Contents

Chapter 9

Two Factor Designs - Single-sized Experimental units - CR and RCB designs

9.1 Introduction

So far we've looked at two different experimental designs, the single-factor completely randomized design (1-factor CRD), and the single-factor randomized compete block design (1-factor RCB).

Both designs investigated if differences in the mean response could be attributed to different levels of a single factor. However, in many experiments, interest lies not only in the effect of a single factor, but in the joint effects of 2 or more factors.

For example:

- **Yield of wheat**. The yield of wheat depends upon many factors - two of which may be the variety and the amount of fertilizer applied. This has two factors - (1) variety which may have three levels representing three popular types of seeds, and (2) the amount of fertilizer which may be set at two levels.
- **Pesticide levels**. The pesticide levels may be measured in birds which may depend upon gender (two levels) and distance of the wintering grounds from agricultural fields (three levels).

- **Performance of a product**. The strength of paper may depend upon the amount of water added (two levels) and the type of wood fiber used in the mix (three levels).

There are many ways to design experiments with multiple factors - we will examine three of the most common designs used in ecological research - the completely randomized design (this chapter), the randomized block design (this chapter), and the split-plot design (next chapter).

As noted many times in this course, it is important to match the analysis of the data with the way the data was collected. Before attempting to analyze any experiment, the features of the experiment should be examined carefully. In particular, care must be taken to examine

- the treatment structure;
- the experimental unit structure;
- the randomization structures;
- the presence or absence of balance;
- if the levels of factors are fixed or random effects; and
- the assumptions implicitly made for the design.

If these features are not identified properly, then an incorrect design and analysis of an experiment will be made.

9.1.1 Treatment structure

The treatment structure refers to how the various levels of the factors are combined in the experiment.

The first step in any design or analysis is to start by identifying the factors in the experiment,their associated levels, and the treatments in the experiment. Treatments are the combinations of factor levels that are 'applied' [1] to experimental units.

The two-factor design has, as the name implies, two factors. We generically call these Factor A and Factor B with a and b levels respectively. We will

[1] Recall that in analytical surveys, the factor levels cannot be assigned to units (e.g. you can't assign sex to an animal) and so the key point is that units are randomly selected from the population

examine only **factorial** treatment structures, i.e. every treatment combination appears somewhere in the experiment. For example, if Factor A has 2 levels, and Factor B has 3 levels, then all 6 treatment combinations appear in the experiment.

Why factorial designs?

Why do we insist on factorial treatment structures? There is a temptation to investigate multi-factor effects using a 'change-one-at-time' structure. For example, suppose you are investigating the effects of process temperature (at two levels, H & L), fiber type (at two levels - deciduous and coniferous) and initial pulping method (at two levels - mechanical or chemical) upon the strength of paper. In the 'change-one-at-a-time' treatment structure, the following treatment combinations would be tested:

```
1.  L deciduous  mechanical
2.  H deciduous  mechanical
3.  L coniferous mechanical
4.  L deciduous  chemical
```

The researcher then argues that the effect of fiber type could be found by examining the difference in strength between (1) and (3); the effect of pulping method could be found by examining the difference in strength between (4) and (1); and the effect of process temperature could be found by examining the difference in strength between (1) and (2).

This is valid **provided that the researcher is willing to assume the treatment effects are additive**, i.e., that the effect of process temperature is the **same** at all levels of the other factors; that the effect of fiber type is the **same** at all levels of the other factors; and that the effect of initial pulping method is the **same** at all levels of the other factors. Unfortunately, there is no method available to test this assumption with the set of treatments listed above.

It is usually not a good idea to make this very strong assumption - what happens if the assumption is not true? In the previous example, it means that your 'effects' are only valid for the particular levels of the other factors that happened to be present in the comparison. For example, the process temperature effect would only be valid for deciduous fiber sources that are mechanically pulped.

A superior treatment structure is the **factorial** treatment structure. In the factorial treatment structure, every combination of levels appears in the experiment. For example, referring back to the previous experiment, all of the following treatments would appear in the experiment:

1.	L	deciduous	mechanical
2.	H	deciduous	mechanical
3.	L	coniferous	mechanical
4.	H	coniferous	mechanical
5.	L	deciduous	chemical
6.	H	deciduous	chemical
7.	L	coniferous	chemical
8.	H	coniferous	chemical

Now, the *main effects* of each factor are found as:

- **main effect of temperature** - treatments 1, 3, 5, 7 vs. 2, 4, 6, 8
- **main effect of source** - treatments 1, 2, 5, 6 vs 2, 4, 7 and 8
- **main effect of method** - treatments 1, 2, 3, 4 vs. 5, 6, 7, 8

Each main effect would be interpreted at the 'average change' over the levels of the other factors.

In addition, it is possible to investigate if **interactions** exist between the various factors. For example, is the effect of process temperature the same for mechanical and chemical pulping methods? This would be examined by comparing the change in (1)+(3) vs. (2)+(4) [representing the effect of temperature for mechanically pulped wood] and the change in (5)+(7) vs. (6)+(8) [representing the effect of temperature for chemically pulped wood]. Can you specify how you would investigated the interaction between temperature and source? What about between source and method of pulping? All of these are known as **two factor** interactions.

The concept of a two-factor interaction can also be generalized to three-factor and higher interaction terms in much the same way.

Why not factorial designs?

While a factorial treatment structure provides the maximal amount of information about the effects of factors and their interactions, there are some disadvantages. In general, the number of treatments that will appear in the experiment is equal to the product of the levels from all of the factors. In an experiment with many factors, this can be enormous. For example, in a 10 factor design, with each factor at 2 levels, there are 1024 treatment combinations. It turns out that in such large experiment, there are better ways to proceed that are beyond the scope of this course - an example of which is a *fractional factorial* design which selects a subset of the possible treatments to run with the understanding that the subset chosen loses information on some of the higher order interactions. If you are contemplating such an experiment, please seek competent help.

As well, in some cases, interest lies in estimating a response surface, e.g. factors are continuous variables (such a temperature) and the experimenter is interested in finding the optimal conditions. This gives rise to a class of designs called **response surface designs** which are beyond the scope of this course. Again, seek competent help.

Displaying and interpreting treatment effects - profile plots

An important part of the design and analysis of experiment lies in predicting the type of response expected - in particular, what do you expect for the size of the main effects and do you expect to see an interaction.

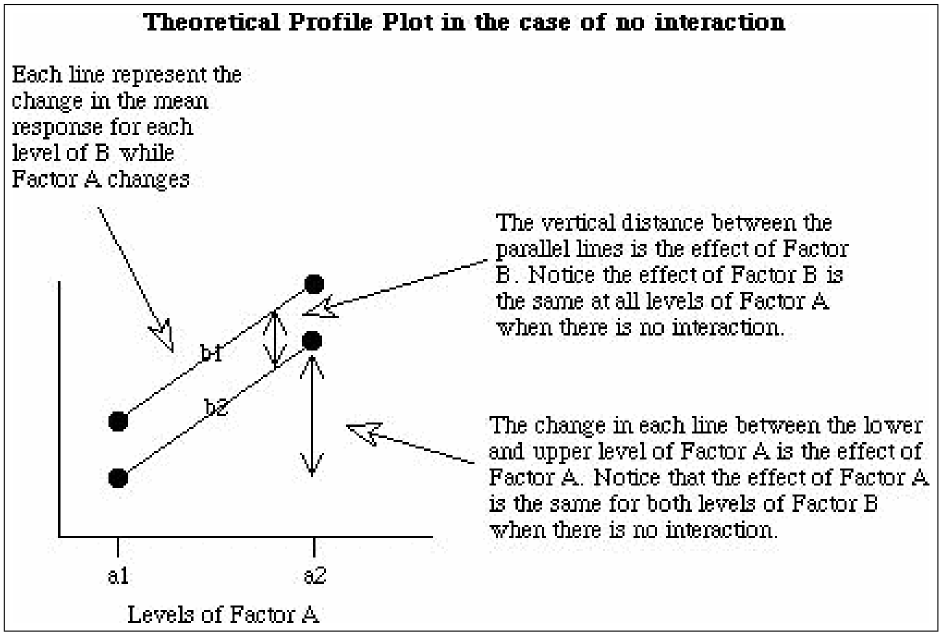
During the design phase, these are useful to determining the power and needed sample sizes for an experiment. During the analysis phase, these values and plots help in interpreting the results of the statistical analysis.

