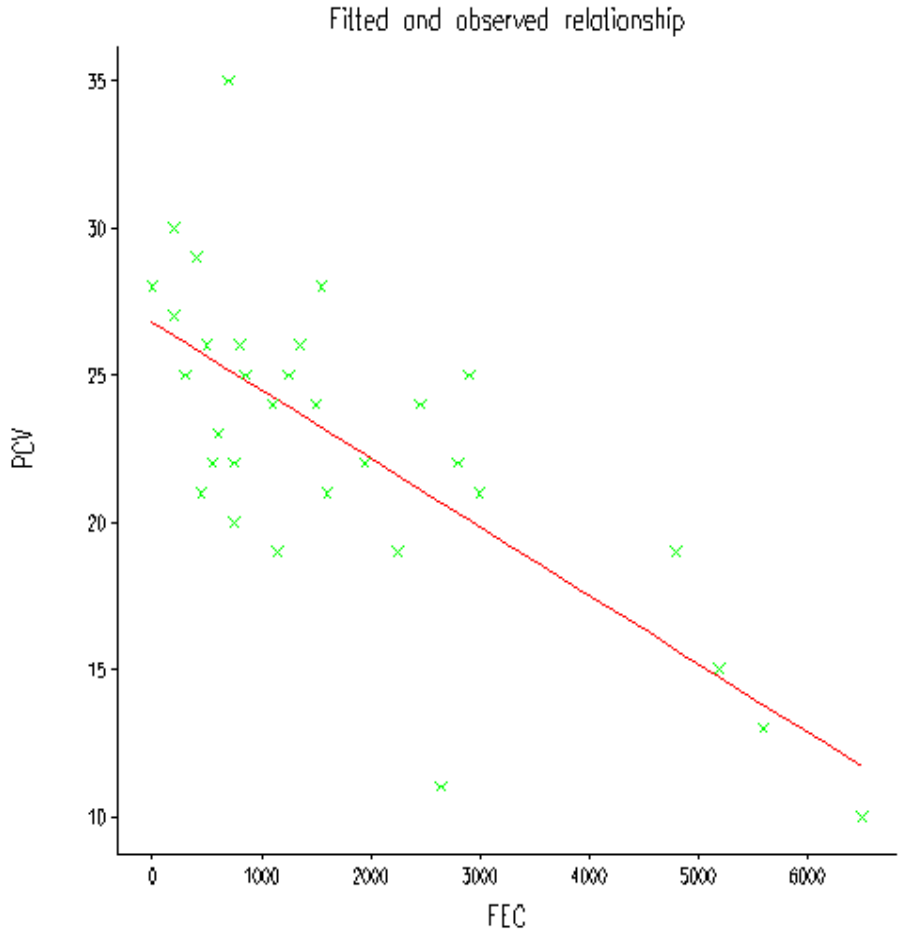


Fig. 5.1 Regression line of PCV on FEC for the 32 lambs in Experiment A. The slope is designated by 'b' and the point where the line intercepts the y-axis by 'a' in the equation $y = a + bx$.

The calculations of a and b are fairly straight forward and it is helpful to understand how these are done in the simple case. Sometimes one may be working in the field, without immediate access to a computer, and so it is useful to know how to do simple calculations on a calculator.

If we write y_i to refer to the ith value of the dependent variable PCV such that i ranges from 1 to 32, and write x_i to refer to the ith value of the independent variable FEC, then we can produce an analysis of variance table to determine the proportion of the total variation in y that can be accounted for by x. Analysis of variance is a descriptive tool in statistical analysis. Its purpose is to describe the amounts of different sources of variation attributable to different attributes and to compare them to the residual, uncontrollable variation. The beauty of this method is that it can look

at different effects together. We shall be considering different forms of analysis of variance later. However, for the time being, consider that analysis of regression is just one form of analysis of variance.

An analysis of variance table comprises three columns : degrees of freedom (df), sum of squares (SS) and mean squares (MS), which equals SS/df. The residual MS is the same as the residual variance. The variance ratio (VR), as we shall see, is the ratio of the regression MS to the residual MS.

Source of variation	df	SS	MS	VR
Regression	1			
Residual	30			
Total	31			

The degrees of freedom for the total are calculated as (32-1), i.e. the total number of observations minus 1. This is precisely as is done for the calculation of a variance. This takes care of the intercept a in the equation. One further degree of freedom is used for fitting the slope b, leaving 30 degrees of freedom for the residual term.

Total corrected sum of squares for PCV = $\sum y_i^2 - (\sum y_i)^2 /32$

$$= 17405 - 727^2 /32$$

$$= 888.4688$$

The above formula is precisely that used for the numerator in the calculation of a variance in the previous chapter.

Let us write this as S_{yy} where S_{yy} signifies the total corrected sums of squares of y ignoring any pattern displayed by the regression line. Thus,

$$S_{yy} = \sum y_i^2 - (\sum y_i)^2 / 32$$

Similarly, total sum of squares for FEC can be written as:

$$S_{xx} = \sum x_i^2 - (\sum x_i)^2 /32$$

$$= 188,382,500 - 56,650^2 /32$$

$$= 88,094,297$$

We also need the total sum of cross products $= \sum x_i\, y_i - (\sum x_i \sum y_i) /32$ which we shall write as S_{xy}.

$$\text{Thus, } S_{xy} = 1,082,200 - (56,650 \times 727) /32$$

$$= - 204,817$$

The analysis of variance table for this regression analysis can then be written as shown below.

The calculation of the sums of squares (SS) for the regression line are as shown

	df	SS	MS
Regression	1	$(S_{xy})^2/S_{xx}$	
Residual	30	$S_{yy} - (S_{xy})^2/S_{xx}$	
Total	31	S_{yy}	

where the expression $(S_{xy})^2/S_x = 204{,}817^2 / 88{,}094{,}297$ represents the sum of squares explained by the regression of PCV on FEC.

The MS column is then calculated as SS/df. Thus it can be seen that the total MS is simply the variance of PCV (i.e. ignoring any fit of a regression line to FEC). The residual MS is an expression for the residual variance of the PCVs left over after fitting the regression line. To understand this further imagine vertical lines drawn from each point in Fig. 5.1 as far as the regression line. The sum of squares of these deviations from the line gives the residual sum of squares. Thus, again, the residual variance represents that variation remaining in the data after the pattern described by the regression line is fitted.

In addition,

Regression coefficient (b) = S_{xy} / S_{xx}

Intercept (a) = $\bar{y} - b\bar{x}$

Regression equation becomes $y = \bar{y} - b\bar{x} + bx$. This is the pattern that we are attempting to describe.

SE of b = $\sqrt{}$ (residual MS/S_{xx})

Correlation coefficient = $\sqrt{[(S_{xy})^2 /(S_{xx}\, S_{yy})]}$

The SE measures the precision with which we can define b. We can use this standard error, just in the same way as has been shown for a mean, for determining a confidence interval for b.

The correlation coefficient ($0 \le r \le 1$) is a measure of association between y and x. If r = 0 there is no association i.e. no pattern; if r = 1 the association is perfect, i.e. all points lie on a straight line. The correlation coefficient is particularly useful for determining the degree of association between two variables x and y when neither is considered as the dependent variable.

The above formulae lead to the following analysis of variance. Note again that the MS is calculated as SS / df

Source of variation	df	SS	MS
Regression	1	476.1952	476.1952
Residual	30	412.2736	13.7425
Total	31	888.4688	28.6603

Regression coefficient (b) = -0.00232. This is interpreted as meaning that PCV decreases by 0.00232 percentage units for one egg per gram increase in FEC, or alternatively as 0.232 percentage units per 100 eggs per gram.

Intercept (a) = 26.8%.

Thus, regression equation is $y = 26.8 – 0.00232x$

Exercise 5.1 Calculate the SE of b in the above example and the correlation coefficient (r or ρ). What you can say about the estimated population regression coefficient given the size of the SE?

Exercise 5.2 Fit the above linear regression using Genstat and study the output. Certain observations are identified in the output as either having influence on the slope of the line or having large residuals. See if you can identify these in Fig. 5.1 of these notes. Fit the line again without the influential points. What effect has this on the slope?

So what can we deduce from all of this? First of all the regression equation has accounted for over half the sums of squares for PCV. Furthermore, the variance has been reduced by over half (from 28.6603 to 13.7425% units^2) when the pattern in our data has been taken into account. The pattern is described by line intersecting the y-axis at a PCV of 26.8% and decreasing by 0.232 % units of PCV for an increase of 100 e.p.g. in FEC. If we multiply by the SE of b by 2 we can see that 95% confidence limits for the true slope are $–0.00232 \pm 2 \times 0.00039$ = (-0.00310 to -0.00154% units / e.p.g.). The correlation coefficient shows that there is a fairly good correlation between PCV and FEC.

