A. Welsh STAT1004 (1995)

5. Elements of Experimental Design

5.1 Designing an experiment

Statistics is concerned not only with analysing data but also with designing procedures for collecting data. It is obviously important to collect data that can be used to address the question(s) of interest and desirable to collect data which will enable this to be done as simply and directly as possible. It is vital to emphasise that no amount of fancy analysis can extract information from a badly designed and badly executed study. It is therefore imperative that the data be collected properly in the first place.

Experiments are a very important method of data collection. An experiment can be thought of as an organised and planned enquiry carried out under at least partially controlled conditions. Experiments may be concerned with the determination of absolute values (such as the speed of light, the gravitational constant etc) or with making comparisons (comparing the effect of exercise regimes on the age at which infants walk alone, comparing the attractiveness of colours to beetles etc) in which case it is described as a comparative experiment. The experiments we have considered so far are all comparative experiments and in the discussion that follows we will concern ourselves exclusively with such experiments.

The first requirement of a good experiment is that it should address questions which are of substantive importance. The formulation of questions of interest is a creative and non–statistical part of science. In general, statisticians tend not to contribute to the formulation of the basic questions of interest unless they are immersed in the field to the extent of being able to carry out their own innovative research in the field.

Once a question of interest has been formulated, the next step is to create a detailed written statement of the problem. The thinking that goes into providing such a statement resolves a number of design questions. In addition to stating the objectives of the experiment (which as a matter of course should specify the underlying population to which the conclusions are intended to apply), the statement needs to specify precisely how these objectives are to be achieved.

We need to specify a **response** variable and how it will be measured. Clearly, the response variable needs to be measurable and preferably to a high degree of accuracy. Moreover, it is desirable to use the same measurement procedure to make all measurements within an experiment (unless of course, a number of methods of measurement are being compared).

We need to specify the **factors** of interest that are to be varied in the experiment as well as any other secondary factors which might affect the response. We need to determine how the factors of interest are to be varied which means that we need to specify which **levels** of the factors we are interested in. The choice of factor levels has a substantial effect on the information provided by an experiment. If the levels are fixed then the inferences apply only to the fixed levels whereas if the levels are sampled from a population of levels, the inferences are then meant to apply to the larger population. In the experiments we have considered so far, the factor levels are of specific interest (the three exercise regimes; the four colours; the amounts of whey and supplement; the tree species and the size of flakes; the quantity of bacteria, temperature and amount of oxygen) and are therefore fixed. The choice of the fixed factor levels is obviously very important and needs careful consideration. Random levels need to be sampled from the population of interest and the details of this procedure need to be specified.

When we run an experiment, we often combine factor levels. If the factor levels are not combined but instead we run a series of experiments in which all but one factor are held constant from experiment to experiment, we have no information on the joint effects of the factors. On the other hand, if we vary the factors together we can obtain information on both their joint and their separate effects, and possibly with greater precision than we can when we study them separately. We call a

factor level or combination of factor levels whose effect is to be measured and compared to that of other factor levels or combinations of factor levels a **treatment**. Thus a treatment may be a level of a single factor or it may consist of combinations of factors at various levels. For example, in the single factor experiments we considered, the treatments are levels of the factor (the four exercise regimes; the four colours) but in the multi–factor experiments, the treatments are the combined levels of factors (amounts of whey and supplement; tree species and different size flakes; the quantity of bacteria, the temperature and the amount of oxygen). In some experiments (such as the walking experiment) we have a null treatment or **control** (representing an absence of treatment) against which we want to compare other treatments.

The basic component of an experiment is the **experimental unit** which is the unit to which one application of a treatment is made. In our examples, the experimental units are a child, a board, a batch of pancake mix, a manufactured board of bonded wood, and a potato respectively. Clearly, the possibilities are endless. Before we can design an experiment, the experimental units and any restrictions on availablity and cost need to be specified.

The units on which a response is actually measured are called the **sampling units**. The sampling units will often be the same as the experimental units but they can also be a part thereof. For example, in the walking experiment, a child is both the the experimental and the sampling unit but in the pancake experiment, the experimental unit is the batch of pancake mixture but the sampling units are the pancakes. (The design was to mix 24 batches of mixture and cook one pancake from each batch.) Particularly when the measurement process is destructive, we usually destroy only a small part of the experimental unit rather than the entire unit and in this situation, the sampling unit differs from the experimental unit. We need to specify the sampling units and, if they differ from the experimental units, the way in which they will be selected.

The basic specifications for the experiments we have considered are given below:

Question of interest: Does vitamin E defficiency affect the storage of Vitamin A?
Response: The amount of vitamin A stored in the liver
Factors (Levels): Diet (normal, deficient)
Treatments: The levels of diet
Experimental Units: A rat
Sampling Units: A rat

Question of interest: Does exercise affect the age of infants walking alone?
Response: The age of first walking alone (months)
Factors (Levels): Exercise regimes (active, passive, control)
Treatments: The levels of exercise
Experimental Units: 1–week old male infants from upper middle class families
Sampling Units: 1–week old male infants from upper middle class families

Question of interest: Which colours are most attractive to cereal leaf beetles?
Response: The number of beetles attracted to a coloured board
Factors (Levels): Colour (blue, green, white, yellow)
Treatments: The levels of colour
Experimental Units: A board
Sampling Units: A board

Question of interest: Does the amount of whey and/or the use of a supplement affect the quality of pancakes?
Response: An expert's rating of pancakes
Factors (Levels): Whey (0%, 10%, 20%, 30%), supplement (used, not used)
Treatments: The eight possible combined levels of whey and supplement
Experimental Units: A batch of pancake mixture
Sampling Units: A pancake

Question of interest: Does the species of tree and the size of flake used in the manufacture of bonded wood affect the strength of the product?
Response: Themodulus of elasticity
Factors (Levels): Species (aspen, birch, maple), flake size (small, large)
Treatments: The six possible combined levels of species and flake size
Experimental Units: A board of bonded wood
Sampling Units: A board of bonded wood

Question of interest: Do the amount of bacteria, storage temperature and oxygen affect the extent of potato rot?
Response: The diameter of the rotted portion of a potato
Factors (Levels): Amount of bacteria (low, medium, high), storage temperature (10°C, 16°C), amount of oxygen during storage (2%, 6%, 10%)
Treatments: The 18 possible combined levels of bacteria, temperature and oxygen
Experimental Units: A potato
Sampling Units: A potato

Statistics does not really contribute to the specification of the basic elements of an experiment as we have set them out above. The main statistical contribution is in specifying precisely how the experiment is to be carried out. The most important aspect is determining the assignment of treatments to units but the statistical contribution may also include determining the number of observations, the order in which the observations are to be made and describing a method of analysing the results. The objective of designing an experiment is to ensure that the relevant conclusions can be drawn in a valid way at the end of the experiment. In particular, the requirements of a good design are that

- the treatment comparisons are free from systematic error and are made sufficiently precisely
- the conclusions apply to the intended population
- the experimental arrangement is as simple as possible and
- the uncertainty in the conclusions is assessable.

It is obvious that the effect of treatments on a response can be masked by faulty experimental techniques which introduce bias or unwanted variation. It is primarily the responsibility of the experimenter to refine the experimental technique to secure uniformity in the application of the treatments, to exercise control over the external infuences so that the conditions are as uniform as possible across treatments and to take care to reduce gross errors (mistakes) in the procedure. Moreover, the experimenter needs to be sure that the measurement of the response variable is free of bias and unnecessary variation.

Even if the experimental technique is refined and the experiment is carefully executed, the effect of treatments can be masked by the inherent variability in the experimental units. These inherent variations are often referred to as "experimental error". Effective control and measurement of the inherent variation are important components of experimental design.

The three key elements of statistical design which are used to achieve these objectives are **randomisation, replication and control.** We will discuss each of these in some detail.

5.2 Randomisation

Randomisation is the use of a probablistic mechanism to assign treatments to units (or, equivalently, of units to treatments) or to determine other allocations including the order in which the experiment will be carried out and the locations of the units in space. Suppose we were going to repeat the walking experiment so we have 18 infants (experimental units) to be assigned to one of the three treatments. We decide to make the assignment at random. The simplest way to do this is to take 18 identical slips of paper, write the name or an identifying number of an infant on each slip of paper, mix them thoroughly in a container and then draw them at random from the container. The first six infants drawn are assigned to the first treatment, the next six to the second and the last six to the third. This simple mechanical process can become tedious to implement so we often use a computer to simulate it. That is, we draw a sample of size 18 without replacement from the 18 integers {1,2, ..., 18} to represent the numbered slips of paper drawn from the container and then proceed exactly as above. Of course, in more complicated experiments, the randomisation process can become quite complicated.

It is very useful to able to draw a picture to represent the allocation of units to treatments graphically and to show the physical layout of the experiment. There is no physical layout for the walking experiment but we can still produce a representation of the allocation. If the 18 integers in random order are 9, 15, 13, 18, 2, 10; 16, 8, 11, 6, 7, 17; 3, 1, 4, 14, 5, 12, we obtain

Infant	Treatment
1	C
2	A
3	C
4	C
5	C
6	B
7	B
8	B
9	A
10	A
11	B
12	C
13	A
14	C
15	A
16	B
17	B
18	A

The main benefit of randomisation is that it circumvents (at least in the long run) the difficulties associated with the alternative approaches. A brief discussion of some of these follows:

Systematic allocation: In an experiment exploring the growth of trees receiving one of two nutrient supplements, we plant the trees in the ground in two adjacent rows such that

all the trees receiving the first supplement are in one row and all the others are in the second. The apparent benefit of such an arrangement is that it simplifies subsequent applications of the supplements. However, if we later find out that a stream runs along one side of the experiment, we have to recognise that the experiment is invalidated by the fact that we cannot separate out the effects of the two supplements and the effects of being nearer the stream because these two factors are **confounded**. Of course, if we are aware of the stream, we can come up with another systematic arrangement. However, there is no guarantee that another confounding factor will not be found.

In experiments involving plants grown in pots in glass houses, it has been suggested that the plots be moved around in the glass house during the experiment in a systematic way. This may be effective in reducing variation but it leaves us with no way of estimating the experimental error.

Subjective allocation: In the famous Lanarkshire milk experiment, 5000 children received 3/4 pint of raw milk per day, 5000 recieved 3/4 pint of pasteurised milk per day and 10,000 received no milk. The children within each school were assigned to each group roughly at random but then the teachers were allowed to try to balance the groups by making sure that no group had too many well–fed or under–nourished children. At the end of the experiment, the growth of children in the control group far exceeded that of the children in the other groups. In his critique of the experiment, Student (1931) suggested that the teachers had unconciously ensured that the ill–nourished children were mainly in the groups which received milk.