With two factors (A and B) each at two levels, you can construct a **profile plot**. These profile plots show the approximate effect of both factors simultaneously.

The key thing to look for is the 'parallelism' of the two lines.

Profile plots with no interaction between factors

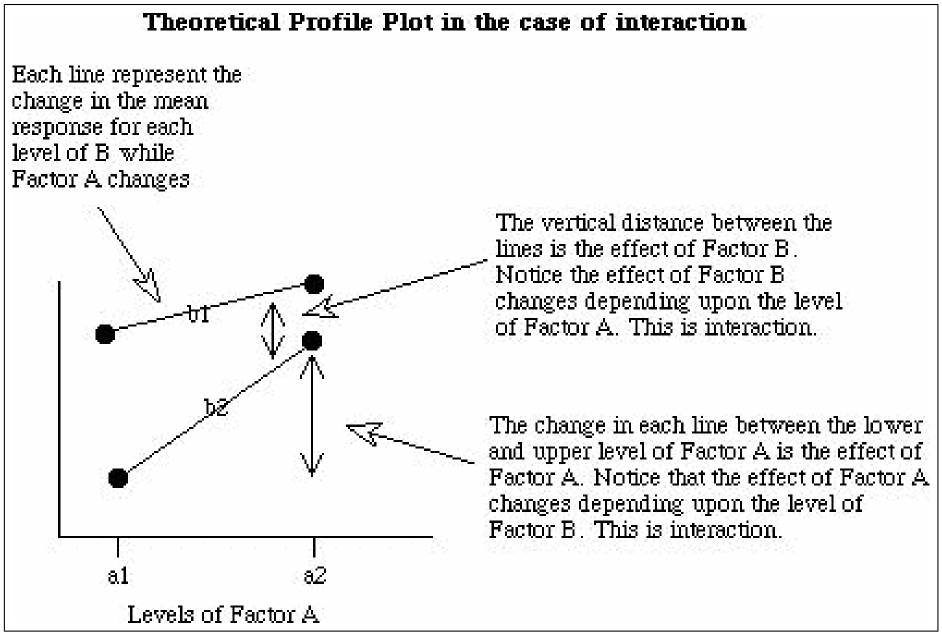
For example, consider the **theoretical** [it is theoretical because it shows the true population means which are never known exactly] profile plot of the mean responses below:


Theoretical Profile Plot in the case of no interaction
Each line represent the change in the mean response for each level of B while Factor A changes
The vertical distance between the parallel lines is the effect of Factor B . Notice the effect of Factor B is the same at all levels of Factor A when there is no interaction.
b1
b2
The change in each line between the lower and upper level of Factor A is the effect of Factor A. Notice that the effect of Factor A is the same for both levels of Factor B when there is no interaction.
a1
a2
Levels of Factor A

In this plot, the vertical distance between the two parallel line segments is the effect of Factor B, i.e., what happens to the mean response when you change the level of Factor B, but keep the level of Factor A constant. The **main effect** of Factor B is the AVERAGE vertical distance between the two lines when averaged over all levels of Factor A. Notice that if the lines are parallel, the vertical distance between the two lines is constant - this implies that the effect of Factor B (the vertical distance between the two lines) is the same regardless of the level of Factor A and the *effect* of Factor B and the *main effect* of Factor B are synonymous. In this case, we say that there is NO INTERACTION between Factor A and Factor B. Similarly, the effect of Factor A is the change in the line between the two levels of Factor A at a particular value of Factor B, i.e., the vertical change in each each line segment. The **main effect** of Factor A is the AVERAGE change when averaged over all levels of Factor B. Notice that if the lines are parallel, the vertical change is the same for both lines - this implies that the effect of Factor A is the same regardless of the level of Factor B and that the *effect* of Factor A is synonymous with the *main effect* of Factor A. Once again, there is no interaction between A and B.

Profile plots with interaction between factors

Now consider the following theoretical profile plot:


Theoretical Profile Plot in the case of interaction
Each line represent the change in the mean response for each level of B while Factor A changes
The vertical distance between the lines is the effect of Factor B. Notice the effect of Factor B changes depending upon the level of Factor A. This is interaction.
b1
b2
The change in each line between the lower and upper level of Factor A is the effect of Factor A. Notice that the effect of Factor A changes depending upon the level of Factor B. This is interaction.
a1
a2
Levels of Factor A

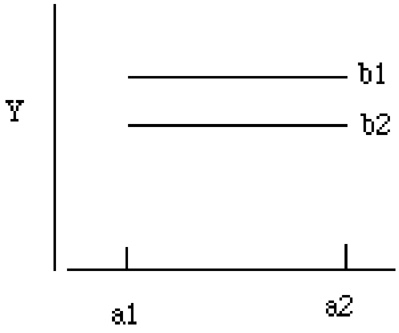
In this plot, the vertical distance between the line segments CHANGES depending on where you are in Factor A. This implies that the effect of Factor B changes depending upon the level of A, i.e., there is INTERACTION between Factor A and B. The main effect of Factor A is the average effect when averaged over levels of B. In this case the main effect is not very interpretable (as will be seen in the plots below). Similarly, the vertical change for each line segment is different for each segment - again the effect of Factor A changes depending upon the level of Factor B - once again there is interaction between A and B.

The plots from an actual experiment must be interpreted with a grain of salt because even if there was no interaction, the lines may not be exactly parallel because of sampling variations in the sample means. The key thing to look for is the degree of parallelism. And it doesn't matter which factor is plotted along the bottom - the plots may look different, but you will come to the same conclusions.

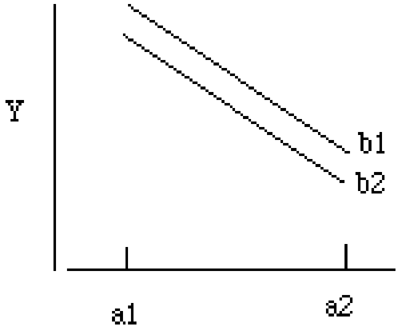
If there is interaction, the line segments may even cross rather than remaining separate.

Illustrations of various theoretical profile plots

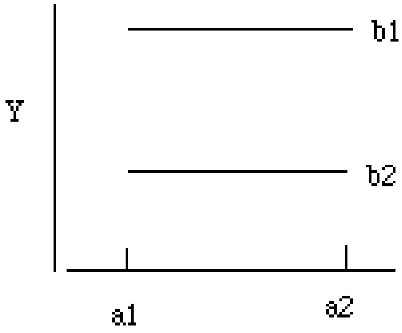
- No main effect of Factor A (average of lines is flat); small main effect of Factor B (if there was no main effect of Factor B the lines would coincide); and no interaction of Factors A and B.



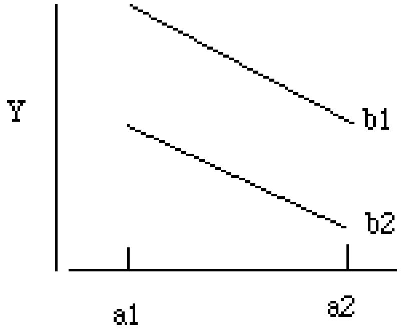
- Large main effect of Factor A; small main effect of Factor B (average difference between lines is small); and no interaction between Factors A and B.



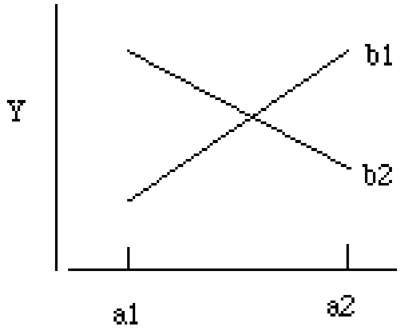
- No main effect of Factor A; large main effect of Factor B; and no interaction between Factors A and B.



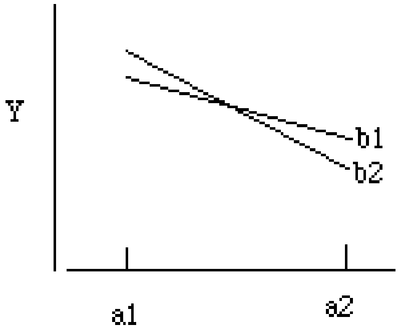
- Large main effect of Factor A; large main effect of Factor B; and no interaction between Factors A and B.