Analysis of variance for designed experiments

Let us now look at the form the analysis of variance takes for designed experiments. Before doing so let us see how we can express our statistical model algebraically. Statistical concepts are often presented mathematically. It is important for biologists not to be frightened by the formulae that sometimes appear in scientific papers and we shall attempt to remove the mystike. Formulae are often the easiest way to describe the

statistical methods that have been applied. As we have already seen there is little difficulty in writing down the formula for a regression line.

If we look again at the structure of our example data set we can see that each variable, 3-month body weight, 6-month body weight, PCV, FEC and weight gain, can be characterised by other attributes. For example, records 1-8 all refer to lambs from breed 1. They are also characterised by the fact that these lambs were not supplemented. Thus, we can say that each record is made up of a component due to breed and a component due to diet. We shall use a letter to refer to each of these components, namely b (breed), d (diet). These components are often referred to as **parameters** or **effects**. It is common to give a subscript to each parameter level. Thus, we have 2 breeds. Let us refer to them as b_1, b_2. Similarly, we have 2 diets, d_1, d_2. Thus, for example, records 1-8 are each made up of $b_1 + d_1$. Similarly, records 9 - 16 are made up of $b_1 + d_2$ and records 25 - 32 of $b_2 + d_2$. It is usual, just as for a regression equation, to use the letter y to refer to the variable to be analysed and to use a combination of subscripts to refer to the particular data value. Thus,

$y_{111} = b_1 + d_1$ for record 1

$y_{124} = b_1 + d_2$ for record 12

$y_{225} = b_2 + d_2$ for record 29

Note that the first subscript refers to breed, the second to diet and third to lamb number (1-8) within each breed-diet category.
Normally, b and d are defined as deviations from the overall mean, which is usually written μ. Also, since there are fewer parameter levels (2 + 2) than there are lambs (32), each observation will deviate from the sum of the 2 parameter levels by a residual amount usually described by the letter e or ε, known as the **error** or **residual** term.

Thus, the above equations are written in full as

$y_{111} = \mu + b_1 + d_1 + e_{111}$

$y_{124} = \mu + b_1 + d_2 + e_{124}$

$y_{225} = \mu + b_2 + d_2 + e_{225}$

Thus, each observation is made up of an overall mean, a term for breed, a term for diet and a residual. These equations can be written more generally as

$$y_{ijk} = \mu + b_i + d_j + e_{ijk} \quad (i = 1, 2; j = 1,2; k = 1....8)$$

This, then is an algebraic formulation of the expression

Data (y) = pattern (μ + b + d) + residual (e)

As has already been mentioned, the purpose of **analysis of variance** is to separate and quantify the different sources of variation. In this case it separates variations due to

breed and variations due to diet from each other, and compares the magnitudes of these different sources of variation with the variation which is left over among residual terms. In using analysis of variance two main assumptions are used: firstly, that the e_{ijk}s have the same distribution for each parameter (i.e. are not more variable for one than another) and, secondly, they are distributed normally. Thus, it is not strictly valid, unless the data set is very large, to use analysis of variance for discrete variables that are associated with binomial or Poisson distributions. Logistic or log-linear models are ones that can be used for these types of distributions. Refereed journals insist on the use of such methods for discrete variables. These models will not be covered in this course.

In developing further the idea of analysis of variance it will be simplest to start with a model based on just the 4 groups of lambs categorised by breed and diet in the data set, but ignoring the breed-diet structure. The model we shall first use, with g signifying group, is

$$y_{ij} = \mu + g_i + e_{ij} \quad (i = 1...4; j = 1....8)$$

It is important to emphasise the need for writing down the statistical model to be fitted to a set of data before embarking on a statistical analysis. It is a great help in representing the pattern in our data and deciding the form the analysis of variance should take.

One way analysis of variance

The model

$$y_{ij} = \mu + g_i + e_{ij}$$

is an example of a **one way analysis of variance.** Such an experiment is often referred to as a completely randomised design. The analysis of variance will take the following form:

Source of variation	df	SS	MS	VR
Among groups	3			
Residual	28			
Total	31			

The **degrees of freedom** (**df**) for the variation among groups is calculated as one fewer than the number of groups (i.e. 4-1). The total degrees of freedom = 32 (number of lambs) -1. The residual degrees of freedom is calculated by subtraction of the among groups df from the total df.

The following calculations calculate the **sums of squares** (**SS**) and **mean squares** (**MS**).

Total. This is the same calculation as that used for the variance, i.e. the sum of squares of individual observations minus the square of the total divided by the number of observations. This is precisely the same as we did for regression analysis. Mathematically we can write this as

$$\Sigma y_{ij}^2 - (\Sigma y_{ij})^2/32 \quad (i = 1, \ldots, 4 ; j = 1, \ldots, 8)$$

We now have two subscripts because we are summing both within and across groups. The latter term $(\Sigma y_{ij}^2)/32$ is often referred to as the **correction factor**.

Thus, if we apply this formula to the weight gains of the 32 lambs, we obtain

$$\text{total SS} = 186.48 - 69.4^2/32$$

$$= 186.48 - 150.5112 = 35.9687$$

Groups The formula for the among group sums of squares uses the totals for each group (which we can write $G_i = y_{11} + y_{12} + ... + y_{18}$), squares them and then divides by the number of lambs in the group. It then substracts the correction factor given above to give:

$$(G_1^2 / 8 + G_2^2 / 8 + G_3^2 / 8 + G_4^2 / 8) - (\Sigma y_{ij})^2 / 32 \qquad (i = 1, \ldots, 4)$$

The among group SS for weight gain is therefore

$$(10.6^2 + 19.1^2 + 11.9^2 + 27.8^2)/8 - 69.4^2/32$$

$$= 173.9525 - 150.5112 = 23.4413$$

The among group mean square equals the among group sum of squares divided by the among group degrees of freedom.

Thus, among group MS $= 23.4413/3$

$$= 7.8138$$

Residual We calculate the residual sum of squares by subtracting the among group sum of squares from the total sum of squares.

Thus, residual SS = 35.9687 - 23.4413 = 12.5274

Residual MS = residual SS/residual df

$$= 12.5274/28 = 0.4474$$

Putting the results of these calculations into the analysis of variance table we get

Source of variation	df	SS	MS	VR
Among groups	3	23.4413	7.8138	17.5
Residual	28	12.5274	0.4474	
Total	31	35.9687	1.1603	

The residual MS estimates the average variance among individuals within groups and is thus an estimate of the average within group variance of weight gains. The among group MS represents the additional variation brought about by the differences in mean weight gains among the 4 groups. We can see that it is 17 times higher than the residual MS, which suggests that there is considerable variation in average weight gain among the four groups. This is as we suspected in our exploratory analysis which we undertook earlier. Indeed the residual variance has been reduced from 1.1603 to 0.4474 kg^2 by accounting for this pattern in our data.

Exercise 5.3 Complete using your calculator a one-way analysis of variance for growth rate from 0 to 69 days to compare differences between the 3 immunisation groups in Experiment B. Comment on what it shows.

Exercise 5.4 Run this one-way analysis of variance for both growth rate from 0-69 days and mean PCV between 49 and 56 days using Genstat. Click 'options' and remove ticks for F-probabilities and Standard Errors of 'Differences ; click Standard Errors of 'Means' instead). Study output. Is there anything in the output that you do not understand?

Factorial analysis of variance

The groups of lambs in Experiment A have been chosen with a particular structure. In other words the 4 groups are composed of combinations of 2 diets and 2 breeds, namely

	Supplementation	
	yes	no
Breed		
1	1	2
2	3	4

This design, sometimes described as a 2x2 factorial design, is a very useful way, as already discussed in the Introduction, of getting as much information from an experiment as possible. Thus, we can look at the effects of dietary supplementation on weight gain of two breeds at the same time. What's more, we can look at the way that dietary supplementation and breed may interact. If there is no **interaction** then we conclude that any effect of dietary supplementation is the same for each breed. If there is an interaction then the effect of supplementation is different for different breeds.

We can rewrite our statistical model as follows:

$$y_{ijk} = \mu + b_i + d_j + (bd)_{ij} + e_{ijk} \qquad (i = 1, 2 ;\ j = 1,2, k = 1, \ldots., 8)$$

where $(bd)_{ij}$ is a notation used to signify the interaction of breed and diet.

If the analysis shows there to be no significant interaction than we could express the model as

$$y_{ijk} = \mu + b_i + d_j + e_{ijk}$$

In other words, the model simply adds to breed the same estimates for diet irrespective of breed. Sometimes we refer to such effects as being **additive**.

In order to estimate sums of squares for breed and diet we split the sums of squares we have already obtained for groups, namely

$$\Sigma G_i^2/8 - (\Sigma y_{ijk})^2/32 \quad (i = 1,2 ; j = 1,2 ; k = 1,8)$$

We now calculate a component for breed

$$\Sigma B_i^2/16 - (\Sigma y_{ijk})^2/32 \quad (i = 1, 2)$$

where B_i is the sum of the ys for breed i,

and a component for diet

$$\Sigma D_j^2/16 - (\Sigma y_{ijk})^2/32 \; j = 1,2)$$

where D_j is the sum of the ys for diet j.