Self selection: A teacher wanted to show the benefits of learning Latin on English performance. The average score on an english–proficiency test of the seniors who studied Latin was much higher than that of the seniors who did not. The teacher concluded that the study of Latin improves proficiency in English. However, the students themselves chose whether to study Latin or not and by and large the students who are better and more interested in languages probably choose to study Latin. We cannot separate out here the effect of studying Latin and the effect of being smart and interested in languages.

Haphazard allocation: Haphazard allocations are allocations that seem random but are not derived from an explicit randomisation mechanism.

Randomisation prevents extraneous factors (whether we identify them in advance of the experiment or not) and unconscious factors from having a systematic effect on the results and hence prevents systematic bias. Randomisation does not in a single experiment guarantee that there is no favourable or unfavourable allocation but it does guarantee that in a sequence of experiments (with different allocations) there is no systematic favourable or unfavourable allocation. That is, in the long run, there is no confounding with uncontrolled factors.

Experience has shown that conscious and haphazard allocations usually lead to bias in the results. Hence a failure to randomise will often attract suspicion and the use of explicit randomisation ensures that allegations of conscious or unconscious bias can be effectively defended.

The experimenters are not necessarily the only source of unconcious bias in an experiment. In experiments involving human subjects, the mere fact of participating in an experiment can have an effect which biases the results. This is most clearly seen in drug trials in which the control subjects are administered a drug which is indistinguishable from the active drugs except that no active ingredient is present. This drug is called a **placebo** and is given to the control group to ensure that its members do not know that they are not receiving an active drug. Such experiments usually show that the placebo is a surprisingly effective treatment. This **placebo effect** represents the effect on a subject of believing that they are receiving an effective drug. Experiments in which the subjects do

not know which treatment they are receiving are described as **blind experiments.** In a **double blind experiment** neither the subject nor the experimenter know which (if any) treatment is being applied. This is important if subjective assessment (for example clinical progress) is part of the experiment.

In addition to ensuring that there is no systematic bias, randomisation also helps ensure that the assumption of independent errors is reasonable and hence that the estimate of error is meaningful. In this context, systematic allocations are arguably not as bad as subjective ones which can never be made explicit and are therefore not repeatable. Nonetheless, the use of systematic designs requires strong uncheckable assumptions (so that we can estimate the error) so it is sensible to randomise unless there is a very good reason not to.

5.3 Replication

Replication is the application of a treatment more than once in an experiment. For example, in the walking experiment, each treatment was replicated six times, in the potato rot experiment, each treatment was replicated three times etc. Replication is intended to

– provide an estimate of the inherent variation

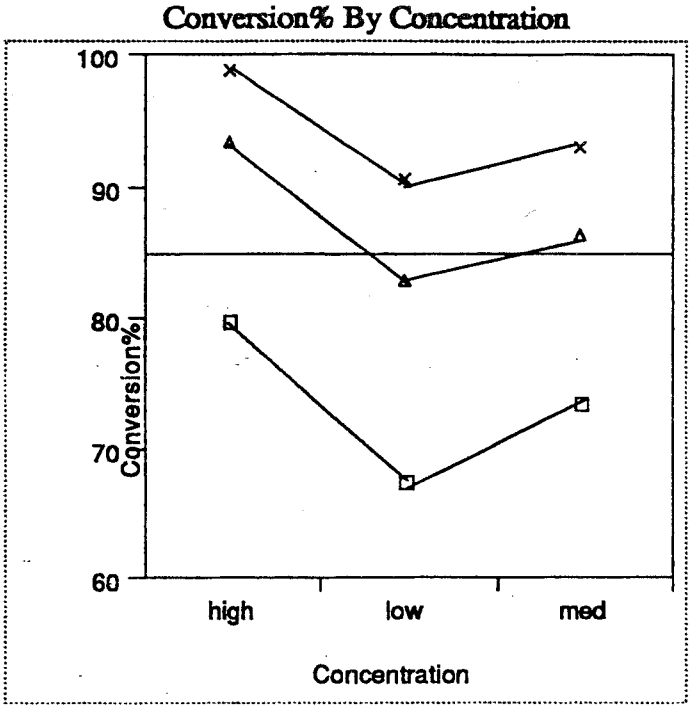
– reduce the standard deviation of the estimates and

– increase the scope of the inference by including more units.

An estimate of the experimental error is needed for inference. If each treatment occurs only once in an experiment, it is not possible to estimate the experimental error unless prior information is available or we make strong assumptions about the data. The difficulty is that there is no way to determine whether observed differences are due to different treatments of different units. That is, the treatments are confounded with the experimental units.

Johnson and Leone reported data from an experiment on the effect of catalyst concentration and temperature on the conversion percentage of the polyesterification process of fatty acids with glycols. The treatments were the concentration of the catalyst (low, medium, high) and the temperature (low, medium, high) so there were $3x3 = 9$ treatment combinations. The treatments are reported as nominal variables. The response is the conversion percentage in the reaction. The question of interest is do the treatments affect the conversion process? The data are available in **Polymer.**

When we examine the data, either directly or on a plot, we notice that there is only one observation at each treatment combination so there is no replication. This happens when due to time or cost it is impractical to replicate the experiment.



The three curves seem to be roughly parallel so there is no evidence of an interaction.

The model with interaction is obtained by representing Y_{ij}, the percentage conversion for the i^{th} concentration at the j^{th} temperature, as

$$Y_{ij} = \alpha + \gamma_i + \delta_j + (\gamma\delta)_{ij} + \varepsilon_{ij}$$

$$i = 1,2,3; j = 1,2,3$$

with ε_{ij} independent N(0, σ^2).

If we try to fit this model, we find that the analysis of variance table is

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	827.42000	103.427	⊃
Error	0	0.00000	⊃	Prob>F
C Total	8	827.42000		⊃

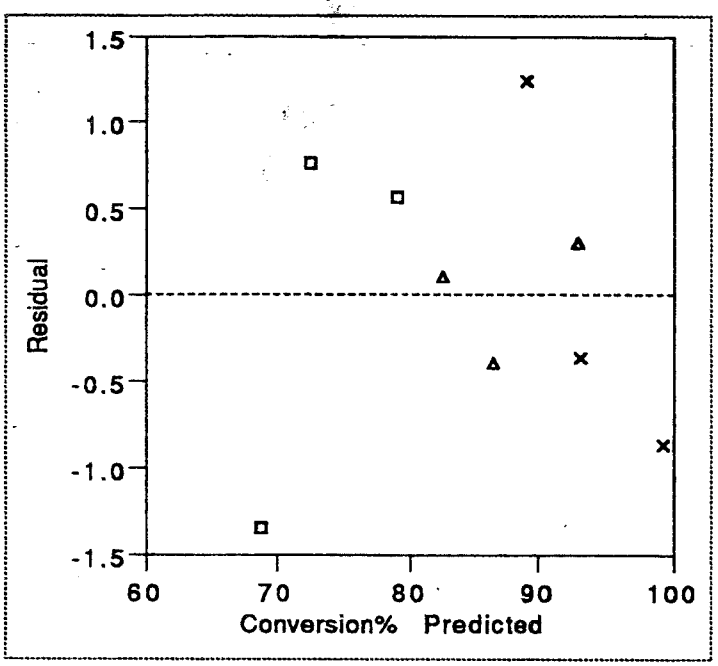
That is, the model fits exactly because there are as many parameters as observations. The exact fit means that there is no estimate of error and hence no standard errors, no test statistics etc.

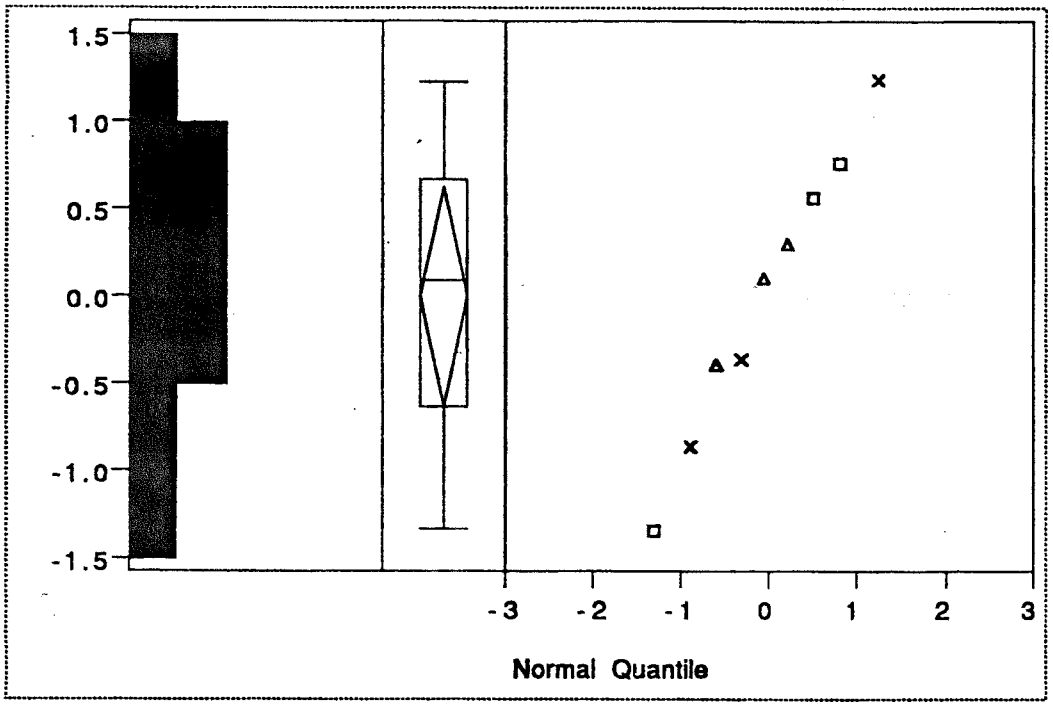
If we design an experiment without replication, we know in advance that we will be unable to estimate both the interaction and the error variability because we cannot separate them out. We say that they are confounded. If there is no interaction (and this is something we have to assume without the benefit of a formal test), then we can estimate the error variability. We have plotted the data and seen that there is little evidence of interaction so we will proceed on this basis. If however, there appears to be an interaction then alternative approaches may be necessary. In particular, we can try to transform the data to reduce the interaction. The difficulty though is in separating the interaction from the variability when we lack an estimate of variability.