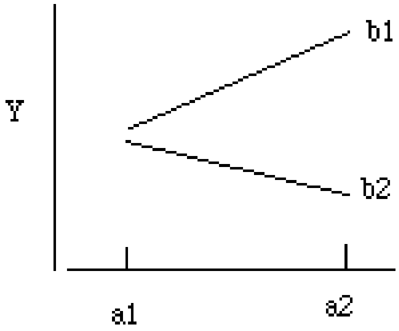
- No main effect of Factor A; no main effect of Factor B; but large interaction between Factors A and B. This illustrates the dangers of investigating 'main effects' in the presence of interaction (why? - a good exam question!).



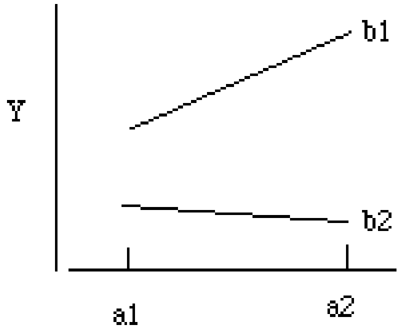
- Large main effect of Factor A; no main effect of factor B: slight interaction. Again, this diagram illustrates the folly of discussing main effects in the presence of an interaction (why?).



- No main effect of Factor A; large main effect of Factor B; large interaction between Factor A and B. As before, there may be problems in interpreting main effects in the presence of an interaction (why?).

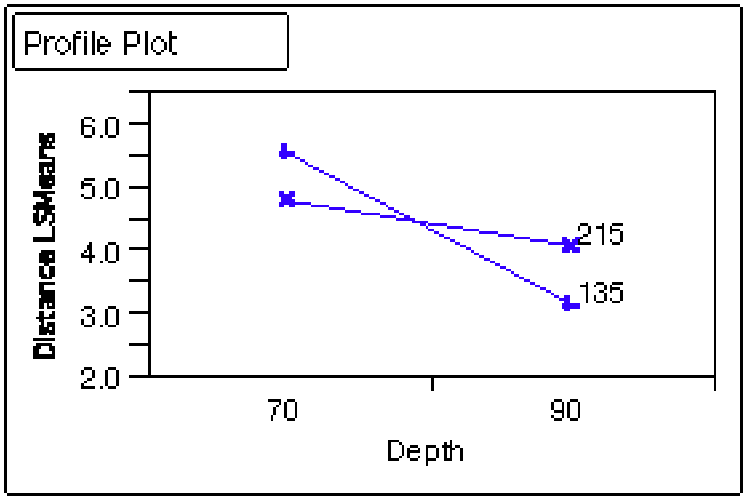


- Small main effect of factor A; large main effect of Factor B; large interaction between Factors A and B. See previous notes about interpreting main effects in the presence of an interaction.



Further examples of profile plots

- Discuss the example of an environmental impact study. Here interaction indicates that there was an impact.
- Discuss profile plots for three factors.
- Here is the profile plot for an experiment to investigate the effect of wing depth and wing width upon the flight of paper airplanes. Based upon the profile plot below, what do you conclude?



The MOF publication Displaying factor relationships[2] also has a discussion

[2] http:../../MOF/pamp55.pdf

on profile plots.

9.1.2 Experimental unit structure

The experimental unit structure is the key element to the design of the experiment and the recognition of an existing experimental design.

In single-factor experiments, there is usually only a single size of experimental unit and the only design choice is blocking or not. However, in multifactor designs, the choices are much greater and the potential problems in design and analysis multiply!

For example, consider the following experimental designs to investigate the effect of light level (at two levels - High and Low), and the amount of water (also at two levels - Dry or Wet) upon the growth of pine tree seedlings in a greenhouse. There are a total of 8 seedlings.

- **Design 1**. Each seedling is put into its own pot. One pot is placed in each of 8 separate growth chambers. Two grow chambers are each assigned to each combination of light level or water amount.
- **Design 2**. Two seedlings are placed in each pot. One pot is placed in each of 4 separate growth chambers. Each grow chamber is given one combination of light level or water amount.
- **Design 3**. Each seedling is put into its own pot. Two pots are placed into each of 4 separate growth chambers. Each growth chamber is assigned one combination of light level and water amount.
- **Design 4**. Each seedling is put into its own pot. Two pots are placed in each of 4 separate growth chambers. Two of the growth chambers are assigned the high light level; two are assigned the low light level. Within each chamber, one pot receives the wet water level; one pot the dry water level.
- **Design 5**. Two seedlings are placed in each pot. Two pots are placed in each of 4 separate growth chambers. Two of the growth chambers are assigned the high light level; two are assigned the low light level. Within each chamber, one pot receives the wet water level; one pot the dry water level. The growth of each seedling is measured.

Without much difficulty, more ways could be found to run this experiment! Each different way of design requires a different analysis!

Which experimental unit structure is better? Design 1 requires the most growth chambers, but is easiest to run. Design 2 requires fewer growth chambers, but suppose that the particular pot had an effect on growth (e.g. the previous researcher used a herbicide in a previous experiment and didn't clean the pot properly). Design 3 requires that two pots be placed in each chamber - is the chamber big enough? There is no best design that fits all problems!

In all cases, the data will consist of 8 measures of growth along with the light level and water level received. **The data will not tell you about the experimental unit structure!** Consequently, it is imperative that you think very carefully about the experimental unit structure and give explicit instructions on how to perform the experiment so that there is no ambiguity. You will see later in the course that every one of the previous designs would be analyzed in a different way!

In this course, we will look at two popular experimental design choices:

1. the Single-size of experimental unit (with and without blocking)
2. the Split-plot Design (with and without blocking) The Split-plot design (which takes its name from its agricultural heritage) is the most common 'complicated' design and, unfortunately, the design that is most often analyzed incorrectly. It is discussed in the next chapter.

The simplest designs have a single size of experimental unit and the observation unit is the same as the experimental unit, i.e. only one measurement is taken on each experimental unit. The greatest advantage of using a single sized unit is that loss of that unit only entails the loss of one data point. If you are conducting multiple measurements on the same unit (e.g. following a unit over time), then the loss of that unit entails the potential loss of much more information. The greatest disadvantage of using a single-sizes unit is that variation in responses may give poor power. However, the simple strategy of blocking is often sufficient to improve power without making the design too complicated.

A common problem is pseudo-replication where the observational unit is not the same as the experimental unit. Hurlbert (1984) should be reread at this point.

In some designs, multiple measurements are taken on the SAME experimental units - typically repeated measurements over time. [3] The most common reason for multiple measurements on the same unit is to have each unit serve

[3] Many experiments have TIME as one of their factors. If the same units are measured repeatedly over time, this is definitely NOT a completely randomized design. A more appropriate analysis would be a *repeated measures* design or a *split-plot-in-time* design. The former is beyond the scope of this course; the latter will be covered in a later section.

as its own control and thereby have greater power to detect changes over the repeated measurements.

9.1.3 Randomization structure

We have already seen two randomization methods:

- **Complete randomization** where treatments are assigned completely at random to experimental units
- **Complete-Block randomization** where experimental units are first grouped into blocks, every block has every treatment, and treatments are randomized to units within blocks.

The actual randomization in practice is a straightforward generalization of that done before for a single-factor design and I won't spend too much time on it but it will be discussed in class.

These will be the only two randomization structures considered in this course. In both cases, there is complete randomization over all units or over all units within a block. This is makes TIME a particularly difficult factor - there is often no randomization to new units at different time points. The problem is that non-randomization often introduces more complex covariance structures among the responses. For example, in repeated measurements over time, measurements that are close together in time would be expected to be more highly correlated than measurements that are far apart in time. In a complete-randomization scheme, the correlation would not expect to change as a function of time separation.

Here are some example of experiments that are **NOT** completely randomized designs.

- Measuring plankton levels at various locations and distances from the shore. At each location, samples are taken at 1, 5, and 10 m from the shore. Here the distance from the shore are not randomized to different locations - each location has all three distances from shore.
- The concentration of a chemical in the blood stream is measured on each rat at 1, 5, and 10 minutes after injection. The time of measurement is not randomized to individual rats - each rat is measured three times.

How could these experiments be redesigned to be CRD's?

9.1.4 Putting the three structures together

We will examine the four most popular experimental designs based upon the above three structures. You may wish to draw a picture of the experimental layouts.