The expression we had earlier for the among group SS was:

$$(10.6^2 + 19.1^2 + 11.9^2 + 27.8^2)/8 - 69.4^2/32$$

The between breed SS becomes:

$$(29.7^2 + 39.7^2)/16 - 69.4^2/32$$

The between diet SS becomes:

$$(22.5^2 + 46.9^2)/16 - 69.4^2/32$$

This results in sums of squares of 23.4413, 3.1250 and 18.6050 for group, breed and supplementation, respectively. The sum of squares for the interaction can be obtained by subtraction of the last two numbers from the first, namely 23.4413 - 3.1250 - 18.6050 = 1.7113.
We can then complete the analysis of variance table as follows:

Source of variation	df	SS	MS	VR
Breed (B)	1	3.1250	3.1250	6.98
Supplementation (S)	1	18.6050	18.6050	41.6
Interaction(BxS)	1	1.7113	1.7113	3.82
Residual	28	12.5274	0.4474	
Total	31	35.9687		

This analysis now partitions the group variation described earlier into components associated with breed and supplementation. We can see that the largest component of the total variation is associated with supplementation and that this is about six times that between breeds. The interaction mean square is approximately half that between breeds. Our exploratory analysis indicated that supplementation might have had a greater impact on weight gain in Red Maasai than Dorpers. This analysis seems to confirm this, but it also shows that any interaction effect is small. Is it significant? In the next chapter we discuss how we can infer whether or not different terms included in our model are statistically significant or not.

Exercise 5.5
Complete using your calculator the following analysis of variance table for PCV in Experiment A.

Source of variation	*df*	*SS*	*MS*	*VR*
Breed	*1*			
Diet	*1*			
BxD	*1*			
Residual	*28*	*560.3750*		
Total	*31*	*888.4688*		

Exercise 5.6 Repeat the above analysis of variance for PCV in Genstat. (Click 'options' and change values as in Exercise 5.4). Compare the output with your calculated results.

We can use the residual MS to calculate standard errors for the different means. The residual MS is simply the average residual variance. We saw in the last chapter that the standard error of a mean is the square root of the variance divided by the number of observations used to calculate the mean.

Mean values for weight gain (kg) are as follows:

	Supplementation		
Breed	No	Yes	Mean
Dorper	1.32	2.39	1.86
Red Maasai	1.49	3.48	2.48
Mean	1.41	2.93	2.17

From the analysis of variance the residual variance = 0.4474 kg^2. Thus the standard error (SE) of any of the 4 means in the body of the table is $\sqrt{(0.4474/8)} = 0.236$. Similarly, SEs for overall means for breed or supplementation (yes or no) are $\sqrt{(0.4474/16)} = 0.167$. If we multiply these SEs by 2 we can use them to determine approximate 95% confidence interval within which the true means lie. Comparisons of these ranges will illustrate the degree of separation between the various means.

Exercise 5.7 Calculate 95% confidence intervals for the 4 means in the body of the table of means for weight gain in Experiment A (Dorper/Red Maasai experiment).

Blocking

If we recall the designs of Experiments A and B we might remember that in both cases some form of blocking was carried out. In experiment A lambs within breed were put into 4 blocks ranked on the basis of body weight before randomly assigning them to treatment. Cattle in Experiment B were separated into 2 age groups before assigning them to treatments. So far we have not considered blocks in the analysis of variance, and we must.

Recall that, without blocks, the model that we used for Experiment A was

$$y_{ijk} = \mu + b_i + d_j + (bd)_{ij} + e_{ijk}$$

where i = 1,2 ; j = 1,2 and k = 1,2 ….., 8.

We now have to incorporate block (for which we shall use the letter r standing for replicate) in the model. This a little tricky for the blocking was done within breed. For each breed there are 4 blocks. We say that blocks are **nested** within breeds. To do this we use a double suffix i and m where m = 1,2,3,4. Thus, r_{im} describes block m within breed i. The complete formula is now.

$$y_{ijkm} = \mu + b_i + r_{im} + d_j + (bd)_{ij} + e_{ijkm}$$

In practice we would reorder the suffices and rewrite the model

$$y_{ijkm} = \mu + b_I + r_{ij} + d_k + (bd)_{ik} + e_{ijkm}$$

where i = 1,2 ; j = 1, …, 4 ; k = 1,2 and m = 1,2 for the two lambs for each diet within each block

Exercise 5.8
Do an analysis of variance for weight gain for the above model in Genstat. Comment on the effect of blocking. Would you block again on body weight in a follow-up experiment? (Normally one would use the Block structure in Genstat to put in a term involving blocks. In this example however, it causes a complication in the output. To keep the output simple include the term for block within breed as BREED.BLOCK in the Treatment structure.)

Exercise 5.9
Write the statistical model for Experiment B including parameters for both immunisation and block. Do an analysis of variance for weight gain in Genstat..

Analysis of covariance
Analysis of covariance is a simple extension of analysis of variance, combined with analysis of regression. Independent variables, not incorporated in the design of the experiment, are included as **covariates**. Thus, in the study of the effect of

supplementation on the weight gains of the two breeds of sheep, weight at 3 months was taken into account in the definition of blocks. Sex, however, was not. We can, however, include sex in the model as a covariate (1,2). Covariates may also be continuous variables. The full model appears as follows:

$$y_{ijkm} = \mu + b_i + d_j + r_k + (bd)_{ij} + \beta x_{ijkm} + e_{ijkm}$$

where β is the regression coefficient of weight gain (y_{ijkm}) on sex (x_{ijkm}).

The analysis of covariance for weight gain appears as follows:

Source of variation	df	SS	MS	VR
Breeds (B)	1	2.9919	2.9919	6.62
Blocks within breed	6	2.9192	0.4865	1.08
Supplementation (S)	1	18.3697	18.3697	40.66
Breed x Supplementation	1	1.7629	1.7629	3.90
Sex	1	0.7870	0.7870	1.74
Residual	21	9.4868	0.4518	
Total	31	35.9687		

The analysis adjusts, or corrects, parameter level means for the differences in numbers of males and females making up each mean. The regression on sex, whilst only having a very small effect (VR = 1.74) has slightly reduced the residual mean square. Interestingly, when adjusted for sex, the F value for blocks is now also greater than 1. Note that the total sum of squares is no longer the sum of the individual squares in the analysis of variance table. This is because each sum of squares is now corrected for differences between sexes.

When there is a choice to use a factor as a covariate or to define it as a blocking factor in a designed experiment, the latter is usually to be preferred, provided adequate numbers of degrees of freedom are retained for the residual term. Had there have been equal numbers of females and males for each breed in this example, then it would have been possible to first block lambs on the basis of their sex within breed and then on the basis of body weight for each sex.

Exercise 5.10 Run the model for analysis of covariance on sex in Genstat and verify that the output matches that given in the notes. Compare mean values for breed and diet with those given when carrying out an analysis of variance without sex and see how they have been adjusted for sex.

So what does β, the regression coefficient for sex in the output, mean? Normally we would think of this as a regression coefficient of a dependent variable y on an independent variable x which is continuous. In this case the independent variable x is

sex, which is a discrete variable taking on the value 1 for male and 2 for female. We can still think of β as a regression coefficient representing the slope of the line joining the mean values of weight gain for the two sexes. The slope represents the difference in weight gain along the y-axis divided by the difference in sex (2 minus 1) along the x-axis. As $2 - 1 = 1$, β is simply the difference in weight gain between the two sexes.

The computer output produced by Genstat gives the regression coefficient (slope) for β to be 0.42 and the s.e. 0.32 kg. Thus, females grow faster by 0.42 ± 0.32 kg than males. This is an unlikely result and probably occurred by chance. Referring to Fig. 4.7 in chapter 4 we note that the box plot indicates a trend in the opposite direction. Note, however, the box plot shows the data grouped by one factor and ignores other factors such as breed and supplementation. Calculation of the sex effect in the analysis of covariance on the other hand corrects for all other factors in the model.

Chapter 6 Writing up and presenting results

The statistical analysis has no value when the results cannot be conveyed to the relevant audience. This can be difficult. The researcher knows just how complex and perhaps confusing the analysis has been, has discovered many nuances to the data, yet has to convey the results in as simple a way as possible that demonstrate how the objectives of the study have been met.

Text becomes difficult to read if too many numbers are included. Thus, a set of numerical results should usually be presented as a table or graph rather than included in the text. Well presented tables and graphs can concisely summarise information which would be difficult to describe in words alone. Poorly presented tables and graphs, however, can be confusing or irrelevant. In general, tables are better than graphs for giving detailed numeric information, whereas graphs are better for indicating trends and making broad comparisons.

Tables and graphs should, ideally, be self-explanatory. In other words, the reader should be able to understand them without undue reference to the text. The title should be informative and rows and columns of tables and axes of graphs should be clearly labelled. Graphs and tables should be as simple as possible while having sufficient detail to be useful and informative.