The full model in this case is

$$Y_{ij} = \alpha + \gamma_i + \delta_j + \varepsilon_{ij}$$
$$i = 1,2,3; j = 1,2,3$$
with ε_{ij} independent N(0, σ^2).

This model is readily fitted in JMP. It has only 5 effective conditional mean parameters so there are $9 - 5 = 4$ degrees of freedom available for estimating σ^2. We can examine the diagnostics in the usual way.





The model fit seems to be adequate.

The effect test table shows that both treatments are significant

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Concentration	2	2	162.26000	60.6202	0.0010

Temp	2	2	659.80667	246.5031	0.0001

The yield is most improved in this example by taking the highest concentration at the highest catalyst.

are
the

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	822.06667	205.517	153.5616
Error	4	5.35333	1.338	Prob>F
C Total	8	827.42000		0.0001

Root Mean Square Error 1.156864

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	84.966667	0.38562	220.34	0.0000
Concentr[high–med]	5.5666667	0.54535	10.21	0.0005
Concentr[low–med]	–4.733333	0.54535	–8.68	0.0010
Temp[high–med]	9.0333333	0.54535	16.56	0.0001
Temp[low–med]	–11.5	0.54535	–21.09	0.0000

The standard deviation of an estimate decreases as the square root of the number of replications so increasing the number of replications can decrease the standard error of the estimates. However, replication introduces more experimental units and hence can increase the inherent variation which tends to counteract the expected decrease in the standard error.

Replication can be used to intentionally include more experimental units and to increase the inherent variation because this extends the scope of the inference. Essentially, the units constitute a larger sample from the underlying population.

A design in which all the treatment differences have equal standard errors is said to be **balanced**. In simple situations, this is equivalent to having equal replication of each treatment.

The choice of the number of replicates depends on providing sufficient information to estimate the experimental error (with between say 12 and 20 degrees of freedom), the availability of experimental units, cost and time available for the experiment. If we have information about the variability, we can do a formal calculation to ensure that we achieve a specified precision. Consider a simple two sample comparison. It is sensible to use balanced designs whenever possible so suppose that we intend to draw $X_1,...,X_r$ and $Y_1,...,Y_r$ independently from $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$ populations respectively. Suppose we want to estimate $\mu_X - \mu_Y$ to within $\pm\Delta$. A 95% confidence interval for $\mu_X - \mu_Y$ is

$$\bar{X} - \bar{Y} \pm 1.96\sigma\sqrt{2/r}$$

so we require

$$1.96\sigma\sqrt{2/r} < \Delta$$

or

$$r > 2(1.96)^2\sigma^2/\Delta^2.$$

The usefulness of this simple formula is affected by the fact that we often do not know σ^2. Standard practice is to substitute an estimate of σ^2 based on preliminary studies, past experience or "expert guesses". In these situations, it is often acceptable to use an upper bound for σ^2 in the formula.

It is important that replication be genuine. **Pseudo-replication**, apparent but not genuine replication, is a common problem in experimentation. It generally arises through confusion over the distinction between experimental units and sampling units. Genuine replication is replication at the experimental unit level; pseudo–replication is the attempt to pass off as replication the use of a large number of sampling units. Suppose for example, the experimental units in an experiment are trees and the sampling units are leaves. An extreme case of pseudo–replication arises when we make observations on a very large number of leaves from a single tree. This may tell us a great deal about this tree but nothing about tree to tree variation which, given the definition of trees as the experimental units, is what we are interested in. Similarly, if the experimental unit is a field of trees, then sampling a large number of trees from a single field is pseudo–replication. Another example would be if, in the pancake experiment, we made 4 pancakes from each of 2 batches of mixture rather than 1 pancake from each of 8 batches.

BACI (Before and After Controlled Impact) trials are sometimes used in ecology. Here two areas which are as similar as possible are chosen and observed. A treatment is applied to one area and both areas are then observed again. Usually a large number of observations are made within each area. However, the area is the experimental unit and it is clear that there are only two experimental units so there is no effective replication.

A more subtle form of pseudo–replication can arise from the allocation of units to treatments. If for example in an experiment in which the experimental units are fields of trees, we adopt a systematic arrangement in which the fields receiving the same treatment are contiguous, then the actual experimental units are the contiguous sets of fields rather than the separate fields and working with the original fields can be viewed as pseudo–replication.

5.4 Control

Control (or reduction) of the inherent variation in an experiment is an important aspect of design. One way to achieve control is to select homogenous experimental units. However, this may make them less representative of the underlying population of interest and so diminish the generality of the conclusions. The need to control the inherent variability arises when we are interested in heterogeneous populations.

The inherent variation in an experiment can often be reduced by arranging the experimental units into **blocks** of closely related or similar units and then applying the treatments to the units in each block so that every treatment occurs in every block. The basic idea is to try to choose the blocks so that the within block variation is much smaller than the between block variation. In the analysis of the results of the experiment, we then remove the between block variation from the total variation and effectively use the smaller within block variation as the experimental error.

Suppose that we have two treatments to compare. One possibility is to simply assign the units at random to one of the two treatments. In this case, the model associated with the design is that Y_{ij}, the response for the i^{th} subject receiving the j^{th} treatment, is

$$Y_{ij} = \alpha + \delta_j + \varepsilon_{ij}$$
$$\text{with } \delta_1 + \delta_2 = 0 \text{ and } \varepsilon_{ij} \text{ independent } N(0, \sigma^2).$$

Here the inherent variability is represented by the error variance σ^2 which is the same as the variance var(Y_{ij}) of an individual response Y_{ij}. On the other hand, if we match the subjects to create

pairs (so that the two subjects in a pair are as similar as possible) and then assign one subject from each pair at random to each treatment, we have a **paired design.** The model can now be written with a pair effect β_i as

$$Y_{ij} = \alpha + \delta_j + \beta_i + \varepsilon_{ij}$$

with $\delta_1 + \delta_2 = 0$, β_i independent $N(0, \sigma_\beta^2)$, the β_i are independent of ε_{ij} which are independent $N(0, \sigma_\varepsilon^2)$.

Notice that we have written the pair effect as a random variable so we think of the pairs as being a random sample from a population of pairs for which the pair effects β_i are normally distributed with mean zero and variance σ_β^2. Although there are a large number of pairs in the experiment, the random effects β_i introduce only a single additional parameter σ_β^2 into the model. (We can think of avoiding fitting a large number of parameters by treating the pair effects as random and modelling their distribution with only a single parameter.) The parameter σ_β^2 represents the between pair variation while σ_ε^2 represents the within pair variation. Under the assumptions of the model, we can show that the variance of an individual response Y_{ij} is

$$\text{var}(Y_{ij}) = \sigma_\beta^2 + \sigma_\varepsilon^2 .$$

Removing the between pair variance σ_β^2 enables us to use σ_ε^2 rather than the total variance $\text{var}(Y_{ij}) = \sigma_\beta^2 + \sigma_\varepsilon^2$ to represent the inherent variation. If the within pair variation σ_ε^2 is much smaller than the between pair variation σ_β^2, then the inherent variation represented by σ_ε^2 is much smaller than the total variance and the pairing has reduced the variation.

An important property of the model with a random pair effect is that the observations are not independent. Pairs are independent but the two observations within a pair are dependent because

$$\text{Cov}(Y_{i1}, Y_{i2}) = \text{Cov}(\alpha + \delta_1 + \beta_i + \varepsilon_{i1}, \alpha + \delta_2 + \beta_i + \varepsilon_{i2})$$

$$= \sigma_\beta^2 .$$

Obviously, our analysis needs to take this dependence into account.

A second important point about the model is that the model is saturated (i.e. there is no replication) if we allow an interaction between the pair effects β_i and the treatment effects. It is standard practice to assume that the random pair effects β_i do not interact with the treatment effects.

Consider an experiment carried out to determine whether epinephrine elevates plasma cholesterol levels in human subjects. Twelve adult males were selected and given both a placebo and epinephrine in random order. Blood samples were taken after injection of the placebo and again after injection of the epinephrine. There is a natural paired structure in this experiment because two measurements are made on each subject. The data are available in **Cholesterol.**

We can analyse data from a paired experiment in two different ways. A simple analysis may be based on the fact that the paired differences

$$Y_{i1} - Y_{i2} = (\alpha + \delta_1 + \beta_i + \varepsilon_{i1}) - (\alpha + \delta_2 + \beta_i + \varepsilon_{i2})$$

$$= \delta_1 - \delta_2 + (\varepsilon_{i1} - \varepsilon_{i2})$$

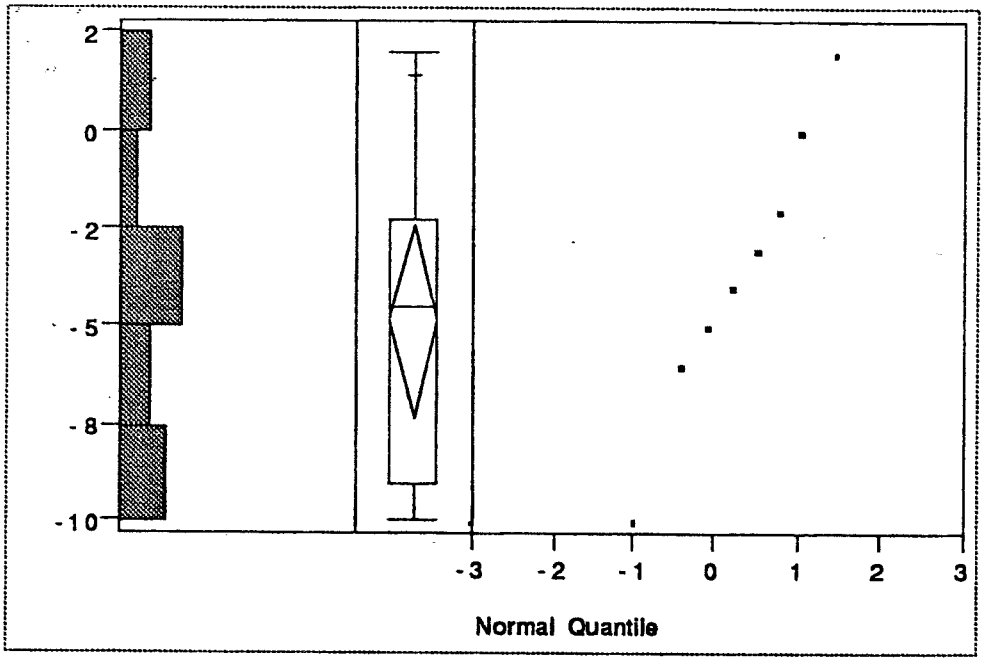
$$= 2\delta_1 + (\varepsilon_{i1} - \varepsilon_{i2})$$

$$= \mu + \omega_i$$

say, with ω_i independent N(0, τ^2), where $\tau^2 = 2\sigma_\varepsilon^2$.