For example, consider the a plant-growth experiment. There are two factors - light level (High and Low), and water level (Wet or Dry). Some possible designs that we will demonstrate how to analyze in this course include:

- **Completely randomized design**. Each seedling is randomly placed in its own pot. Each pot is randomly placed in its own greenhouse. Each greenhouse is randomly assigned one of the four treatments at random. The experiments are all run at the same time or run in random order.
- **Randomized complete block design**. Each seedling is randomly placed in its own pot. Four pots are placed in each greenhouse. Within each greenhouse, each of the four pots is randomly assigned to one of the four treatment combinations. [This may require some modification to the greenhouse so that the two light levels can be applied. within each greenhouse]
- **Split-plot - variant A**. It may be too difficult to modify the greenhouses to have both light levels in each greenhouse. Therefore, two greenhouses are randomly assigned to each light level. Within each greenhouse, two pots are used. These are randomly assigned to the two watering levels.
- **Split-plot - variant B**. Four green houses are not available at one site, but we have two sites available, each with two greenhouses. Therefore, one greenhouse at each site is randomly assigned to each light level. Within each greenhouse, two pots are used. These are randomly assigned to the two watering levels.

Further reading Refer to MOF publication What is the design?[4]

9.1.5 Balance

Balance is a statistical property of a design. Balance used to be much more important in the days of hand computations when the computations for balanced designs were particularly easy to do. This is less important today in the age of computers but the unwary traveler may hit a few pot holes as will be seen in later sections.

[4] http:../../MOF/pamp17.pdf

The very simplest balanced design has an equal number of replicates assigned to each treatment combination. Balance in the experiment will give the greatest power to detect differences among the various treatments. As well, it makes the analysis particularly simple and most computer packages will do a good job of the analysis.

However, because of deliberate decision or demonic interference, unbalanced designs (unequal numbers of replicates for each treatment) can occur. Fortunately, the analysis of such designs in the simple case of a two-factor completely-randomized design is straightforward, but there are a few subtle problems that will be pointed out by example. Some computer packages will give incorrect answers in the case of unbalanced data.

As you will see in a future section, the greatest danger in unbalanced designs is the lack of a complete factorial treatment structure. These type of experiments are extremely difficulty to analyze properly.

9.1.6 Fixed or random effects

In some cases, the choice of levels for a factor is also of concern. If the experiment were to be repeated, would the same levels be chosen (fixed effects) or would a new set of levels be chosen (random effects). Or, is interested limited to the effects of the levels that actually occurred in the experiment (fixed effects) or do you wish to generalize to a large population of levels from which you happened to choose a few for this experiment (random effects).

As an illustration, consider an experiment on the effects of soil compaction on subsequent tree growth. Suppose that the experimenter obtained seedlings from several different seed sources. This experiment could be viewed as having two factors - the level of soil compaction, and the seeding source.

Presumably, if the experiment were to be repeated, the same levels of compaction would be of interest. As well, these levels of compaction are of interest in their own right. Hence, compaction would be treated as a **fixed effect**.

However, what about the factor *seed source*. If the experiment were to be repeated, would the same sources of seeds be used? Are these the only sources of seeds available, or are there many other sources, of which only a few were chosen to be in this experiment? Do you want to extend your inference to other seed sources, or are these the only ones that you are really interested in?

Usually, you will want to argue that your conclusions should extend to other seed sources. If this is the case, then you must be able to argue that the sources you used are in some sense “typical of the ones to which you want to extend your inference". The simplest way to do this is to argue that the sources you selected were essentially a random sample of all possible seed sources to which you wish to extend your inference. This factor is then a **random effect**.

We will start with demonstration of the analysis of experiments where ALL EFFECTS are fixed effects. You can still proceed to analyze experiment with random effects the same way as before — up to a point. It turns out that this seemingly innocuous change to a factor has dramatic implications for the analysis of the experiment! As well **MANY poorly written packages will give WRONG RESULTS!**. In addition, contrary to the impression that statistics is a static science, the whole area of the analysis of models with fixed and random effects is undergoing a revolution in the statistical world. Many of the newer techniques are not discussed in textbooks and certainly not in the published literature. Even experienced statisticians have difficulty in keeping up with advances in this area.

For all but the simplest cases, seek help with models containing combinations of fixed and random effects (often called *mixed models*).

The crucial first step in this model building is deciding which factors are fixed and which factors are random effects.

A factor is a **fixed effect** if :

- the same levels would be used if the experiment were to be repeated;
- inference will be limited ONLY to the levels used in the experiment;

A factor is a **random effect** if:

- the levels were chosen at random from a larger set of levels
- new levels would be chosen if the experiment were to be repeated
- inference is about the entire set of potential levels - not just the levels chosen in the experiment.

Typical fixed effects are factors such as gender, species, dose, chemical. Typical random effects are subject, locations, sites, animals.

9.1.7 Assumptions

Each and every statistical procedure makes a number assumptions about the data that should be verified as the analysis proceeds. Some of these assumptions can be examined using the data at hand; other, often the most important can only be assessed using the meta-data about the experiment. Fortunately, many of the assumptions are identical to those seen in previous chapters. Please consult previous chapters on details on how to verify the assumptions.

The most important assumptions to examine are:

- **The analysis matches the design!** Enough said in past chapters.
- **Equal variation within treatment groups** All the populations corresponding to treatments have equal variances. This can be checked by looking at the sample standard deviations for each group (where each group is formed by one of the treatment combinations). Unless the ratio between the standard deviations is larger than about 5:1, this is not likely a problem. Procedures are available for cases where the variances are not equal in all groups. Fortunately, ANOVA is fairly robust to unequal variances if the design is balanced.

  Often you can anticipate an increase in the amount of chance variation with an increase in the mean. For example, traps with an ineffective bait will typically catch very few insects. The numbers caught may typically range from 0 to under 10. By contrast, a highly effective bait will tend to pull in more insects, but also with a greater range. Both the mean and the standard deviation will tend to be larger. A transformation may be called for (e.g. take logarithms of the response).
- **No outliers** There are no outliers or unusual points. Look at the side-by-side dot plots formed by the treatment groups. Examine the residual plots after the model is fit.
- **Normality within each treatment group** If the sample sizes are small in each group, then you must further assume that each population has a normal distribution. If the sample sizes are large in all groups, you are saved by the 'central limit theorem'. Normal probability plots within each treatment group, or the residuals found after the model fitting procedure can be examined. However, these likely have poor power when the sample sizes are small and will detect minute differences when sample sizes are large. Hence, they are often not very informative.
- **Are the errors are independent?** Another key assumption is that experimental units are independent of each other. For example, the response of one experimental animal does not affect the response of another experimental animal.

9.1.8 General comments

The key to a proper analysis of any experiment is recognizing the design that was used and then specifying a statistical model that incorporates the sources of variation in the design.

As you will see, this statistical model will have terms representing the main effects and interactions of the factors and terms for every size of experimental unit in the experiment. [The latter will become important when we analyze a split-plot design.]

Once the model is specified, then the analysis of variance method (ANOVA) partitions the total variation in the observed responses into *sources* - one for each component representing a main effect or an interaction, and one for every size of experimental unit. [Again, the latter will become more important in split-plot designs.] For example, in the single factor CRD design, the ANOVA table consisted of a line for total variation which is then split into sources representing the contributions from the single factor (the treatment sum of squares), and a contribution for experimental unit effects (the error sum of squares). In the single factor RCB design, the ANOVA table introduced yet another entry for the contribution from blocks (the block sum of squares).

As you will see later in this chapter, two factor design will have lines in the ANOVA table corresponding to the interaction of the two factors and their respective main effects.

Then, starting with interaction, you successively test the hypothesis of 'no effect' from that source, i.e., you first test the hypothesis of no interaction effects, and then, depending upon the results of the test, you may or may not wish to test the main effects.

Rarely, if ever, are tests performed on experimental unit effects - it would be quite rare to expect that the experimental units are exactly identical!

Again, the hypothesis tests only tell you that some effect exists - it doesn't tell you where the effect lies. You may need to explore the responses using multiple comparison procedures and/or confidence intervals for the marginal means or contrasts among means.

As before, you should always assess that your model adequately fits the data, and as well before performing the experiment, determine if the sample size is adequate to detect biologically important effects.

When reporting results in a paper or thesis, try not to overburden the reader - no one is interested in the minute details - they want a broad picture - everything else can likely go into an appendix.

Be sure you carefully describe the experimental design so that some one can verify how the experiment was done.

I would recommend that you ALWAYS show a profile-plot of the means of the various treatment combinations along with approximate 95% confidence intervals - this often tells the entire story.