Statistical information, such as standard errors and significance levels (determination of significance levels will be described in the next chapter) is usually required in scientific papers. This may not be necessary for articles for a more general readership or for slides or transparencies for use in a lecture. Such statistical information should always be presented in such a way that it will not obscure the main message of the table or graph.

One often finds tables in scientific papers clustered with little superscripts a, b, ab, c, etc, or symbols *, *** to signify significant differences. Authors need to ask themselves whether these are really necessary, as they tend to obscure the main ingredients of the table. In some cases they will feel they are necessary and help with the understanding of the table. Authors should think twice, however, before doing this. Normally tables are much better if they are simple and present the data at their face value. Often only means and standard errors are required. Significance levels can be quoted in the text as a guide to the reader, but the reader should be able to draw his own conclusion on how 'significant' he/she feels the results are.

Exercise 6.1. Write, in no more than three sentences, a report of the results shown in the analyses of weight gain and PCV done so far for Example A (Dorper/Red Maasai). Do this without using any statistical significance tests.

Exercise 6.2. Do likewise for Experiment B (immunisation experiment).

The number of digits and decimal places presented in a table should be the minimum number that is compatible with the purpose of the table. Thus, three decimal places for weight gain in the lamb experiment, e.g. 1.325, may be too many, since the third

decimal place may have little practical significance; conversely, one decimal place e.g. 1.3, 2.4, could lose too much information. As a general rule, standard errors should be written with the same or one more decimal place than the mean.

Graphical presentation often helps to highlight certain aspects of the results. Individual standard errors for each diet can be presented as a vertical line above a bar. If it is necessary to include standard errors of difference between means, these can be included as a footnote or in the text. A graph should never duplicate the results displayed in a table in the same report.

Exercise 6.3. Prepare a table illustrating the mean results for weight gain and PCV in Example A assuming that it has been decided to ignore the interaction. Include measures of variation in this table.

Exercise 6.4. Prepare a similar table for Example B. Which graphs would you include in a paper to supplement this table?

The course so far has dealt only with estimation and not with hypothesis testing (see next chapter). As statistics is only a tool to help the researcher put over his results in a way that gives some reassurance to the reader, we ought to be able to write a convincing report without resorting to over-use of statistical significance levels. The term 'statistically significant' is indeed often over-used in statistical reports of experiments. Firstly, 'significant' has a common usage quite different from the statistical one. Some readers may get confused. Secondly there is no clear boundary between 'significant' and 'not significant' (those that are used, e.g. $P < 0.05$, $P < 0.01$, are artificial). Thus, using phrases such as 'clear evidence for ….', 'some evidence for ……' and 'little reason to believe that ……' could reflect the real state of the knowledge more fairly. Furthermore, to state whether a difference is statistically significant may not be sufficient. The objective of a study may more likely need to determine whether the difference observed is large enough to be useful in some way.

Avoid repeating lists of results that can be seen more clearly in a table. Whole paragraphs that just reproduce in long-hand the data in a table are just a waste of space and effort. Some results in each table or graph must be referred to in the text. A table, or part of a table, that is ignored in the text can usually be omitted. Also avoid discussing the results in the Results section. Discussion of results should be held back until the Discussions section. Of course, any experimental results that you wish to discuss should be included in the Results section.

When writing a report it will usually be necessary to include a description of the statistical methods that have been used. This is mainly done so that a reader can see whether he/she feels that the methods are appropriate or not. If the data are archived then it should be possible to repeat the analysis from the description given in the paper. Sometimes in simple cases it is possible to give a clear description very briefly – e.g. the data were analysed by analysis of variance for a randomised block design. However, in more complex situations it is difficult to be brief. The analysis may have taken various steps or the statistical model may be complicated. Sometimes it may be easiest to express the statistical model algebraically. A biometrician should be able to help the researcher in difficult situations such as this.

Exercise 6.5. Write a statistical report describing the data analysis carried out in Experiment A.

Exercise 6.6. Write a similar report for Experiment B. In this case the decisions to omit 3 animals and the calculations of the summary statistics for growth rate and PCV used in the final analysis will need to be emphasised.

Chapter 7 Ideas of simple statistical inference

We have already drawn some ideas of statistical inference when we described confidence intervals. From a knowledge of the variation of the observations within our sample we make inferences about how closely we could estimate the range within which the population mean is likely to lie. We can, however, use statistical inference in other ways to assess the significance of the results from an experiment. There are a number of statistical tests that allow us to do this. Each of these tests are concerned with the examination of a null hypothesis that there is no difference in the means being compared and with determination of the probability level at which the null hypothesis can be rejected. The null hypotheses relate to the objectives set out when the experiment was planned.

t-test

The simplest and best known test is the t-test. It is normally used for determining the statistical difference between two means.

The formula for the t-test is

t = (difference between 2 means)/(SE of difference between 2 means)

The **SE of the difference between 2 means** (**SED**) is calculated as the square root of [the variance multiplied by $(1/n_1 + 1/n_2)$] where n_1 and n_2 are the numbers of observations respectively making up the 2 means to be compared. When $n_1 = n_2 = n$ the formula reduces to $\sqrt{}$[2 x variance x (1/n)] or $\sqrt{2}$ SE.

To illustrate better the use of the t-test let us return to Example A in which we used the 4 group means to form a 1-way analysis of variance.

Thus, to examine the increase in weight gain brought about by supplementing the Dorper, we get

t = (2.39-1.32)/($\sqrt{}$2x0.236) where 0.236 is the variance determined from the residual MS in the analysis of variance

= 1.07/0.334 = 3.20

We compare this value with t-values in the t-table using the number of degrees of freedom, namely 28, for the residual variance. With 28 df we get t-values of 2.05, 2.76 and 3.67 at 5%, 1% and 0.1% levels of significance, respectively. The value of 3.20 is less than 3.67 but greater than 2.76, and so the difference in weight gain brought about by supplementation is significant ($P<0.01$). This means that the probability of rejecting the null hypothesis that there is no effect of supplementation when the null hypothesis is true less than 1 in 100. In other words with a probability greater than 99 in 100 we can deduce that there is a real difference between the means. However, if we compare non-supplemented Red Maasai with non-supplemented Dorper we get

$t = (1.49 - 1.32)/(\sqrt{2} \times 0.236) = 0.51$

which is not significant. In other words we are unable to reject the null hypothesis that there is no difference between the unsupplemented means.

Exercise 7.1. Use Genstat to select the 16 Dorper sheep (Breed 1) in Experiment A. Then carrying out a t-test on this data subset to examine the effect of supplementation. Compare the t-value with the value of 3.20 above. Explain why they are different.

Least significant difference

It is sometimes more convenient to calculate the **least significant difference** (**LSD**) needed for the difference between 2 means to be significant. This is calculated as

LSD = t x SED

Thus, for the 4 means in the body of the table

LSD = 2.05 x 0.334 = 0.68 at the 5% level of significance
= 2.76 x 0.334 = 0.92 at the 1% level of significance
= 3.67 x 0.334 = 1.23 at the 0.1% level of significance

We can then compare differences between the respective means with these values to determine their significance. Thus, for example,

2.39 - 1.32 = 1.07 : significance level P<0.01
3.48 - 1.89 = 1.59 : " " P<0.001
1.49 – 1.32 = 0.51 : not significant

Note that the t-value in the table decreases as the number of degrees of freedom for the residual term increases. This is because the greater the number of degrees of freedom the more precisely is the variance known and the closer it is to the true population value. This is important in relation to the design of experiments. Very small experiments yield few degrees of freedom for the residual variance making it more difficult to detect statistical significance. It is also important to emphasise here that the correct formula for a confidence interval is mean ± t x se, not the approximate formula mean ± 2 x se used earlier. As can be seen from the t-tables, however, t approximates to 2 for large degrees of freedom.

Exercise 7.2 Run again the analysis of variance for Experiment A weight gain in Genstat but this time request for SEs of differences, not for means. Use the appropriate SED to calculate LSDs at 5, 1 and 0.1% levels for comparing the means in the body of the breed x diet table.

F-test

This test is used for assessing the statistical significance of variance ratio (VR) values in analysis of variance. Unlike the t-test, it has a pair of degrees of freedom, the first referring to those for the term in the numerator and the second to those for the residual. For example, let us consider the one-way analysis of variance for group in the previous chapter. We put

F = VR = 7.8138 / 0.4474 = 17.5

We use this value to compare with values in F-tables with 3 and 28 degrees of freedom. If we look at the column headed 3 df at the top of each of the F-tables at the end of these notes and then come down this column until we reach 28 df on the left hand side, we find the values 2.95, 4.57, and 7.19 for 5%, 1% and 0.1% levels of probability, respectively. If our observed F-value exceeds any of these values then we say it is statistically significant. The value of 17.5 exceeds the value of 7.19. Thus, the analysis of variance shows that there is highly significant variation in liveweight gain among the 4 groups of lambs at the 0.1% level of statistical significance. This is normally described using the notation ($P<0.001$). In other words, the probability of there being no differences among the 4 groups of lambs is less than 1 in 1000. It should be noted here that we are non the wiser as to which groups differ from one another. All that we know is that there are differences. To ascertain where the differences lie we go on to apply the t-test between pairs of means.