The act of differencing has eliminated the pair effect so the set of paired differences $Y_{i1} - Y_{i2}$ are independent and identically distributed observations from a normal distribution and we are interested in making inferences about their underlying mean μ which corresponds to precisely the difference in the effects of the two treatments.

This analysis is readily carried out in JMP. The quantile plot for the differences is roughly linear indicating that we can treat their distribution as normal.

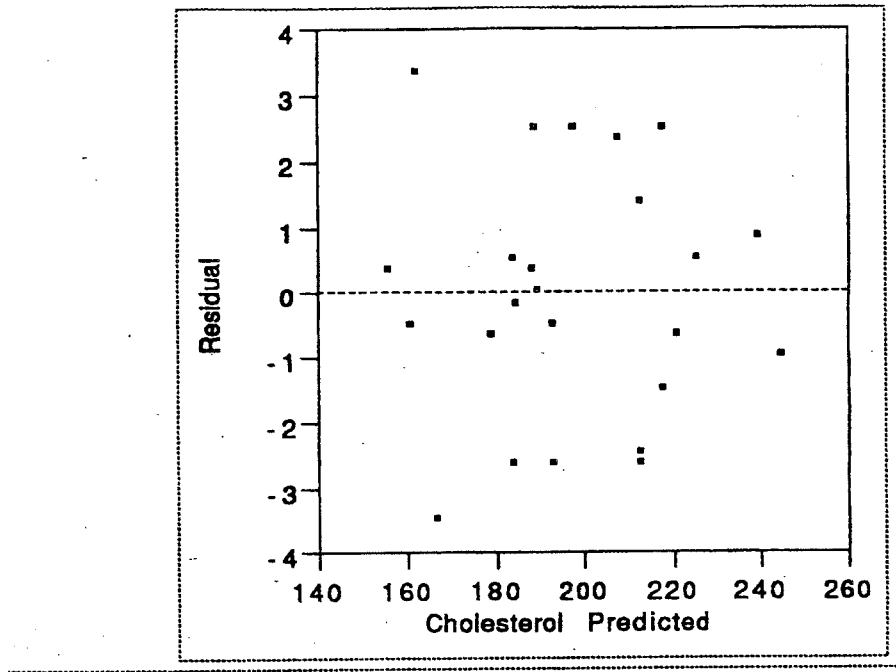


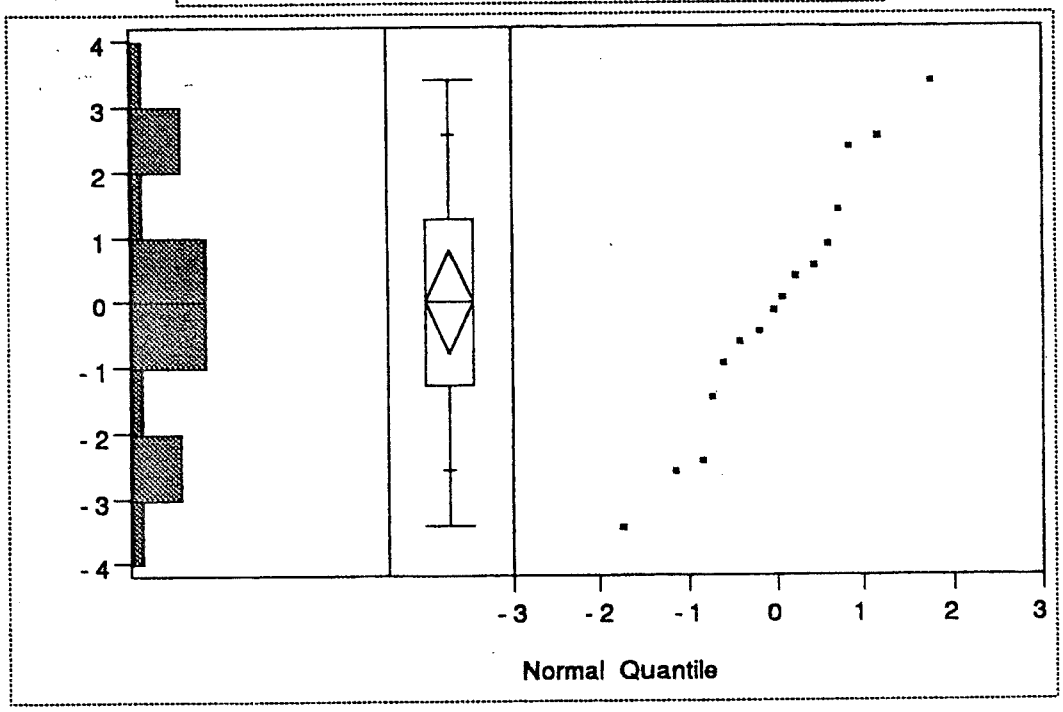
Moments

Mean	–4.83333
Std Dev	3.88080
Std Err Mean	1.12029
upper 95% Mean	–2.36758
lower 95% Mean	–7.29909
N	12.00000
Sum Wgts	12.00000

A 95% confidence interval for the mean μ is given by $-4.833 \pm 1.120 \times t_{11}(0.975)$. i.e. (–7.3, –2.4).

A second approach is to actually fit the model with the random pair effect. The residual plot looks acceptable and the normal quantile plot shows that the residuals have a normal distribution.





To obtain a confidence interval for the difference in the means, we can refer to the table of means under the effect test table for the treatment variable.

Least Squares Means

Level	Least Sq Mean	Std Error	Mean
Epinephrin	199.5000000	0.7921649150	199.500
Placebo	194.6666667	0.7921649150	194.667

A 95% confidence interval for the difference in the means is

$$199.5 - 194.667 \pm t_{11}(0.975)\sqrt{0.7921^2 + 0.7921^2}$$
$$= 4.833 \pm 2.2 \text{x} 1.120$$
$$= 4.83 \pm 2.464$$

or (2.4, 7.3).

Alternatively, we can refer to the parameter estimates table. The subject effects are really estimates of random variables rather than parameters so they really should not appear in the table. The relevant row is that for the treatment parameter.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	197.08333	0.56015	351.84	0.0000
Treatmen[Epineph-Placebo]	2.4166667	0.56015	4.31	0.0012
Subject[1-12]	-16.08333	1.85779	-8.66	0.0000
Subject[2-12]	44.416667	1.85779	23.91	0.0000
Subject[3-12]	12.916667	1.85779	6.95	0.0000
Subject[4-12]	-10.58333	1.85779	-5.70	0.0001
Subject[5-12]	-2.083333	1.85779	-1.12	0.2860
Subject[6-12]	-11.08333	1.85779	-5.97	0.0001
Subject[7-12]	-39.08333	1.85779	-21.04	0.0000
Subject[8-12]	25.916667	1.85779	13.95	0.0000
Subject[9-12]	17.916667	1.85779	9.64	0.0000
Subject[10-12]	-33.08333	1.85779	-17.81	0.0000
Subject[11-12]	-7.083333	1.85779	-3.81	0.0029

Recall that the difference between the two means in this parametrisation is $2\partial_1$ so a 95% confidence interval for $2\partial_1$ is

$$2.41 \times 2 \pm t_{11}(0.975)\ 2 \times 0.56$$

which is the same as before.

Finally, we get estimates of the variance components.

Variance Component Estimates

Component	Var Comp Est
Subject	598.3636
Residual	7.530303

We see that between subject variance is much larger than the within subject variance. The pairing has therefore controlled much of the variability.

An advantage of fitting the random pair effect explicitly is that the approach generalises in a straightforward manner to problems with more than two units in each block.

Consider now an experiment described in the Minitab Student Handbook (p200) to explore the effects of oven position on the baking of meatloaves. A batch of 8 loaves was mixed. The loaves were then randomly assigned to one of 8 positions in the oven. The loaves were baked and the driploss (amount of liquid which dripped out of the loaf divided by the original weight of the loaf) recorded. Another 2 batches of loaves were prepared and tested in the same manner. The loaves are the experimental units. Each batch is a block and we randomise the units within each block to treatments which are the positions in the oven. Notice that the randomisation is carried out within blocks and is therefore a form of restricted randomisation. This relationship between randomisation and control makes it difficult to discuss them separately: we try to recognise possible sources of variation and control for them but where we cannot control, we randomise. The data are available in **Meatloaf**.