In terms of the actual statistical computations, usually the F-statistic s are reported along with the p-values, but rarely are all the ANOVA tables shown except in appendices or simple tables. In this day of the WWW, the raw data are often made available on a Web site in case someone else wishes to verify your analysis.

9.2 Completely randomized design - all levels fixed

This is the simplest of the two-factor designs and serves as a template for the analysis of more complex designs. As noted many times in this course, it is important to match the analysis of the data with the way the data was collected. Before attempting to analyze any experiment, the features of the experiment should be examined carefully. In particular, the treatment, experimental unit, and randomization structures; the presence or absence of balance; if the levels of factors are fixed or random effects; and the assumptions implicitly made for the design.

We will proceed by example.

9.2.1 Example - Effect of photo-period and temperature on gonadosomatic index

The *Mirogrex terrau-sanctae* is a commercial sardine like fish found in the Sea of Galilee. A study was conducted to determine the effect of light and temperature on the gonadosomatic index (GSI), which is a measure of the growth of the ovary. [It is the ratio of the gonad weight to the non-gonad weight.] Two photo-periods – 14 hours of light, 10 hours of dark and 9 hours of light, 15 hours of dark – and two temperature levels – 16 and 27 C – are used. In this way, the experimenter can simulate both winter and summer conditions in the region.

Twenty females were collected in June. This group was randomly divided into four subgroups - each of size 5. Each fish was placed in an individual tank, and received one of the four possible treatment combinations. At the end of 3 months, the GSI was measured.

Here are the raw data:

Temperature	Photo-period	
	9 hours	14 hours
27 C	0.90	0.83
	1.06	0.67
	0.98	0.57
	1.29	0.47
	1.12	0.66
16 C	2.31	1.01
	2.88	1.52
	2.42	1.02
	2.66	1.32
	2.94	1.63

Design issues

There are two factors in this experiment - photo-period with 2 levels; and temperature also with 2 levels.

What is the treatment structure?

All of the 4 possible treatment combinations (which are?) appear in this study - hence it has a *factorial* treatment structure.

Now the purpose of this experiment was to simulate summer and winter conditions - however, two of the treatment combinations seem unnatural. Why were these treatment combinations used? How could you run this experiment if you really were interested only in the summer and winter conditions? Is any confounding taking place?

What is the experimental unit structure?

The experimental units were individual tanks and the observational units were the individual fish within a tank. There is only one observation unit per experimental unit.

There are a total of 20 fish each of which was placed in an individual tank. This seems kind of wasteful - 20 tanks are needed as five of the tanks are needed for each treatment to get the same photo-period and temperature treatment combination. What is the problem if you used only 4 tanks with 5 fish in each tank? [Hint - what are the experimental and observation units - and is this pseudo-replication?]

What is the randomization structure?

The article was not very clear, but the treatments appear to be completely randomly assigned to the tanks, etc.

Balance

The design is balanced as an equal number of replicates was performed for teach treatment combination.

Fixed or random factors?

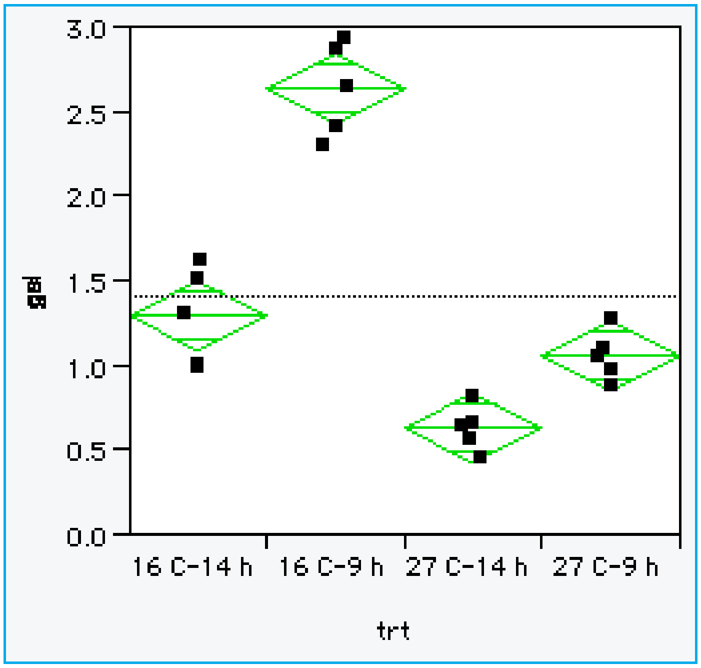
Are the factors to be considered *fixed effects*? In this case, you would use exactly the same levels of both factors - therefore both of the factors are fixed effects.

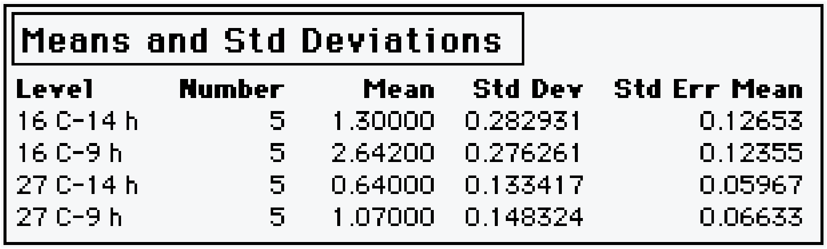
Preliminary summary statistics

Before doing any formal analyses, it is always advisable to do some preliminary plots and compute some simple summary statistics - even if these don't fully tell the whole story. Here are some simple plots and summary statistics [Note that the above data must be converted to a data file in standard format with the appropriate scale of measurements.

Temperature	Photo-period	GSI
27C	09h	0.90
27C	09h	1.06
27C	09h	0.98
27C	09h	1.29
27C	09h	1.12
16C	09h	2.31
16C	09h	2.88
16C	09h	2.42
16C	09h	2.66
16C	09h	2.94
27C	14h	0.83
27C	14h	0.67
27C	14h	0.57
27C	14h	0.47
27C	14h	0.66
16C	14h	1.01
16C	14h	1.52
16C	14h	1.02
16C	14h	1.32
16C	14h	1.63

Sometimes it is easier to create a ‘pseudo-factor’ consisting of the actual treatment levels to make simple plots and to find simple summary statistics. [Illustrate how to do this in JMP - a similar procedure would be done in SAS]. Because this is a completely randomized design, there is no conceptual difference between a two factor design (each with 2 levels) and a single factor design with 4 levels. In more complex designs, this is not true.





Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean
16 C-14 h	5	1.30000	0.282931	0.12653
16 C-9 h	5	2.64200	0.276261	0.12355
27 C-14 h	5	0.64000	0.133417	0.05967
27 C-9 h	5	1.07000	0.148324	0.06633

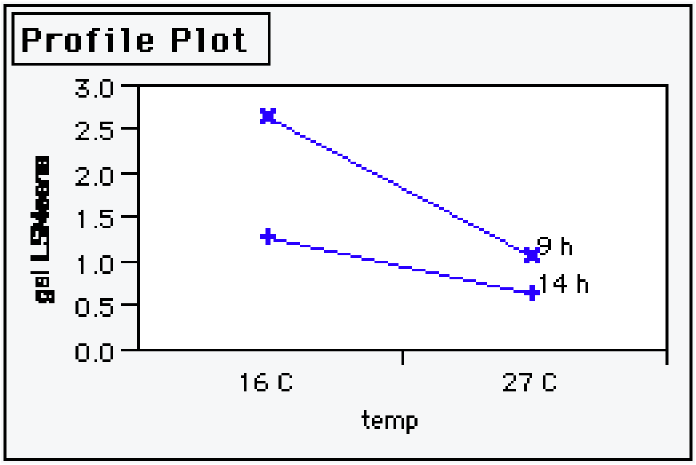
Because the overall design is a CRD, the standard errors reported are sensible. If a blocking factor was available, the *Analyze->Fit Y-by-X* platform would also have computed proper standard errors after block centering. In all other cases, the reported standard errors would not be sensible as the assumed design in the *Analyze->Fit Y-by-X* platform (an RCB or CRD) didn't match the actual design.

Hmmm.. the standard deviation seems to show that the variability at 27C is about 1/2 of that at 16C. This is an interesting effect in its own right - however, the change in standard deviation is small enough that it shouldn't be too much of a concern for this problem. [As a rough rule of thumb, unless the ratio of standard deviations from small samples is on the order of at least 3 to 5x times different, there is likely nothing to worry about.]