If we now consider the interaction between breed and supplementation in the factorial analysis of variance given in Chapter 5, and put $F_{1,28} = 3.82$ we find that the P-value is just above P = 0.05 ; indeed we can guess it to be 0.06. Thus, the probability of an interaction occurring by chance = 6/100. The question is - is the size of the interaction sufficiently important to report this P-value, or is the size of insufficient biological importance? This is for the researcher to decide. It is important in any data analysis to distinguish between statistical significance and practical importance.

Exercise 7.3. Rerun the analysis of variance of weight gain in Experiment A using the statistical models fitted in Chapter 5. This time do not click 'options'. Look at the P-values for the interaction, for block and for sex and decide whether to include any of these in the final model to use for the reporting of the results.

In summary, statistical analysis provides a number of tools that allows the researcher to draw statistical inferences about the patterns he observes in his data. In writing up results the researcher should remember that the role of statistics is to simply provide some of the numerical evidence to support the arguments being made. Thus, in the respect it is just a tool. The researcher is the best judge of the practical importance of his results – he therefore uses the results of statistical analysis to provide some basis for making his/her conclusions.

It should also be noted that there are certain assumptions that are made in the use of analysis of variance and the F and t-tests. For instance, it is assumed that observations in the data being analysed are independent and random, follow a normal distribution and all groupings of the data are drawn for populations with the same residual variance. However, F and t-tests are **robust** in the sense that they can handle slight deviations

from normality and slight variations in variation across the population without affecting the general inferences that can be drawn.

In general if an F-test is not significant then one should not go on and compare different individual means by the t-test. But this rule does depend on the initial null hypotheses established when the experiment was planned. A common sense approach is needed. What is inappropriate is to carry out a 'fishing' exercise searching for statistical significance between different means without regard to the null hypotheses.

One-tailed tests

When a t-test is used it is usually applied as a 2-tailed test. This means that we are considering significant deviations in both directions, i.e. treatment A could have a mean that is higher or lower than treatment B. Our null hypothesis is treatment A = B and our alternative hypothesis is A ≠ B. Sometimes we may set up an experiment, often with a control, in which *a priori* we have decided to ascertain whether the treatment is 'better' than the control. In other words we are evaluating the alternative hypothesis treatment A > B. This is known as a one-tailed test. For example, we may wish to test whether a vaccine results in some degree of protection compared with unaffected controls, or we may wish to evaluate whether a new form of therapy is better than are existing one. If we decide that a one-tailed t-test is appropriate then we look up the 10% rather than the 5% value, or the 2% rather than the 1% value, in the statistical table. Thus, we double the size of each tail so that we look at only one side of the distribution. If we detect a significant difference in the opposite direction then we ignore it as it is inconsistent with our alternative hypothesis.

Excercise 7.4 Run one-way analyses of variance in Genstat to compare mean PCVs and weight gains for animals in the 3 immunised groups in Experiment B. Apply t-tests to compare difference between means. Combine the means for the ARF1 and p32 immunised groups and calculate the t-value for comparing with the negative control. Which of your comparisons are consistent with the null hypotheses defined when the experiment was designed? Do you think that we should be using one-tailed or two-tailed t-tests in this example?

Paired t-test

A paired t-test is sometimes applied. This is for situations where animals are blocked into pairs for assignment of two treatments, one to each pair, or where an observation is recorded on an individual before and after some treatment or intervention is applied. This is analogous to a randomised block analysis of variance with two treatments. Suppose for example 5 cows were infected with trypanosomes and then treated with a trypanocidal drug. If we are interested in extent of recovery of PCV then we would measure PCV at two time points before and after treatment. The experimental unit in this case is each individual sampling occasion within the cow and the statistical model can be written:

$y_{ij} = \mu + c_i + t_j + e_{ij}$

where c_i (i = 1-5) is the effect of cow and
t_j (j = 1,2) is the effect of treatment (or sampling time before or after treatment).

The analysis of variance table therefore looks like

Source	df
Cow	4
Treatment	1
Residual	4
Total	9

Alternatively, we can calculate the difference in PCV measured on the two occasions to obtain five new values which represent the change in PCV for each cow. We can use these values to calculate the mean difference and the standard deviation of these differences, and apply a t-test to compare the mean difference against zero. Thus if $\bar{x}_{diff}$ represents the mean difference then

$$t_4 = \bar{x}_{diff} / \sqrt{\text{variance}(\bar{x}_{diff})/5}$$

One can demonstrate that the above variance is the same as the residual MS in the analysis of variance table. Indeed if one compares $F_{1,4}$ and t_4 values (say at the 95% level) in the statistical tables it can be seen that $F_{1,4} = t_4^2$. Thus, the paired t-test is an alternative to the analysis of variance for a randomised block experiment when there are only 2 treatment levels. Likewise, a non-paired t-test is an alternative method for analysing a completely randomised design with two treatments.

Exercise 7.5 Using the data in Exercise 7.1 carry out a one-way analysis of variance to compare supplementation versus non-supplementation and show that the same results are obtained as for the t-test applied in Exercise 7.1.

Heterogeneity of variance

What do we do when variances vary across the populations we are comparing? In comparing the Red Maasai/Dorper experiment the variances for weight gains among individuals can be calculated separately for each breed-diet group and shown to be 0.3793, 0.9041, 0.2955, 0.2107 kg^2, respectively. The residual variance used in the analysis of variance is the average of these. However, the individual variance, for groups 1, 3 and 4 are smaller than for group 2. The analysis of variance we have undertaken assumes that each variance calculated for each group estimates the same 'population' variance. This may not be true. Even if it is not true, do we have to worry about it? Probably not. It is likely that the weight gains of lambs in group 2 were by chance more variable than in the other groups. As already mentioned the analysis of variance technique is robust. In other words it handles data for which assumptions, such

as homogeneity of variance of symmetric, normal distributions, do not strictly apply. But what alternative methods are at our disposal? Faecal egg count (FEC), for instance, has a very skewed, non-normal distribution and an analysis of variance undertaken on these data may not be justified.

Transformations

Firstly we could transform FEC to a different scale by using a calculation such as log (y), log (k + y) or $\sqrt{y}$. Each of these calculations reduces the skewness of the data and makes the residual variance less dependent on the mean. The logarithm transformation is stronger than the square root one. The log (k + y) transformation, where k is a constant >0, is used instead of log (y) when individual values of y occur which equal 0. The disadvantage of this method is that, although it may produce a more valid statistical analysis, it is often not so easy to present the results. This is because the means and standard errors calculated during the statistical analysis are now on a different scale.

Exercise 7.6 Run an analysis of variance for FEC in Genstat, both on the original values and after transforming the data to $\log_e$ (50 + FEC). Compare the results. What difference, if any, is there in the interpretation of the significance of terms in the analysis of variance?

Sometimes a transformation is not appropriate; an alternative method might be to split the data into separate parts and to do a separate analysis of variance on each part.

It is conventional to analyse FEC on the log scale. An arcsin ($\sin^{-1}\sqrt{y}$) or square root transformation is more appropriate for disease prevalence data, the square root in particular when the prevalence is low. However, there are also alternative methods using what is known as **logistic regression** or **log linear models** which are more suitable for such data. These methods are not described here, being somewhat too advanced for this course.

Results can be expressed on the original scale by calculating the geometric mean (antilog of the transformed mean) and the antilog of the 95% confidence limits for the transformed mean. Take, for example, the Dorper lambs (breed 1) supplemented with cottonseed cake and bran meal. The mean and standard error for FEC calculated on the log scale ($\log_e$ y) are 7.26 and 0.36, respectively. The 95% confidence range for this mean is $7.26 \pm t \times 0.36$ where t, with 7 degrees of freedom, = 2.36. Thus, 95% confidence interval for mean = (6.41, 8.11). If we transform these values back to the original scale using the antilog (or e^x) transformation, then mean = 1422, range = (608 to 3327). Thus, we can say that the geometric mean is 1422 eggs per gram (epg) and that 95% confidence interval within which the true 'population' mean lies is between 608 and 3327 epg.

Three important points need to be highlighted:

- It is not permitted to calculate an antilog for the standard error to produce a standard error on the original scale.

- The upper and the lower confidence interval values are not equidistant from the mean.

- If $\log_e (50 + y)$ is used for the transformation for instance, then 50 must be taken off the antilogs after transformation back to the original scale in order to calculate the correct geometric mean and range.

Another variable that is often transformed to logarithms is antibody titre. Serum is usually diluted in a geometric series, i.e. with a constant ratio between successive dilutions, e.g. 1/2, 1/4, 1/8, 1/16 etc. Thus, frequency distributions of titres tend to be lognormal and so the logarithm transformation is usually necessary.

Analysis of discrete data.