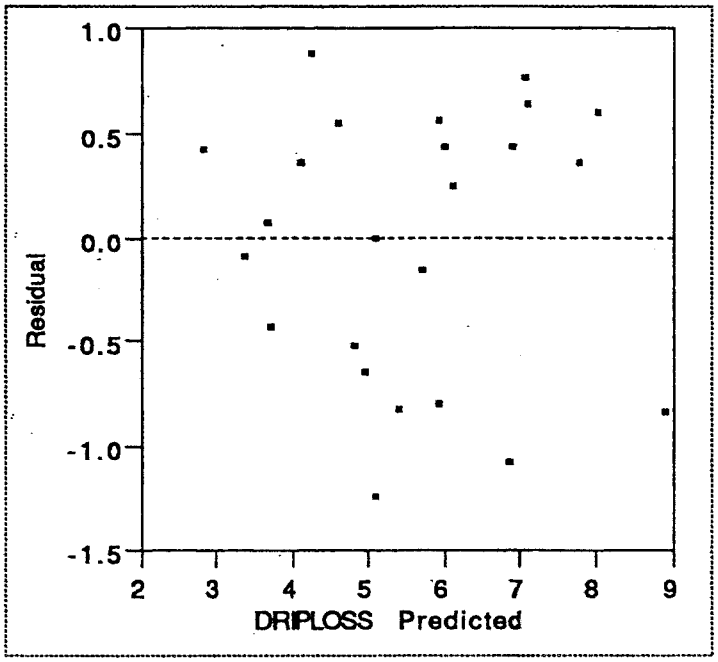
A plausible model for the driploss is

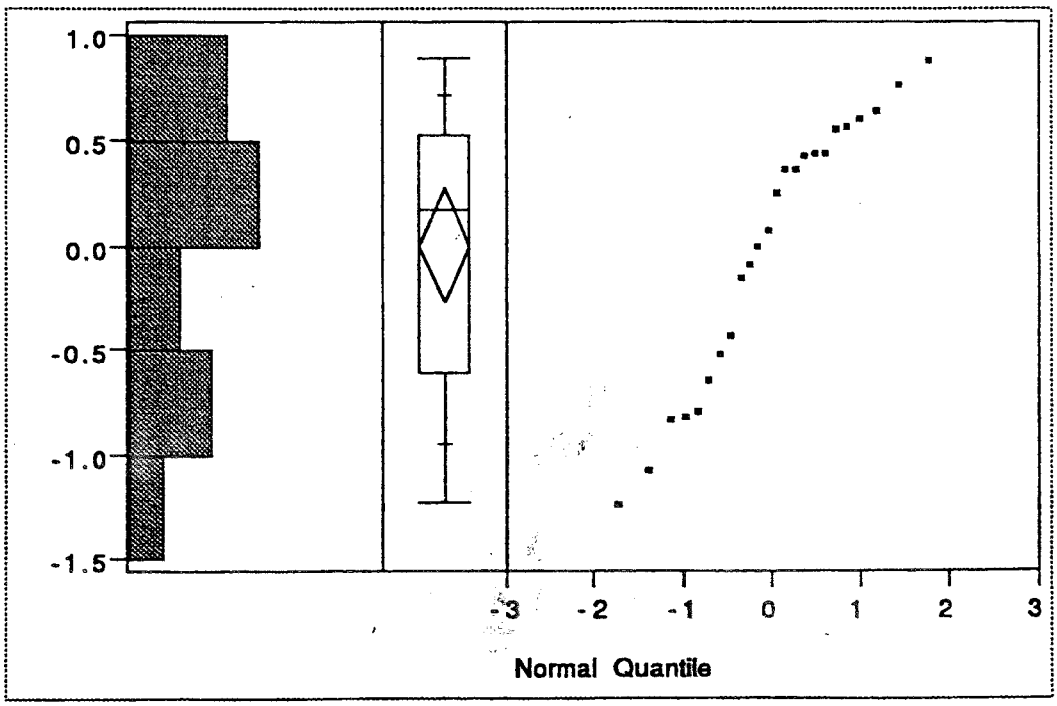
$$Y_{ij} = \alpha + \delta_j + \beta_i + \varepsilon_{ij}$$

with β_i independent $N(0, \sigma_\beta^2)$ independent of ε_{ij} independent $N(0, \sigma_\varepsilon^2)$.

Here α is the common oven effect and δ_j is the departure from the common effect due to being in the j^{th} position. The random variables β_i represent the random block effect so σ_β^2 is the between block variance.

The model is straightforward to fit.





The effect test table shows that the effects of batch and position are both significant.

Tests wrt Random Effects

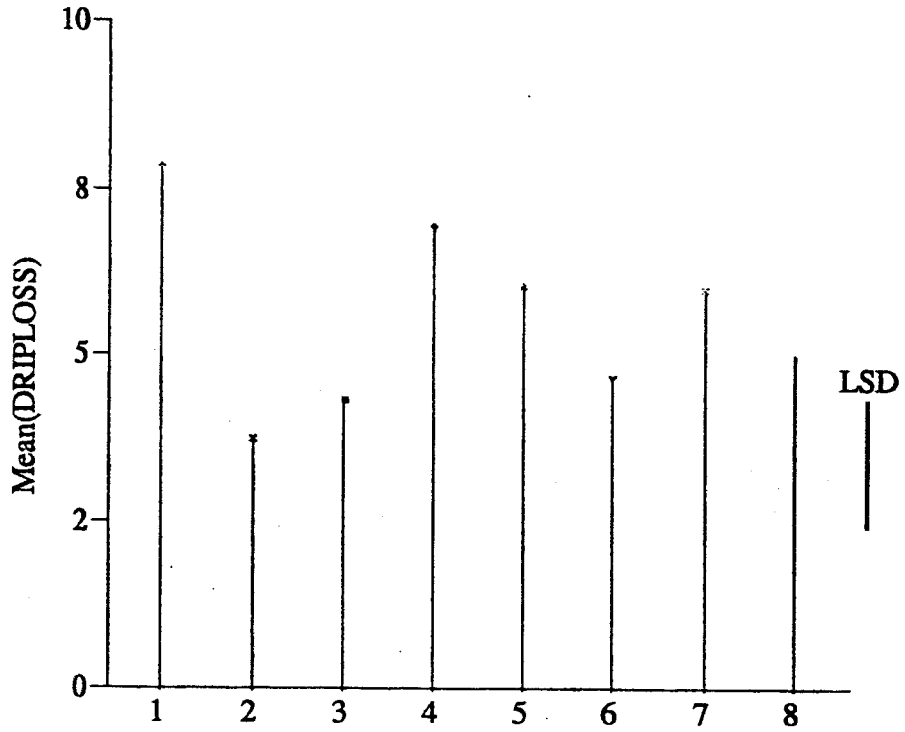
Source	SS	MS Num	DF Num	F Ratio	Prob>F
BATCH	16.2593	8.12965	2	12.2516	0.0008
POSITION	40.3957	5.77081	7	8.6968	0.0003

We can therefore compare the effects of the different positions using either the parameter estimates or the least squares means for position.

Least Squares Means

Level	Least Sq Mean	Std Error	Mean
1	7.833333333	0.4703044860	7.83333
2	3.740000000	0.4703044860	3.74000
3	4.316666667	0.4703044860	4.31667
4	6.943333333	0.4703044860	6.94333
5	6.016666667	0.4703044860	6.01667
6	4.650000000	0.4703044860	4.65000
7	6.000000000	0.4703044860	6.00000
8	5.016666667	0.4703044860	5.01667

A barchart of these means facilitates comparison. We see that the greatest driploss is incurred at positions 1 and 4 and the least at positions 2 and 3.



Finally, the variance component estimates show the effectiveness of the blocking in controlling the variation.

Variance Component Estimates

Component	Var Comp Est
BATCH	0.933262
Residual	0.663559

As a final example on blocking, consider the data from a carrot experiment described in Mead, Curnow and Hasted (1993, p 108). The objective was to investigate the effect on yield of sowing rate (at four levels denoted A, B, C, D in increasing order) for two stocks of seed. The treatments were the 2x4 = 8 crossed levels of stock and sowing rate. These eight treatments were applied to experimental units (presumably fields) arranged in three blocks (possibly farms). The data is available in **Carrot** and has the structure shown below.

Stock	Sowing Rate	Block		
		I	II	III
T	A	4.20	4.94	4.45
T	B	4.36	3.50	4.17
T	C	5.40	4.55	5.75
T	D	5.15	4.40	3.90
H	A	2.82	3.14	3.80
H	B	3.74	4.43	2.92
H	C	4.82	3.90	4.50
H	D	4.57	5.32	4.35

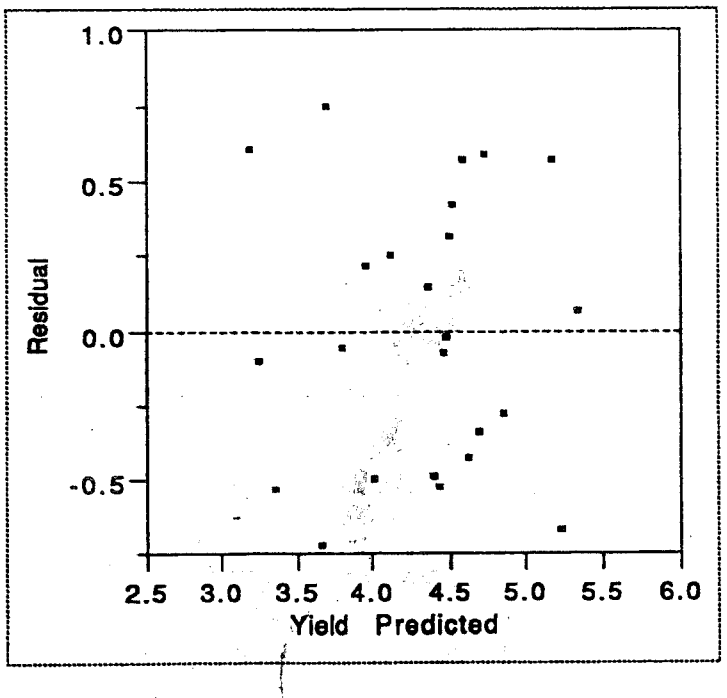
Note that this table does not represent the physical layout of the experiment. (It would be an unsuitable physical layout because the units have not been randomised within the blocks.)

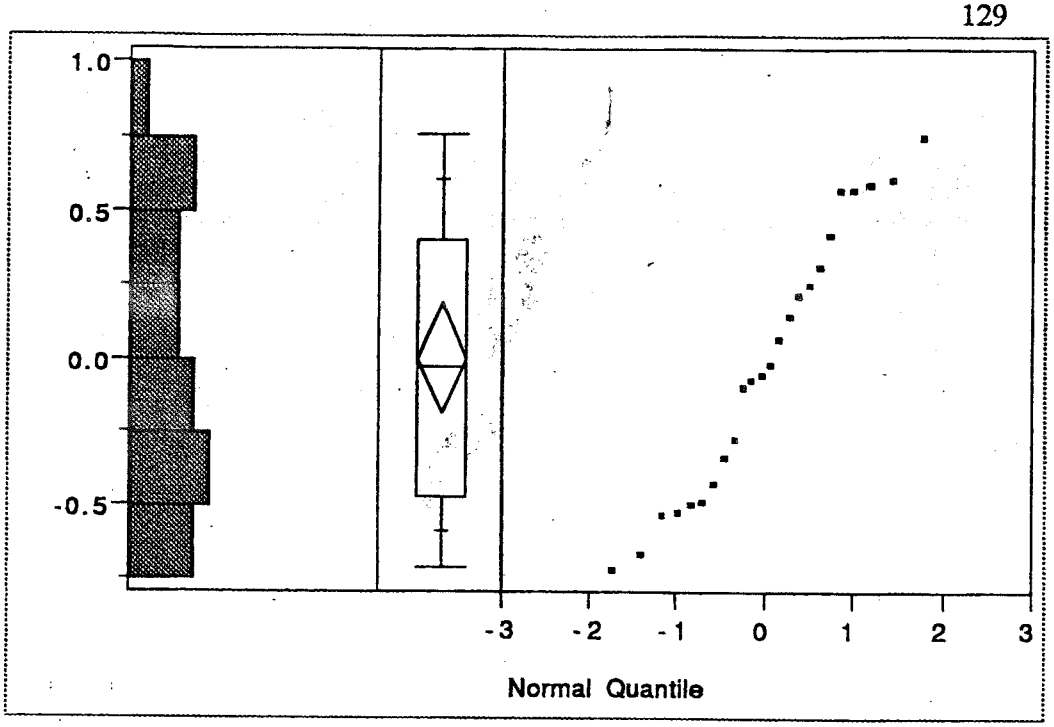
On an appropriate scale, a model for Y_{ijk} the yield of the i^{th} stock with the j^{th} sowing rate in the k^{th} block is

$$Y_{ijk} = \alpha + \delta_j + \theta_i + (\delta\theta)_{ij} + \beta_k + \varepsilon_{ijk}$$

with β_k independent $N(0, \sigma_\beta^2)$ independent of ε_{ijk} independent $N(0, \sigma_\varepsilon^2)$.