The design is balanced - every treatment has the same number of replications

- this makes the analysis easier. Also every treatment combination has some data - missing cells where some cells have no data are a **REALLY MESSY PROBLEM**. Most statisticians even have difficulty in analyzing such experiments - beware!

You should also draw a preliminary profile plot to get a sense of the level of Interaction, if any. This can be done by hand or using Excel. If you are using JMP, you must actually fit the model first to get the interaction profile plot which seems backwards.



It appears that there may be a bit of interaction between the two factors - the lines are not parallel. It would be easier to assess interaction if the approximate 95% c.i. were drawn for each mean - why most packages don't do this is beyond me.

Looking at the profile plot above, what is the effect of photo-period at 16C? at 27C? What is the effect of temperature at 9 h? at 14 h?

The statistical model

The statistical model for any design has terms corresponding to the treatment, experimental unit, and randomization structure. Fortunately, in simple designs, the latter two are often implicit and do NOT have to be specified by the analyst.

Any factorial treatment structure will have terms corresponding to interactions and main effects.

In cases where there is only one size of experimental unit, and no subsampling or pseudo-replication, and no blocking, then it is not necessary to specify any term for experimental unit effects (this corresponds to the MSE line in the ANOVA table).

In cases of complete randomization, there is no need to specify anything further in the model. In cases of non-randomization (e.g. repeated measurement over time, you might specify the covariance structure of the observations).

This gives a model often written as:

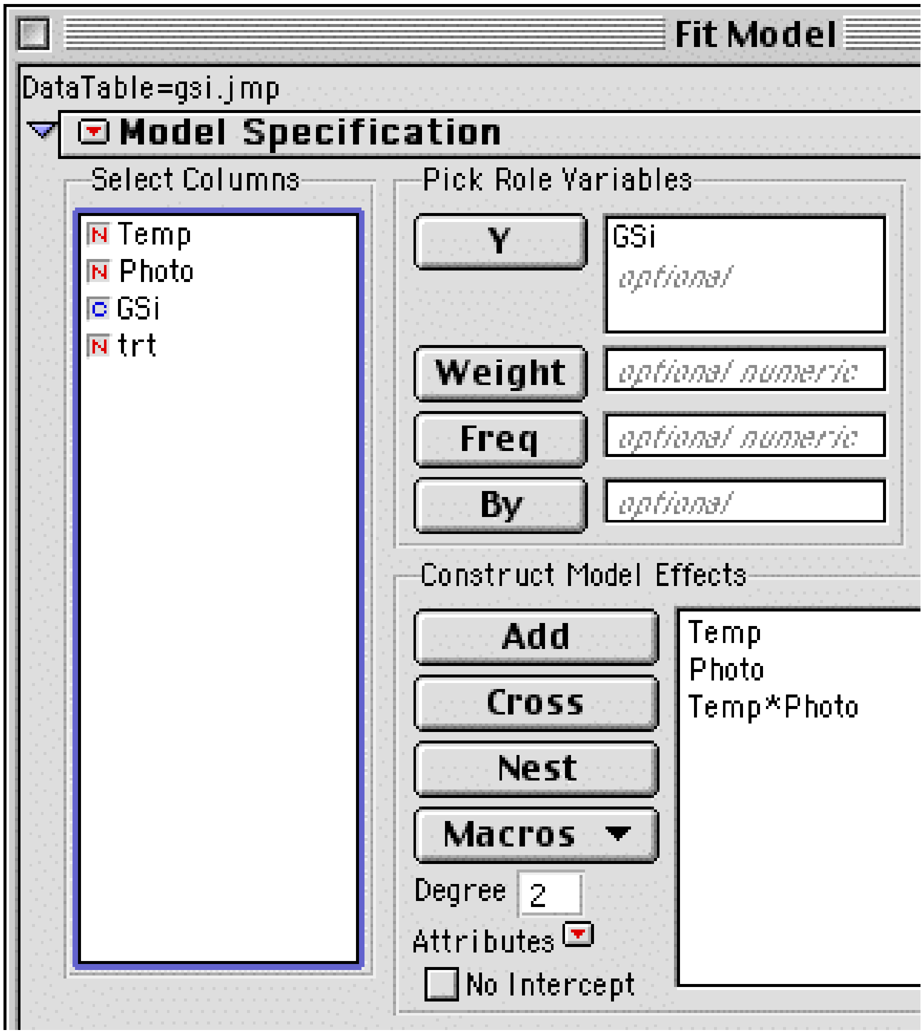
$$GSI = temp \;\; photo \;\; temp * photo$$

What the statistical model says is that we recognize that the observed GSI response values (left of equals sign) are not all the same. What are the various sources of variation in the observed responses? These appear to the right of the equal sign. Well, we expect some differences due to the main effects of *temperature*, some differences due to the main effects of *photo-period*, some differences possibly caused by an interaction between photo-period and temperature. Note that the '*' does NOT imply multiplication, but rather an interaction between two factors. The terms can be written in any order.

There are NO terms representing experimental units (this implies there is a single size of experimental unit), nor any terms representing randomization effects (complete randomization is assumed).

Fitting the model

In order to fit this model, we must use the *Analyze->Fit Model* platform of JMP and complete the dialogue entries as shown below.



DataTable=gsi.jmp

Select Columns: Temp, Photo, GSi, trt

Y: GSi

Model Effects:
Temp
Photo
Temp*Photo

Degree 2

□ No Intercept

The response variable, *GSI*, is entered in the *Y* box. It must be continuous (interval or ratio scales).

Every term in the statistical model must have a corresponding term in the *MODEL EFFECTS* box. The main effects are entered by selecting each variable in turn and then pressing the *ADD* button. Interactions are entered by selecting BOTH variables simultaneously, and then pressing the *CROSS* button. The order of effects in the EFFECTS box is not important - it will give output in a slight different order, but no substantive changes. [5]

[5] You can also enter a factorial structure by jointly selecting the two factors and using the *Macro* button to select a factorial structure

Some of the other buttons are for special cases such as NO INTERCEPT (it is very, very rare to check this box), or if you don't with standard least squares, etc. These are beyond the scope of this course.

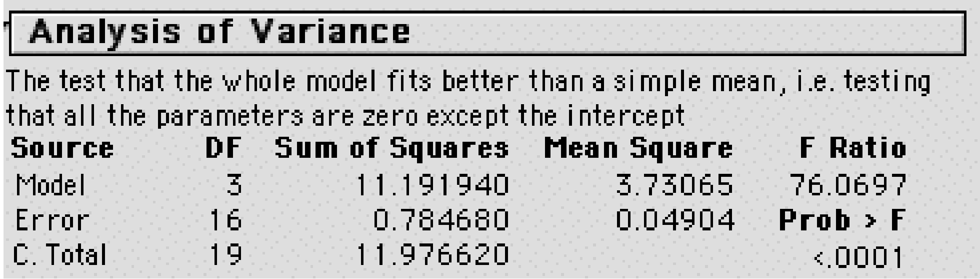
The *RUN MODEL* button, then fits the model to the data.

Hypothesis testing and estimation

The output from the Fit-Model platform is voluminous. It is divided into sections corresponding to the whole model (in the left most column) and then sections corresponding to each effect in the model.

Here are the ANOVA tables from fitting the model. In JMP the various pieces are all over the place - you may wish to sit at a terminal to reproduce the output below. Most packages will give similar output.

The first output below is not very useful - it is a **Whole Model** test which simply examines if there are any statistically significant effects anywhere in the experiment. It is rarely useful.

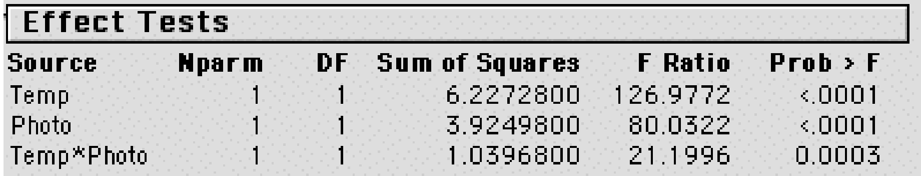


Analysis of Variance

The test that the whole model fits better than a simple mean, i.e. testing that all the parameters are zero except the intercept

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	11.191940	3.73065	76.0697
Error	16	0.784680	0.04904	**Prob > F**
C. Total	19	11.976620		<.0001

The second table below breaks down the Model line in the whole model test into the components for every term in the model. Some packages give you a choice of effect tests. For example, SAS will print out a Type I, II, and III tests - in balanced data these will always be the same. In unbalanced data, these tests will have different results - which test is the 'correct' test is still an item of controversy among statisticians and can (and do) results in fist-fights among the various camps (and you thought statistics was dull!)



Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Temp	1	1	6.2272800	126.9772	<.0001
Photo	1	1	3.9249800	80.0322	<.0001
Temp*Photo	1	1	1.0396800	21.1996	0.0003

Start the hypothesis testing with the most complicated effects (usually interactions) and work towards simpler terms (main effects).

In this case, we start with the test for no interaction effects.

Our null hypothesis is
H: no interaction among photo-period and temperature in their effects on the mean GSI level
A: some interaction among photo-period and temperature in their effects on the mean GSI level.

Our test statistic is F=21.2, the p-value (.0003) is very small. There is very strong evidence of an interaction among the two factors in their effects on the mean GSI level. This is not surprising, the profile plots showed that the lines didn't appear to be too parallel.

What does a statistically significant interaction mean? It implies that the effect of temperature upon the mean GSI is different at the the various photo-period levels. Similarly, the effect of photo-period upon the mean GSI index is different at the two temperature levels.

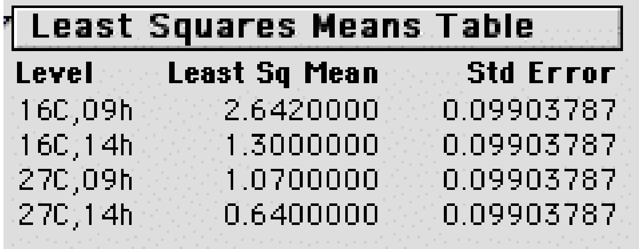
If you detect an interaction, it usually doesn't make much sense to continue along to test main effects because, by definition, these are not consistent - e.g. the effect of temperature is different at the two photo-periods.

What to do if an interaction is present?

There is no single way to proceed after this point. Some authors suggest that you now break up the data into two mini- experiments and analyze each separately. For example, analyze each photo-period separately and analyze each temperature level separately to estimate the effects at each of the various levels. As these mini-experiments are now simply single-factor CRDs (in this case two-sample t-test) all the machinery that we had before can be brought into bear. The disadvantage of this approach is that you forgo pooling of the error variances from all four groups.

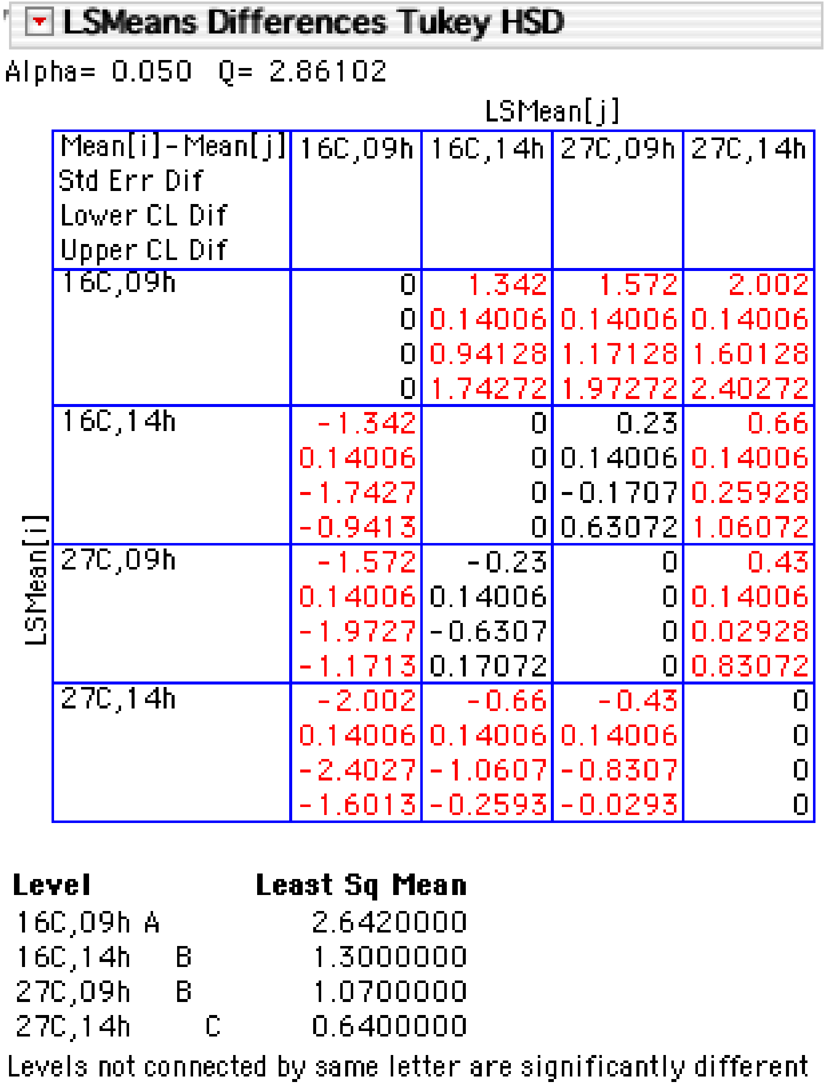
The output from JMP provides additional information about the interaction.

In the column corresponding to the interaction effect, estimates of the means corresponding to each combination of levels, and estimates of the differences between pairs of means (adjusted for multiple comparisons are available).



Least Squares Means Table

Level	Least Sq Mean	Std Error
16C,09h	2.6420000	0.09903787
16C,14h	1.3000000	0.09903787
27C,09h	1.0700000	0.09903787
27C,14h	0.6400000	0.09903787



LSMeans Differences Tukey HSD

Alpha= 0.050 Q= 2.86102

LSMean[j]

Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	16C,09h	16C,14h	27C,09h	27C,14h
16C,09h	0 0 0 0	1.342 0.14006 0.94128 1.74272	1.572 0.14006 1.17128 1.97272	2.002 0.14006 1.60128 2.40272
16C,14h	-1.342 0.14006 -1.7427 -0.9413	0 0 0 0	0.23 0.14006 -0.1707 0.63072	0.66 0.14006 0.25928 1.06072
27C,09h	-1.572 0.14006 -1.9727 -1.1713	-0.23 0.14006 -0.6307 0.17072	0 0 0 0	0.43 0.14006 0.02928 0.83072
27C,14h	-2.002 0.14006 -2.4027 -1.6013	-0.66 0.14006 -1.0607 -0.2593	-0.43 0.14006 -0.8307 -0.0293	0 0 0 0

LSMean[i]

```
Level              Least Sq Mean
16C,09h A              2.6420000
16C,14h   B            1.3000000
27C,09h   B            1.0700000
27C,14h     C          0.6400000
Levels not connected by same letter are significantly different
```

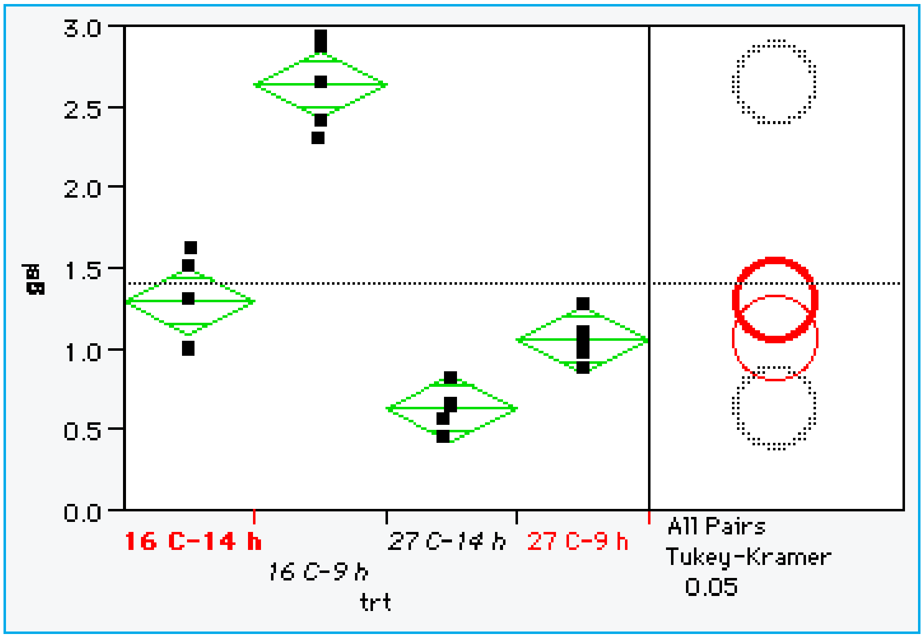
The LSMeans estimates are equal to the raw sample means – this is will be true ONLY in balanced data. In the case of unbalanced data (see later), the LSMEANS seem like a sensible way to estimate marginal means.