Some of the more general methods for the analysis of discrete data (e.g. logistic regression) are beyond the scope of this course. Discrete data are data that are not continuous. For example, variables which take on the values 0 or 1 are often referred to as binary variables. Mortality is such an example. Others are disease incidence, conception (yes or no) etc. Statistical models can be prepared using discrete variables as the response variable in just the same way as continuous variables. When the statistical model is very simple the data can often be analysed using a χ^2 test.

Chi-square (χ^2) test

In simple experiments data can often be grouped into a two-way table and a χ^2-test performed. Thus, for example, if the lambs used in this experiment were drawn from a group of 55 Dorper and 42 Red Maasai lambs and 14 and 4, respectively, died before weaning, then we can compare difference in mortality by forming a table, often known as a **contingency table,** as follows.

Breed	Alive at weaning		
	No	Yes	Total
Dorper	14	41	55
Red Maasai	4	38	42
Total	18	79	97

The above values are the ones observed. The next step is to calculate the values that would have occurred had the mortality rates in the two breeds been the same. These are known as expected values. Expected values for each of the 4 cells in the body of table are obtained by dividing the corresponding totals in the right hand column by the grand total to give the proportion of animals of that breed) multiplied by the corresponding total in the bottom row. Thus, expected values are, respectively:

55/97 x 18 55/97 x 79
42/97 x 18 42/97 x 79

$$= \begin{matrix} 10.2 & 44.8 \\ 7.8 & 34.2 \end{matrix}$$

We are interested in knowing whether the observed values differ significantly from the above expected values calculated assuming no association between breed and mortality. We do this by use of what is known as a χ^2 test.

The formula for the χ^2 test is $\sum(o_i - e_i)^2/e_i$
where o_i = observed value and e_i = expected value (i = 1, 2, 3, 4)

$$\text{Thus } \chi^2 = (14 - 10.2)^2/10.2 + (41 - 44.8)^2/44.8 + (4 - 7.8)^2/7.8 + (38 - 34.2)^2/34.2$$
$$= 1.416 + 0.322 + 1.851 + 0.422$$
$$= 4.01$$

The number of degrees of freedom = (2-1) x (2-1) = 1. The reader will have noticed that only one cell in the body of the table needs to be independently calculated. Values for the other cells can be derived by substracting the first cell from the adjacent totals. Thus, the number of degrees of freedom is 1.

The 5% and 1% values for χ^2 with 1 degree of freedom are 3.84 and 6.63, respectively. Thus, we can say that the mortality rate before weaning in the Dorpers of 14/55 (25%) was significantly higher than that of 4/42 (10%) in the Red Maasai ($P < 0.05$).

The χ^2 test is an approximate test and can only be satisfactorily applied when the numbers in the table are reasonably large. When any value of e_i is less than 5, the above formula should be replaced by:

$$\sum(|\, o_i - e_i \,| - 0.5)^2/e_i$$

where $|\, o_i - e_i \,|$ is the positive value of the difference between o_i and e_i. This is because the χ^2 test in its unadjusted form over-estimates the statistical significance of any association. In other words the test rejects more often than it should the null hypothesis that the values in the table are independent and occur by chance. Note that this method of adjustment is only valid for 2 x 2 tables and not for tables with more than 2 rows or columns.

Exercise 7.7
Calculate the numbers of animals in Example B that had to be treated and removed from the experiment because of low PCV. Create a 3 x 2 spread sheet in Genstat showing the numbers of animals surviving and not surviving for each treatment group. Carry out a χ^2 test on this table. How many degrees of freedom are there? Combine the two immunised groups and repeat the χ^2 test on the 2 x 2 table. Comment on the outputs.

Fisher's exact test

You may have noted that the Fisher's exact test is an alternative option to the chi-square test for the analysis of contingency table data in Genstat. This is especially useful for experiments involving small numbers of animals and especially when the experiment is planned to determine whether a treatment is better than the control in the response it produces. It has already been stated that the χ^2 test is only an approximation for small sample sizes. In experiments involving cattle the researcher can often only afford few animals. This applies particularly to vaccine experiments where levels of protection are being evaluated. In such cases it is reasonable to apply a one-tailed test, in other words the alternative hypothesis is that the level of protection offered by the vaccine is greater than zero. The Fisher's exact test provides this possibility. This test calculates for all possible 2-way tables the probability of each table occurring by chance, and then adds the probability of the table observed to the probabilities of those tables occurring that are more extreme, both in one direction only (1-tailed test) or in both directions (2-tailed test). With a small sample size we would normally have designed an experiment with a 1-tailed test in mind. The Fisher's exact test is superior to the χ^2 test for small sample sizes. Furthermore, it is not possible to think of the χ^2 test as providing alternative one-tailed and two-tailed tests. It behaves in a similar way to a 2-tailed Fisher's exact test.

Exercise 7.8 Apply the Fisher's exact test to the contingency table representing numbers of Dorpers and Red Maasai alive or not at weaning and compare the result with that of the χ^2 test (Ignore the mid-point P value given by Genstat for the Fisher's exact test.)

Chapter 8 Sample size

Having now acquired an understanding of the development of statistical models for the purpose of carrying out analysis of variance it is useful to briefly return to the question of experimental design. Once the plan for an experiment has been made and the number of animals decided it is useful to sketch out the skeleton of the shape that the analysis of variance will take by listing the numbers of degrees of freedom. As described earlier it is important to ensure that there are reasonable numbers of degrees of freedom for the residual term, not only to ensure that a reasonable estimate of the residual variance is obtained, but also to ensure that t and F significant values are not over-inflated. Also it is important to ensure that there is sufficient replication to provide standard errors that are small enough to detect as significant the required differences in means. The following exercise illustrates these points.

Exercise 8.1 You plan to carry out a farm trial to assess the effect of anthelmintic treatment on growth rate of lambs to weaning. You have 10 flocks, each on a different farm, at your disposal. It will be possible to divide 4 of the flocks so that each half grazes in a different paddock. The flocks in the other farms cannot be subdivided in this way. It has been decided that it would not be desirable to mix treated and untreated sheep in the same grazing pasture. Design two experiments, one based on just the 4 farms with 2 paddocks, the other based on all 10 farms, using one paddock from each farm, to carry out the objective. Write down the statistical models in each case and sketch the form of the analysis of variance table. Which are the experimental units in each case which design do you prefer?

Determination of sample size

Before conducting an experiment it is important to decide on how many animals, flocks etc. one may need in order to estimate treatment means to a desired precision. In other words, we need to know how many replicates or blocks we need in order to have a reasonable chance of detecting a real difference at a given level of statistical significance.

We have already seen that the standard error of the difference between 2 means, each of n observations, is given by the formula:

$$\text{SED} = \sqrt{2 \times \text{variance} / \text{n}} = \quad 2\text{SE}$$

One might have a preconceived idea of the approximate size of the standard deviation to be expected. Thus, if the experiment described in our example were to be repeated using two different breeds, we could assume that we might expect a residual variance of about 0.5 (It was 0.4474 in the experiment). Suppose we decided that we needed to determine a difference of 1.0 kg in weight gain between two groups to be statistically significant ($P<0.05$). We can then apply the formula.

$$LSD = t \times SED$$

$$= t\sqrt{2 \times \text{variance} / n}$$

Squaring both sides and manipulating terms we get

$$n = 2\,t^2 \text{ x variance} / LSD^2$$

Substituting t = 2 (approximate value for P = 0.05 with large number of degrees of freedom)
we get

$$n = (2 \text{ x } 4 \text{ x } 0.5)/1.0^2$$

$$= 4.0$$

Therefore we can expect to require 4 animals per group to show a difference in weight gain of 1.0 kg as significant between diets (P<0.05).

With a completely randomised design with 4 groups (breed x supplementation) we would have 12 degrees of freedom (i.e. 4 x 3 degrees of freedom per group) for the residual mean square. $t_{12} = 2.18$ (P = 0.05). Replacing t^2 with 2.18^2 in the above formula we can revise our sample size estimate to

$$n = (2 \text{ x } 4.75 \text{ x } 0.5)/1.0^2 = 4.75$$

Therefore we would, in fact, need 5 animals per diet.

Exercise 8.2 Run an analysis of variance for minimum mean PCV in Example B. It has been decided to repeat the experiment to compare again the effect of p32 immunisation against the control. Decide what minimum difference in PCV you would like to detect as being significant (P < 0.05). By using the Genstat output to provide an estimate of s^2 calculate how many animals will be needed within each group.

Combining experiments

When experiments are replicated it is often possible to analyse them together within the one analysis of variance framework. This is a useful and practical way of increasing sample size.

Suppose, for example, that the immunisation experiment illustrated in Example A is repeated with 8 animals inoculated with the p32 form of the vaccine and 8 negative controls. The hypothesis now is that p32 itself offers protection. We can now combine the two experiments as follows:

	Number of animals	
Experiment	p32	Control
1	7	5
2	8	8
Total	15	13

Thus, total sample sizes are 15 and 13, respectively. The analysis of variance structure for analysing both experiments together will be as follows:

Source of variation	df
Experiment	1
Immunisation	1
E x I	1
Residual	24
Total	27

Exercise 8.3 Calculate how many sheep per group you would need to determine a difference of 0.5 kg as significant ($P<0.05$) in the breed x supplementation experiment (Experiment A) if your estimated standard deviation was 0.8 kg. Suppose that the number you obtain is far more than the number of sheep available to the researcher. Discuss what advice you might give to the researcher.