We can fit this model in JMP and examine the diagnostics. These seem reasonable so we conclude that the model is adequate.





If we examine the table of tests, we see that the interaction between stock and sowing rate is not significant. The block effect is also not significant but we do not need to react to this other than to note that it indicates that the blocking was not particularly effective.

Tests wrt Random Effects

Source	SS	MS Num	DF Num	F Ratio	Prob>F
Stock	1.73882	1.73882	1	5.1339	0.0398
Sowing Rate	4.41463	1.47154	3	4.3448	0.0232
Stock*Sowing R	1.98235	0.66078	3	1.9510	0.1678
Blocks	0.0991	0.04955	2	0.1463	0.8652

The fact that the interaction is not significant means that an additive model fits these data. Thus we can use the tables of stock and sowing rate means to summarise the results.

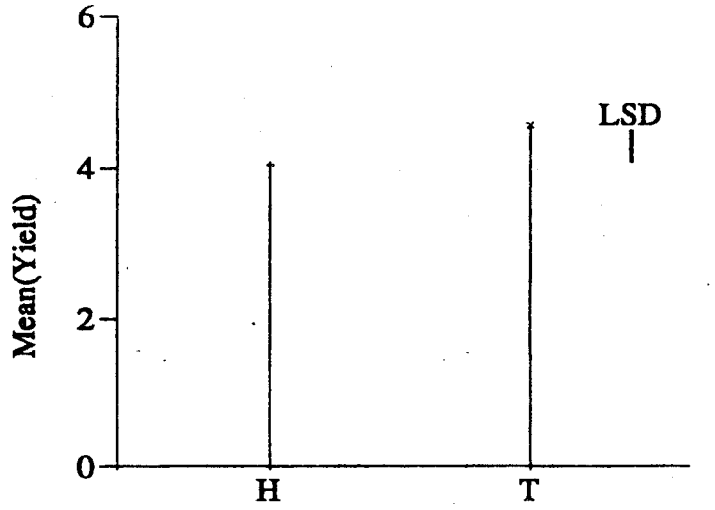
Least Squares Means

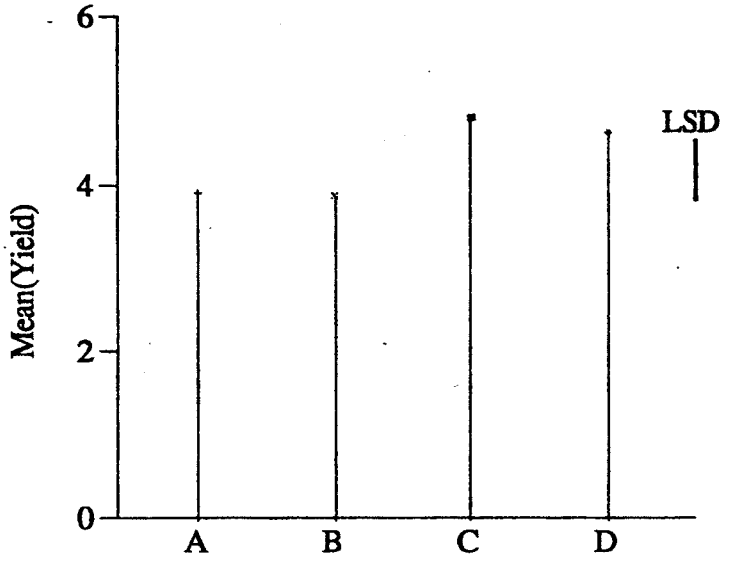
Level	Least Sq Mean	Std Error	Mean
H	4.025833333	0.1680012046	4.02583
T	4.564166667	0.1680012046	4.56417

Least Squares Means

Level	Least Sq Mean	Std Error	Mean
A	3.891666667	0.2375895821	3.89167
B	3.853333333	0.2375895821	3.85333
C	4.820000000	0.2375895821	4.82000
D	4.615000000	0.2375895821	4.61500

The standard error for a pairwise difference in means is $\sqrt{0.168^2+0.168^2} = 0.24$ for the stock and $\sqrt{0.238^2+0.238^2} = 0.34$ for the sowing rate. Thus differences of less than $t_{14}(0.975)*0.34 = 2.14*0.34 = 0.73$ in the means for sowing rate are not significant. We see that the stock T gives a higher mean yield than H and that sowing rate C also gives the highest mean yield though it is not significantly different from D. Presumably higher sowing rates are more expensive so the use of stock T at sowing rate C is recommended.





Finally, the variance component estimates are available.

Variance Component Estimates

Component	Var Comp Est
Blocks	-0.03614
Residual	0.338693

The negative variance is a result of the method of estimation and suggests that the between block variability is small, as we have already noted. In this situation, it is common to truncate the estimate to zero. It is worth noting that the fitting method used by JMP to estimate the variances of random effects can produce this kind of anomaly and that it does not work well when the experiment is unbalanced. In this case, a statistician should be consulted.

When we have a large number of treatments, it may not be feasible to include every treatment in every block. In this case, we can still arrange similar units in small groups which do not receive all treatments. Such groups are called incomplete blocks. While this technique reduces the inherent variability, since not all treatments appear in each block, some treatment comparisons are impossible. Essentially if we need to compare treatments from different blocks, we cannot distinguish between the effect of the different treatments and the effect of the different blocks so these two effects are confounded. In such cases, it is sensible to allocate treatments so as to confound only comparisons of secondary interest.

When the variation between experimental units is due to measurable **covariates** which are not sufficiently controllable to be a basis for blocking, regression techniques can be used in the analysis of the data to adjust for these covariates. It is important that the covariates have an effect on the response because otherwise the adjustment will be useless. It is also vital that the covariates be measured before the application of the treatments or be known to be unaffected by the treatments because otherwise the adjustment may be detrimental. From the practical point of view, it makes sense at the design stage to identify and measure covariates with a view to subsequent adjustment.

Both blocking and covariate adjustment can be used in a single experiment to control the inherent variation. However, the issues raised by this are well beyond the level of this course.

The final stage of the design phase is to derive the model which will be used to describe the experiment and to specify the statistical analysis to be carried out in terms of the model. For example, hypotheses to be tested, comparisons of parameters and the confidence intervals required should be explicitly stated. Once this has been done, the complete advance description of the analysis can be written into a **protocol**.

5.5 Limitations on the conclusions

One reason for experimentation is to try to establish **causal relationships**. That is, to establish that a treatment causes the response to change. The principle on which this is based is entirely straightforward:

> If everything else is held constant so that the only difference between two groups is in the treatment they receive, then differences in the responses of the two groups must be caused by the different treatments.

The statistical analysis is designed to deal with the inherent variability and determine whether there are differences in the responses or not. The major practical problem is of course to try to hold everything else constant.

We have discussed the use of randomisation to prevent (at least in the long run) confounding effects. However, we always need to be aware of the possibility of unfortunate confounding factors which arise from the limitations in our capacity to hold everything else constant when we evaluate an experiment. In "Concepts and Controversies", Moore reports on a psychologist who carried out an experiment to study the effect of propaganda on attitudes. He devised a test of attitudes towards the German government which he administered to a group of American students. He then asked them to read German propaganda for several months and retested them. Unfortunately, the experiment was carried out in 1940 and Germany invaded and conquered France in between the two tests. The effects of the invasion and the effects of the propaganda are clearly confounded. There is nothing that the psychologist could do about the confounding after it had happened or could have done to prevent it.

Although in the ideal experiment we are supposed to identify the population to which our conclusions are intended to apply and then ensure that they do by chosing a simple random sample of experimental units from this population, this is only rarely feasible in practice. Usually, the experiment is carried out simply with the available experimental units. For example, drug trials are usually carried out on the available patients rather than on a random sample of all possible patients, many experiments are carried out on University students (a danger of studying medicine or psychology!) or simply on available experimental animals bred for that purpose. The de facto population in these examples is always much more limited than desired by the experimenter. The burden is on the experimenter to convince us that the conclusions are in fact more widely applicable. These arguments depend more on subject matter knowledge than statistics and depend on the nature

of the experiment. Medical experiments on students may be more generalisable than the reactions of psychology students.

A different limitation on the validity of the conclusions is brought about by lack of realism in the experiment. For example, the strength of poisons is measured in various ways but one assessment is by the toxicity of a standard dose towards mice, rabbits or dogs. Thus we can express the toxicity as the number of animals killed by a standard dose. Now in general, we are interested in the toxicity towards humans. While the measurement in animals is thought to be useful in this regard, funnelweb spider venom is by a strange quirk non–toxic to rabbits and dogs so experiments on these animals with funnelweb spider venom do not generalise to humans.

It is always important to assess whether the experiment was competently implemented or not. In large part, this means finding out exactly what was done rather than what was intended to be done. For example, changes to the protocol in the middle of an experiment may have serious consequences. Even if the protocol is followed, changes during the experiment (such as to the person making the observations) may make a difference. Outright incompetent experimental procedure also undermines the validity of an experiment. For example, in an experiment on the effect of red dye on rats which led to restrictions on its use as a food additive in the USA, some rats were mixed up during the experiment so that it was impossible to tell which treatment they had received and rats which had died during the experiment were left in the cages so that by the end of the experiment they were too badly decomposed to be evaluated. The end result was that the results from only 94 out of the initial 500 rats in the experiment were usable in the analysis.