Example A

Record	ID	breed	sex	supp	block	wt_3mo	wt_6mo	pcv	fec	wt_gain
1	349	1	2	1	1	8.0	8.9	10	6500	0.9
2	326	1	2	1	1	9.0	10.1	11	2650	1.1
3	393	1	1	1	2	12.0	12.6	22	750	0.6
4	71	1	1	1	2	12.3	14.6	15	5200	2.3
5	271	1	1	1	3	13.0	13.7	19	4800	0.7
6	382	1	2	1	3	15.5	16.8	24	2450	1.3
7	85	1	2	1	4	16.3	18.2	27	200	1.9
8	176	1	2	1	4	15.9	17.7	21	3000	1.8
9	286	1	2	2	1	11.0	13.6	21	1600	2.6
10	183	1	1	2	1	9.9	11.7	21	450	1.8
11	21	1	2	2	2	11.6	13.1	25	2900	1.5
12	122	1	1	2	2	12.5	14.8	25	300	2.3
13	374	1	1	2	3	14.6	17.9	19	2250	3.3
14	32	1	2	2	3	14.2	16.9	22	2800	2.7
15	282	1	2	2	4	16.3	20.2	20	750	3.9
16	94	1	1	2	4	16.7	17.7	13	5600	1.0
17	127	2	2	1	1	7.5	8.1	26	1350	0.6
18	216	2	2	1	1	8.2	9.3	19	1150	1.1
19	133	2	1	1	2	10.1	11.7	30	200	1.6
20	249	2	1	1	2	8.8	10.4	28	0	1.6
21	123	2	2	1	3	11.6	12.6	23	600	1.0
22	222	2	2	1	3	11.3	13.5	24	1500	2.2
23	290	2	2	1	4	12.3	14.3	22	1950	2.0
24	148	2	1	1	4	13.1	14.9	26	500	1.8
25	142	2	2	2	1	8.2	11.5	25	850	3.3
26	154	2	2	2	1	8.5	12.2	35	700	3.7
27	166	2	1	2	2	9.7	12.8	29	400	3.1
28	322	2	1	2	2	8.6	12.0	26	800	3.4
29	156	2	1	2	3	10.2	13.0	28	1550	2.8
30	161	2	2	2	3	11.2	14.6	22	550	3.4
31	321	2	1	2	4	12.1	15.9	25	1250	3.8
32	324	2	1	2	4	13.8	18.1	24	1100	4.3

Example B-WT

ANIM	DOB	BLOCK	GROUP	SEX	INITPCV	D7	D9	D10
4	6/4/1999	1	A	F	39.20	38.30	35.60	31.60
6	7/18/1999	2	A	M	33.60	34.00	31.60	31.30
8	7/1/1999	2	A	M	32.50	34.70	30.10	30.10
11	7/4/1999	2	A	F	30.70	33.40	31.60	29.80
16	6/12/1999	1	A	F	31.15	31.30	31.60	28.60
19	7/8/1999	2	A	M	36.50	35.30	35.30	31.00
20	6/24/1999	1	A	F	29.80	28.00	27.10	27.10
1	7/5/1999	2	B	F	31.75	31.60	31.30	30.40
5	6/1/1999	1	B	F	36.80	37.40	30.40	35.30
9	7/5/1999	2	B	M	34.85	33.10	33.40	29.80
12	7/15/1999	2	B	F	29.95	32.50	30.40	30.40
14	6/2/1999	1	B	F	36.65	37.40	35.60	33.10
17	7/4/1999	2	B	F	36.20	35.60	34.70	34.70
18	6/20/1999	1	B	M	33.70	34.00	31.00	28.90
2	6/28/1999	2	C	M	30.55	32.50	31.60	30.40
3	6/6/1999	1	C	F	28.85	31.60	32.50	31.30
7	7/28/1999	2	C	F	31.90	31.30	32.20	30.70
10	6/10/1999	1	C	M	25.25	27.70	25.20	24.60
13	6/9/1999	1	C	M	30.40	31.60	31.00	30.40
15	7/2/1999	2	C	F	35.60	34.00	33.10	30.10
21	7/1/1999	2	C	M	33.25	34.40	30.70	30.40

D11	D12	D13	D14	D15	D18	D19	D20
34.40	33.40	36.50	35.30	34.40	33.70	31.60	32.20
31.00	30.40	29.80	29.80	28.00	26.80	25.80	26.80
28.60	29.50	27.70	29.20	25.80	25.20	26.10	26.10
28.30	25.80	27.70	27.10	26.10	23.70	23.40	24.00
28.00	27.10	28.00	27.40	24.60	22.20	21.90	21.60
29.80	31.90	28.90	28.90	29.20	28.90	26.80	26.80
26.10	24.00	24.60	21.60	19.80	18.80	18.80	17.60
31.00	30.70	28.90	28.60	25.20	25.50	25.20	25.20
35.30	35.30	30.10	33.10	30.10	27.10	27.10	24.60
30.10	32.20	30.10	31.30	31.00	28.00	26.80	26.40
28.90	27.70	28.60	28.90	27.10	27.70	25.50	23.40
34.70	35.30	34.00	31.90	31.90	29.50	29.50	28.60
33.70	33.40	31.30	31.90	31.30	29.50	28.30	26.40
27.40	26.40	27.40	27.70	28.30	26.40	26.80	24.00
30.40	28.90	27.40	24.90	26.10	24.60	24.30	24.30
31.30	31.30	30.10	28.90	29.20	28.30	29.20	26.80
28.90	27.70	26.40	28.90	29.20	29.20	29.20	28.60
24.00	24.90	24.60	24.90	23.70	24.60	22.20	22.50
28.00	30.10	29.20	29.20	29.20	29.80	28.30	31.00
29.50	27.40	28.00	27.70	25.80	24.60	23.70	24.90
28.00	29.80	27.40	28.90	28.90	28.60	27.70	25.50

D21	D22	D25	D28	D32	D36	D39	D42
30.40	31.90	28.30	30.10	26.10	26.10	24.00	20.40
25.50	25.50	24.60	25.50	23.70	23.70	19.80	20.40
26.10	25.20	22.50	24.30	22.50	22.50	17.90	16.70
24.00	24.30	23.10	23.70	22.20	21.30	20.10	15.80
22.80	22.20	19.80	19.80	16.40	17.00	14.60	14.90
27.40	27.70	22.50	22.80	19.80	18.80	16.70	16.10
18.20	18.20	15.80	15.50	13.40	14.00	12.50	12.50
22.20	22.20	18.20	17.30	13.10	14.30	13.40	12.50
25.80	24.90	20.70	22.80	19.50	19.50	17.90	16.70
26.10	25.50	26.10	26.80	21.00	19.50	18.20	19.80
26.40	25.50	23.10	24.00	20.10	23.40	20.40	20.40
28.30	28.90	26.10	29.50	25.50	23.10	20.10	21.00
27.70	28.60	24.60	24.00	22.80	22.80	21.30	19.80
26.10	25.50	22.80	23.10	22.20	21.00	18.50	18.20
22.50	23.70	22.20	20.40	16.40	15.80	14.30	13.70
27.10	26.80	24.30	26.40	26.10	23.40	22.20	21.00
29.20	27.70	28.30	30.70	31.90	31.90	30.70	29.20
22.20	24.30	22.80	23.70	20.70	21.00	17.90	16.70
27.70	29.20	31.00	30.10	28.90	31.90	31.00	30.40
23.70	23.10	20.70	20.40	18.20	18.20	17.30	15.50
25.50	25.20	23.10	22.20	21.30	18.50	16.40	13.40

D46	D49	D53	D56	D60	D63	D66	D69
19.80	19.80	17.90	18.20	18.80	18.50	18.80	19.20
21.00	20.40	18.50	22.50	21.30	21.60	22.80	24.00
16.10	16.40	14.90	17.00	15.50	15.80	16.10	16.10
21.90	19.20	20.70	22.80	21.30	22.50	21.30	22.50
14.60	13.70	12.20	12.20	12.50	14.00	12.50	12.20
16.70	16.10	15.50	16.10	16.70	15.80	17.90	16.10
12.20	12.20	11.90	11.90	11.90	12.50	12.80	12.50
13.40	13.10	12.80	14.30	12.80	14.00	14.00	13.40
16.10	17.60	19.20	19.20	18.80	17.30	18.20	20.40
18.80	19.20	20.40	20.40	19.50	19.50	24.00	22.80
21.60	21.30	19.80	21.60	21.90	21.90	21.60	22.80
21.00	18.50	17.60	18.20	19.20	17.30	18.20	17.30
19.50	21.00	18.20	21.00	21.30	21.60	23.70	21.90
18.50	19.50	21.00	20.10	20.70	20.10	21.60	20.70
13.10	14.30	13.40	11.90	13.70	12.80	14.30	14.30
17.00	16.40	14.30	14.90	16.70	17.00	17.30	17.30
28.30	30.70	29.20	28.00	28.00	27.40	35.60	29.20
15.20	14.30	14.00	14.60	14.30	14.30	14.60	13.70
31.60	31.60	31.30	31.30	31.60	30.10	31.60	30.10
13.70	15.20	14.60	15.20	17.30	17.00	16.70	15.80
12.20	12.50	12.20	13.10	12.80	12.50	0.00	0.00

D73	D76	D80	D83	D87	D90	D94	D97
18.50	19.20	17.90	20.70	18.50	18.50	20.10	21.30
22.80	22.20	21.30	23.10	21.60	22.50	23.70	23.40
14.60	15.80	12.20	13.70	12.80	13.70	14.00	13.70
21.00	21.00	20.70	21.00	19.50	20.70	21.60	22.50
11.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16.40	15.50	14.90	15.80	15.50	14.00	16.10	14.90
11.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18.80	19.80	19.20	19.20	21.30	20.10	19.80	24.00
20.10	19.50	18.20	19.20	18.20	19.50	21.90	23.10
21.60	21.60	24.00	24.90	23.40	26.80	23.10	23.10
18.50	19.20	16.70	17.00	16.10	17.00	19.20	20.10
23.70	24.30	23.70	25.20	25.80	24.90	23.70	27.40
21.60	20.10	21.60	22.20	20.40	21.60	21.30	23.10
12.20	14.90	13.70	11.60	12.80	11.90	11.60	12.20
16.40	15.50	15.80	15.80	15.50	15.80	16.70	16.10
29.20	28.30	28.00	29.80	29.50	27.40	25.20	30.70
14.60	13.40	12.80	13.70	0.00	0.00	0.00	0.00
29.50	28.30	28.90	28.30	28.30	29.20	29.80	29.20
16.40	16.40	15.80	15.80	16.10	18.20	16.70	15.80
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

D101	D104	D108	D111	D115	D118	D122	D125
20.70	21.00	18.80	21.90	19.50	21.60	21.90	22.50
21.90	26.40	22.50	22.50	21.00	22.50	21.00	22.50
13.10	13.70	12.80	13.40	0.00	0.00	0.00	0.00
22.50	25.50	22.20	23.70	21.00	22.20	23.70	22.50
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13.70	13.70	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21.30	21.90	21.60	20.70	20.40	19.50	21.60	19.80
20.40	21.30	21.30	21.30	19.80	20.70	18.50	19.80
23.10	23.70	21.90	23.70	22.50	23.70	23.70	23.40
19.50	19.80	18.20	17.90	17.00	18.20	18.20	17.30
25.80	27.70	26.80	25.80	28.60	26.40	27.40	27.40
22.50	22.20	21.30	22.80	20.70	22.20	20.70	21.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15.80	15.50	14.90	17.00	14.60	14.00	14.60	13.70
28.30	28.90	27.40	27.70	24.30	24.90	24.30	28.60
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30.10	30.70	29.80	27.10	26.80	27.10	26.10	27.10
19.50	18.20	17.00	16.70	17.30	19.50	16.40	16.40
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

D129	D132	D136	D139	D143
22.80	23.40	23.40	25.50	22.20
22.20	24.60	21.60	22.80	24.90
0.00	0.00	0.00	0.00	0.00
23.40	24.00	20.10	21.90	23.40
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00
20.70	19.80	21.00	19.50	22.20
21.00	21.60	21.60	22.80	21.30
22.20	24.30	24.60	25.20	24.00
16.40	19.80	17.90	19.50	20.40
25.80	29.20	27.40	28.30	25.50
20.70	22.80	20.40	18.20	17.60
0.00	0.00	0.00	0.00	0.00
14.60	15.20	15.80	0.00	0.00
28.00	28.30	28.30	28.30	26.40
0.00	0.00	0.00	0.00	0.00
26.40	26.40	28.60	27.40	24.90
16.10	17.30	17.90	19.80	16.70
0.00	0.00	0.00	0.00	0.00

Example B-PCV

ANIM	GROUP	SEX	DOB	BLOCK	INITWT	WT0	WT12	WT19
4	A	F	6/4/1999	1	142	180	178	180
6	A	M	7/18/1999	2	112	148	140	148
8	A	M	7/1/1999	2	120	150	142	152
11	A	F	7/4/1999	2	122	156	148	152
16	A	F	6/12/1999	1	132	160	116	110
19	A	M	7/8/1999	2	122	160	152	160
20	A	F	6/24/1999	1	108	147	124	132
1	B	F	7/5/1999	2	114	139	136	138
5	B	F	6/1/1999	1	160	198	194	198
9	B	M	7/5/1999	2	136	172	168	174
12	B	F	7/15/1999	2	100	118	116	120
14	B	F	6/2/1999	1	140	200	200	200
17	B	F	7/4/1999	2	136	210	200	200
18	B	M	6/20/1999	1	112	130	142	152
2	C	M	6/28/1999	2	122	152	144	152
3	C	F	6/6/1999	1	64	144	140	144
7	C	F	7/28/1999	2	110	132	126	128
10	C	M	6/10/1999	1	124	120	120	126
13	C	M	6/9/1999	1	120	148	144	148
15	C	F	7/2/1999	2	114	138	146	154
21	C	M	7/1/1999	2	126	164	156	158

WT28	WT36	WT42	WT49	WT56	WT63	WT69	WT76	WT83
190	190	190	188	182	182	184	178	178
150	156	158	162	160	162	162	162	164
150	154	152	148	142	140	134	138	128
156	160	160	162	162	170	170	172	172
104	108	102	102	98	96	98	0	0
164	164	160	160	152	156	150	152	146
132	132	130	132	132	128	124	0	0
148	144	138	140	138	136	132	0	0
200	204	200	200	196	198	196	202	198
172	180	178	176	174	182	176	180	178
124	126	128	132	138	138	138	142	144
204	210	212	212	202	206	200	196	194
202	208	210	210	200	204	206	206	208
154	144	152	148	146	148	152	152	152
158	158	152	144	146	146	138	140	138
148	154	144	146	140	146	134	136	130
136	142	140	142	144	150	150	156	158
122	122	120	118	114	110	112	108	108
142	152	156	166	160	162	160	164	164
140	138	136	136	132	130	134	126	130
160	162	158	160	148	148	0	0	0

WT90	WT97	WT104	WT111	WT118	WT125	WT132	WT139
180	182	182	180	182	182	184	182
166	164	166	172	170	168	170	170
124	122	122	116	0	0	0	0
178	172	174	178	178	180	180	182
0	0	0	0	0	0	0	0
146	142	140	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
198	195	202	196	198	202	200	200
170	166	174	172	170	170	168	168
148	148	148	148	150	148	152	148
198	194	198	198	198	194	196	196
210	210	210	218	214	216	214	216
148	150	150	154	156	154	154	154
134	134	0	0	0	0	0	0
124	126	124	118	118	116	116	0
162	154	156	158	160	162	162	160
0	0	0	0	0	0	0	0
172	170	172	174	176	176	178	180
128	130	128	128	126	124	124	122
0	0	0	0	0	0	0	0

Example B-Summary

GROUP	SEX	DOB	BLOCK	INITWT	WT0	INITPCV	GRTH_69	GRTH_139
B	F	7/5/1999	2	4/23/1900	139	31.8	-66	-66
C	M	6/28/1999	2	5/1/1900	152	30.6	-143	-212
C	F	6/6/1999	1	3/4/1900	144	28.9	-58	-275
A	F	6/4/1999	1	5/21/1900	180	39.2	64	-18
B	F	6/1/1999	1	6/8/1900	198	36.8	1	10
A	M	7/18/1999	2	4/21/1900	148	33.6	304	191
A	M	7/1/1999	2	4/29/1900	150	32.5	-165	-331
B	M	7/5/1999	2	5/15/1900	172	34.9	118	-39
C	M	6/10/1999	1	5/3/1900	120	25.3	-168	-198
A	F	7/4/1999	2	5/1/1900	156	30.7	275	236
B	F	7/15/1999	2	4/9/1900	118	30	361	271
B	F	6/2/1999	1	5/19/1900	200	36.7	72	-78
C	F	7/2/1999	2	4/23/1900	138	35.6	-214	-166
B	F	7/4/1999	2	5/15/1900	210	36.2	11	97
B	M	6/20/1999	1	4/21/1900	130	33.7	175	92
A	M	7/8/1999	2	5/1/1900	160	36.5	-88	-196
A	F	6/24/1999	1	4/17/1900	147	29.8	-156	-156
C	M	7/1/1999	2	5/5/1900	164	33.3	-183	-183

MPCV49_56

13.4
13.2
15.2
18.6
18.7
20.5
16.1
20.0
14.3
20.9
20.9
18.1
15.0
20.1
20.2
15.9
12.0
12.6