Finally, it is an unfortunate fact that some experiments are compromised by fraud on the part of the experimenter. This can range from psychics using standard magicians tricks to simulate psychic activities, to medical researchers drawing tumours in black ink on rats, to making up the data. The possibilities are endless and not restricted to any particular field. Unfortunately, genuine fraud (as opposed to incompetence) can be difficult to detect.

5.6 Ethical Issues.

Experiments on human subjects raise a number of serious ethical issues. Perhaps the most controversial experiments on human subjects are randomised clinical trials which are used to ensure that treatments are effective or better than the existing treatments and don't have serious side effects. In such trials, subjects who meet certain eligibility criteria are assigned to a treatment at random and, whenever possible, the experiment is made double blind to reduce the possibility of confounding subjective effects.

The first ethical principle guiding clinical trials is that treatments which are known to be harmful should not be used. This is consistent with medical ethics which require doctors to "do no harm" to patients. The second principle is that the informed consent of the participants be obtained. This entails first of all informing the participants about the study, the nature of the study and all possible risks and benefits before they are asked to consent in writing to be randomised to a treatment. (Patients who refuse to participate in the study are simply given the standard treatment.) The practice of obtaining informed consent arose in the Nuremburg Code and was adopted by the World Health Organisation in the Declaration of Helsinki (1964). Its origins indicate clearly the kind of concerns the practice is meant to address.

The requirement that doctors "do no harm" implies that the patients in the study need to be selected so that patients in categories for which treatments are known to be harmful or at least not beneficial are excluded from the study. For example, pupils who are allergic to milk should not be included in studies such as the Lanarckshire Milk Experiment.

The requirement that doctors "do no harm" can also be interpreted to mean that they should prescribe the best therapy for a patient because prescribing an inferior therapy can do less good (more harm) than the best therapy. It is clear that in a randomised clinical trial, many patients will not receive the best therapy and therefore will be done harm. However, we do not, in advance, know which group will be done harm. Indeed, the goal of the trial is to determine the answer to this question and there is no other way of discovering which treatment is superior. It is not a solution to abandon clinical trials because refusing to experiment could leave all present patients with an inferior treatment and deprive all future patients from access to a superior treatment and thereby do considerable harm. By and large, trials are carried out when there has been a steady accumulation of anecdotal and other evidence in favour of different treatments and there is a need to resolve the debate so that all future patients can receive the best treatment.

Although at the beginning of a trial, we do not know for certain which treatment is best, it may become clear during the course of a trial that one treatment is superior or that one is harmful. The problem then is whether to complete the study to the predetermined sample sizes or to stop the study early. Stopping the study early can complicate the analysis and make the conclusions less clear cut but has to be balanced against the potential benefits or harm to the remaining patients in or entering the study.

Informed consent can be difficult to achieve. The controversy over the RU486 abortion pill trial in Victoria, while confused by strong beliefs about the propriety of abortion and the type of research in which some of the participants had earlier been involved, was basically about the informed consent. The critics argued that the possible risks of taking the pill were not presented in writing on the informed consent form. The defenders of the trial argued that the informed consent form was only presented after several separate counselling sessions in which the possible risks were discussed in detail. The current status is that the trial has been suspended pending review of the informed consent procedures.

Informed consent can make trials difficult to implement because there are situations where many people refuse to participate. A number of experiments have been carried out on groups of patients who have limited opportunities to refuse their consent. For example, trials have been carried out in the Third World, in poor communities, on soldiers or on prisoners to "simplify" the informed consent procedure. Even if the participants in these trials are properly informed, there is often doubt as to the validity of their consent by virtue of the pressures on them to consent. These may be due to limited access to alternative treatments, the offer of free treatment etc.

Clinical trials in Australia are governed by NH&MRC guidelines which require ethics committees to review and approve trials. The principle function of these committees is to ensure that the rights and welfare of patients are protected. These committees are typically made up both of medical researchers and lay representatives from the community. A secondary issue in the RU486 debate was the make up of the ethics committee which approved the trial. (The religious representative apparently did not attend the meeting at which the approval was granted though, according to the spokesperson, a quorum was present and the members had all received the proposal in advance.) The working of such committees is far from perfect but they do provide at least minimal review of proposed research and the opportunity for community input.

It is probably worth noting that there are some incentives to participating in a clinical trial. These include:

- access to a new treatment (e.g. AZT in the AIDS trial)
- better attention because participants in trials are often looked after better than ordinary patients
- sometimes the promise of access to the winning treatment at the end of the trial
- the knowledge that one is making a scientific contribution.

Behavioural experiments also often involve human subjects and are therefore subject to ethical considerations. Here however, the risks to the subjects are less clear than in clinical trials and it is often not possible to have informed consent and still preserve the experiment. There are situations in which describing the purpose of the experiment would change the way people react and so invalidate the experiment. The role of the review committee is therefore very important here. Proposals for experiments may be vetoed on the grounds that they may have negative emotional effects, undermine human dignity or are simply in poor taste.

Ethical issues also arise in experiments with animal subjects. Proposals for experimentation on animals need ethical review. Typically, a certain amount of harm to the animals is permitted but this has to be balanced against the importance of the knowledge gained from the experiment.

5.7 Experiments versus Observational studies

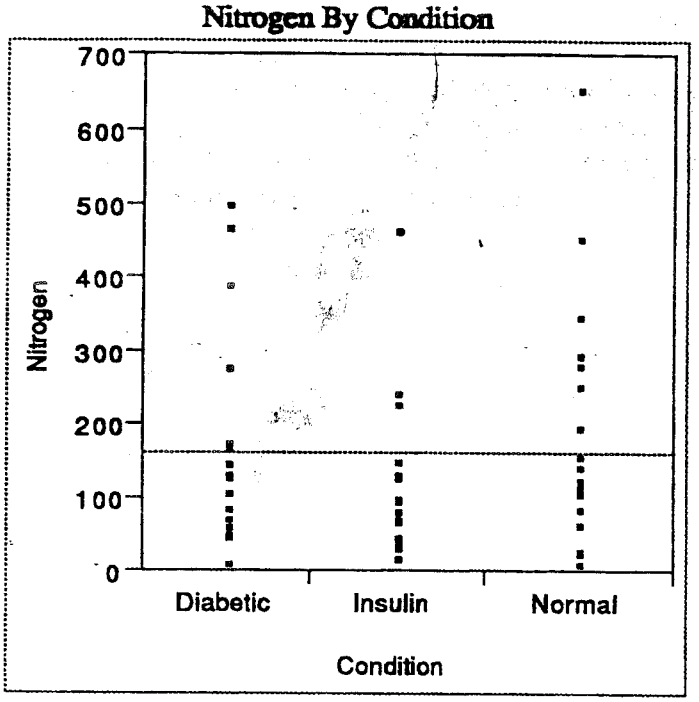
In an experiment, the system under study is, in principle, under the total control of the investigator. The investigator determines the allocation of treatments to units and the physical layout of the experiment. Provided these allocations involve randomisation, it is possible to conclude that any clear cut differences in the response between the treatments is a consequence of the treatments.

On the other hand, in observational studies the investigator observes the data but has no control over it. The main difference between observational studies and experiments is that observational studies include intrinsic explanatory variables which are a property of the units (such as gender) rather than assigned to them. While the analysis of data from an observational study is often the same as analysis of data from an experiment, the interpretation of the results is different. Observational studies are more difficult to interpret because observed differences in the response could possibly be explained by factors which are not known and not measured by the investigator. Moreover, at least in terms of terminology, intrinsic explanatory variables are not usually regarded as causing effects but simply as being a property of the units.

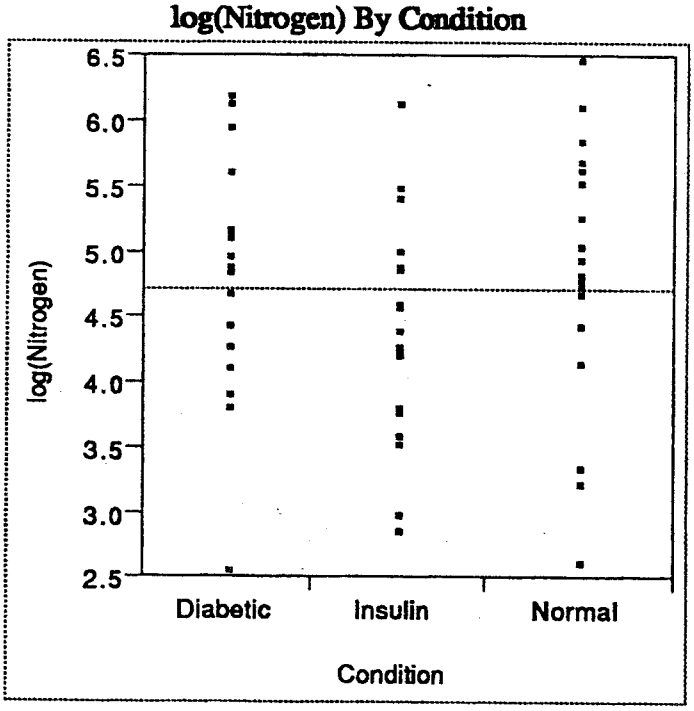
Dolkart, Halpern and Perlman (1971) published data given in **Mouse diabetes** on the antibody responses in alloxan diabetic, insulin–treated diabetic and normal mice injected with bovine serum albumin. The mice in the study represent samples from populations of alloxan diabetic, insulin–treated diabetic and normal mice. This is not an experiment because we cannot randomise a mouse to one of the three groups – the population to which a mouse belongs is simply a property of the mouse. The mice were bled from the orbital sinus and the serum analysed for antigen binding capacity. The data are reported in micrograms of BSA nitrogen bound by 1 ml of undiluted serum. The question of interest is whether or not the diabetic groups have different nitrogen binding capacities from the normal mice.

We can interpret the antigen binding capacity as the response variable and the factor giving the diabetic status of the mice as the explanatory variable. It seems reasonable to assume that all the mice are independent.

A simple plot of the data shows that the distributions of the antigen binding capacity are fairly similar across the three groups. The distributions are all asymmetric with a long right tail so a transformation may make them more normal.



On the log scale, the response distributions are similar and much more normal looking.

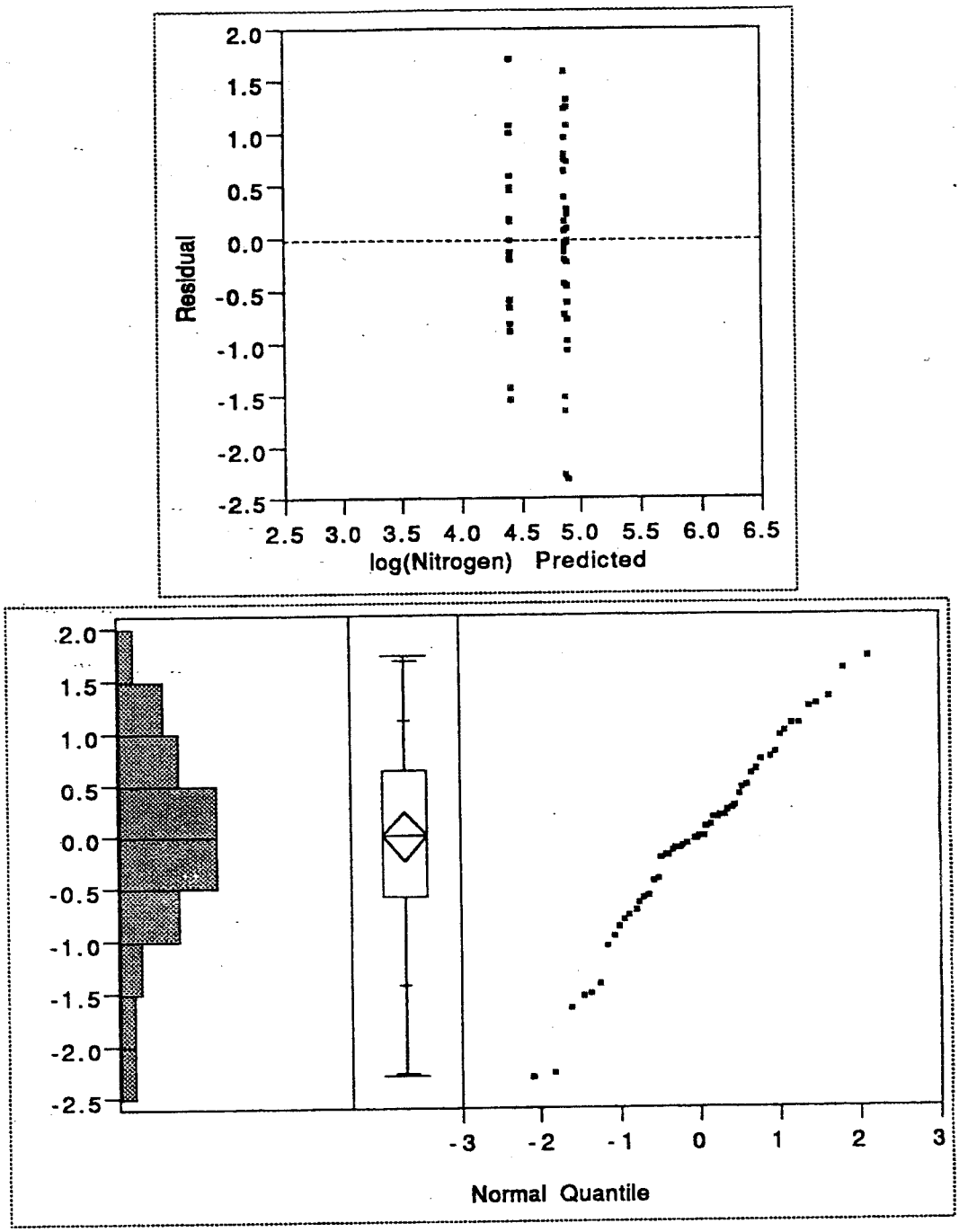


We can now fit a model to describe the differences between the distributions. A plausible model for the log antigen binding capacity is

$$Y_{ij} = \alpha + \delta_j + \varepsilon_{ij}$$

with ε_{ij} independent $N(0, \sigma^2)$

The diagnostics for the model show that it is adequate.



Note that the residual plot shows two vertical bars rather than three because the fitted values are very similar for two of the groups. It is important to realise that the vertical bars are simply due to the structure of the explanatory variable and do not invalidate the model.

Since the model is adequate, we can now test the hypothesis that the distributions are the same by testing $H:\delta_1 = \delta_2 = 0$. The F–test from the analysis of variance table has a p–value of 0.19 so is not significant. In other words, there is no evidence of any difference between the groups.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	2.801929	1.40096	1.6892
Error	54	44.784877	0.82935	**Prob>F**
C Total	56	47.586806		0.1943

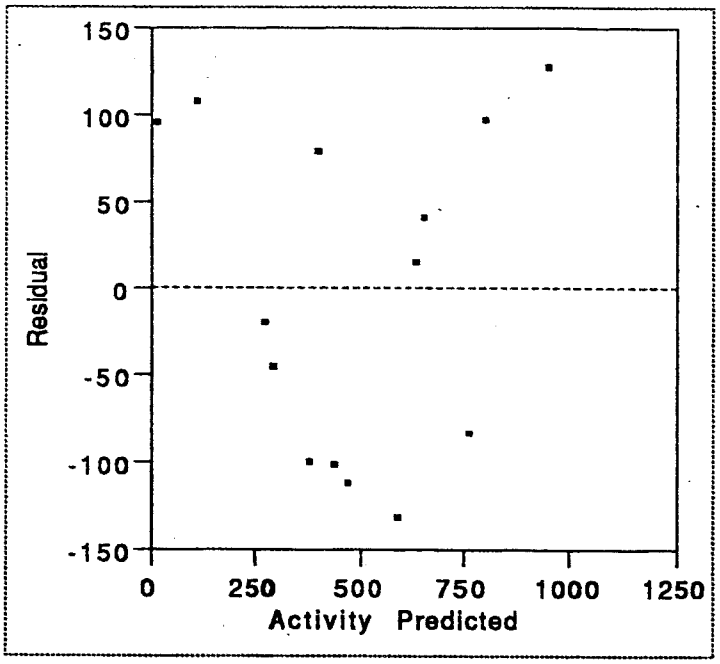
As a second illustration of an observational study, consider the experiment carried out by Elena Gaertner to compare levels of gene expression between leaves and stems of clover plants. The researchers introduced a gene into clover plants and then used flourescence methods to measure the expression levels of the gene. The data consists of the measured flourescence for the leaves and stems from 7 clover plants. There is a natural paired structure in this experiment because the roots and stems are taken from the same plant. However, we cannot randomise between stems and leaves so they are not two levels of a treatment and this is an observational study. The data is available in **Gene expression**.

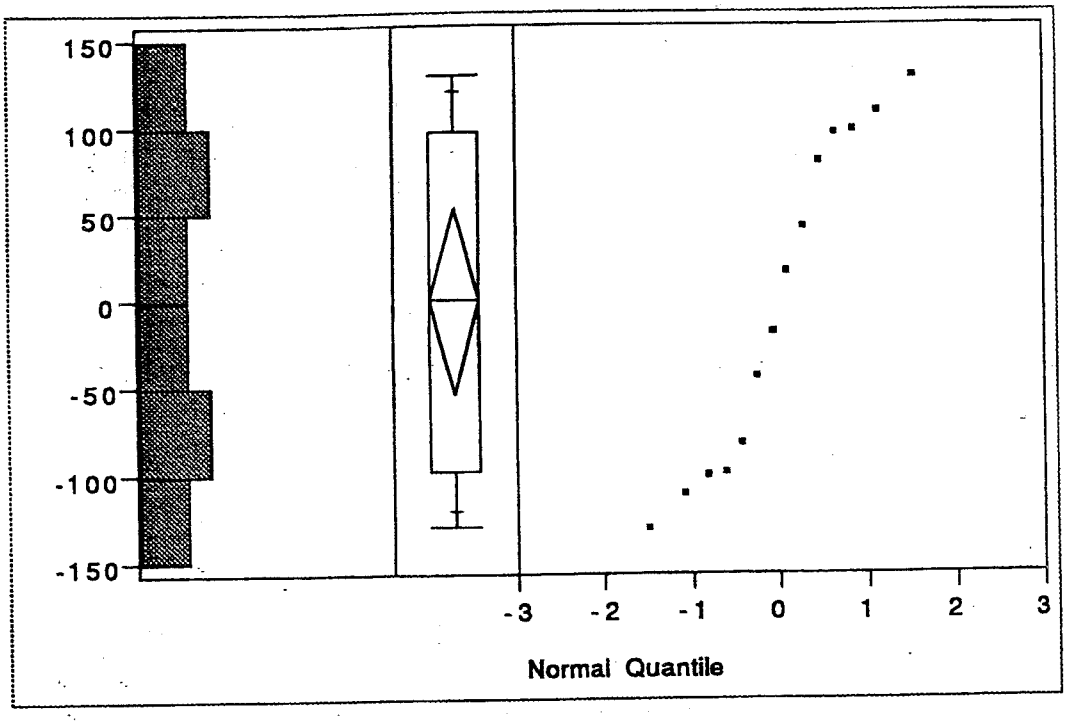
We can analyse data by differencing to remove the paired structure or by fitting a model which includes a random plant effect. That is, the flourescent activity (on some scale) Y_{ij} of the j^{th} site (i.e. stems or leaves) on the i^{th} plant is

$$Y_{ij} = \alpha + \delta_j + \beta_i + \varepsilon_{ij}$$

with β_i independent $N(0, \sigma_\beta^2)$ independent of ε_{ij} independent $N(0, \sigma^2)$.

With only fourteen points the residual plot looks acceptable and the normal quantile plot shows that the residuals have a distribution that is slightly shorter tailed than the normal distribution.





To obtain a confidence interval for the difference in the means, we can refer to the effect test table for the source variable.

Least Squares Means

Level	Least Sq Mean	Std Error	Mean
leaves	296.4285714	52.14396607	296.429
stems	658.1428571	52.14396607	658.143

Warning: Std Err calculated with respect to Synthetic Denominator.

A 95% confidence interval for the difference in the means is

$$296.43 - 658.14 \pm t_6(0.975)\sqrt{52.1440^2 + 52.1440^2}$$

$$= -361.71 \pm 2.447 \text{x} 73.742$$

$$= -361.71 \pm 180.45$$

or (–542.16, –181.26).

Finally, we get estimates of the variance components.

Variance Component Estimates

Component	Var Comp Est
plant	28748.1
Residual	19032.95

We see that between plant variance is greater than the within plant variance so the pairing has reduced the variability.