The above output can be used to see which means appear to differ from each other.

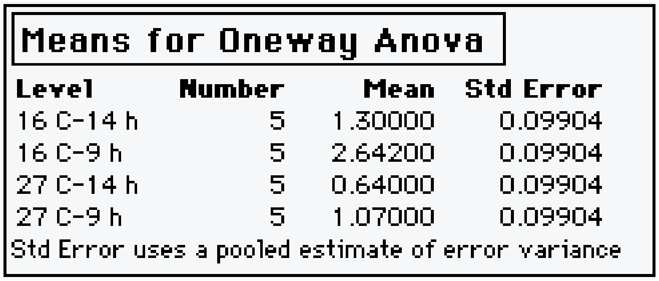
This is also where the profile plot shown earlier is produced - it is a pity that the plot doesn't show the confidence intervals.

Alternately you can return to the pseudo-factors that you defined earlier. This again converts the experiment from a two-factor CRD to a single factor CRD with 4 levels. This approach is only valid because the original design is CRD.

Below is the result of such an analysis:



The ANOVA table is identical to Overall Model ANOVA table earlier. Here are the estimated means for each treatment and the estimated standard errors – identical to the above analysis.

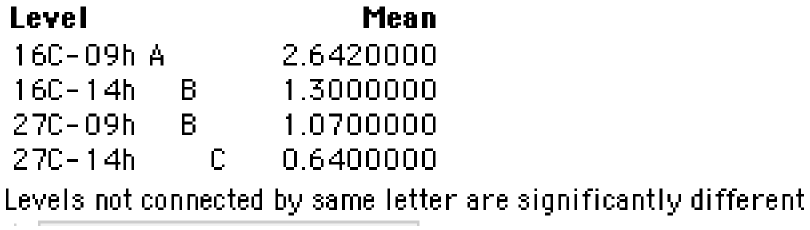


Means for Oneway Anova

Level	Number	Mean	Std Error
16 C-14 h	5	1.30000	0.09904
16 C-9 h	5	2.64200	0.09904
27 C-14 h	5	0.64000	0.09904
27 C-9 h	5	1.07000	0.09904

Std Error uses a pooled estimate of error variance

We perform a multiple comparison procedure (also refer to the comparison circle above) which gives the same results as found previously.



```
Level                  Mean
16C-09h A         2.6420000
16C-14h   B       1.3000000
27C-09h   B       1.0700000
27C-14h     C     0.6400000
Levels not connected by same letter are significantly different
```

What can you conclude from these analyses? In particular, can you see why interaction was detected? Which means appear to be different from the others?

Analysis using SAS

Here is a link to a sample SAS program[6] and output[7]. You should find that most of what we found above also appears somewhere in the SAS output.

9.2.2 Example - Effect of gender and species upon chemical uptake

Several persistent chemicals accumulate up the food chain. Different species may differ in the amount of of chemicals accumulated because of different prey availability or other factors. Because of different behavior, the accumulated amount may also vary by gender.

A survey was conducted to investigate how the amount of PCBs varied among three different species of fish in Nunavut (the new Canadian territory just to east of the Restofit and just north of Ulofit). Samples were taken from four fish of each sex and species and liver PCB levels (ppm) were measured.

Here are the raw data:

[6] ../MyPrograms/gsi.sas

[7] ../MyPrograms/gsi.lst

PCB	sex	Species
21.5	m	sp1
19.6	m	sp1
20.9	m	sp1
22.8	m	sp1
14.5	m	sp2
17.4	m	sp2
15.0	m	sp2
17.8	m	sp2
16.0	m	sp3
20.3	m	sp3
18.5	m	sp3
19.3	m	sp3
14.8	f	sp1
15.6	f	sp1
13.5	f	sp1
16.4	f	sp1
12.1	f	sp2
11.4	f	sp2
12.7	f	sp2
14.5	f	sp2
14.4	f	sp3
14.7	f	sp3
13.8	f	sp3
12.0	f	sp3

Design issues

There are two factors in this experiment - sex with 2 levels and species with 3 levels.

What is the treatment structure? All of the 6 possible treatment combinations (which are?) appear in this study - hence it has a *factorial* treatment structure.

What is the experimental unit structure? Hmmm . . . an interesting question. In observational studies it is often not clear what are the experimental and observational units. For example, is this like a 'fish tank' study where all the fish in a particular location are subjected to the same treatments (i.e., deposited PCBs). Or is each fish subjected to its own experience?

This is very common problem in observation studies and you should be very careful about the dangers of pseudo-replication that we explored earlier.

For now, lets treat the experimental units as individual fish and the observational units as the individual fish. There is only one observation unit per experimental unit.

Finally, what is the randomization structure? Again, this is often not clear in observational studies. First, it is quite impossible to randomly assign sex or species to fish. You must view the randomization as arising from the selection process. Are these fish randomly selected from the entire population of fish of each species and sex? Or is the sample a convenience sample - i.e., the fish closest to the research station that are easiest to catch?

In any observational study, you must be careful that the units measured are a proper random sample from the relevant populations.

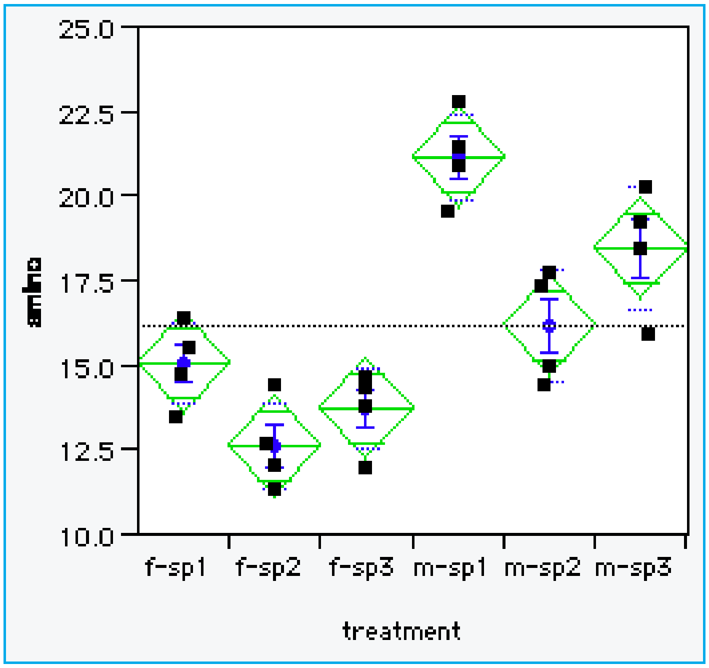
Are the factors to be considered *fixed effects*? Does it seem reasonable that if you were to repeat the survey, you would select the same sexes and species? In this case, you would use exactly the same levels of both factors - they are fixed-effects.

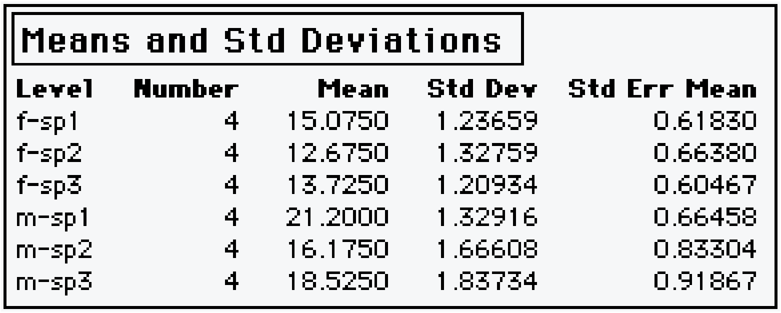
Hence, this experiment appears to satisfy the requirements for a two-factor fixed-effects CRD. In particular, the "randomization" was to individual experimental units and the observational unit is the same as the experimental unit.

Preliminary summary statistics

Again, create some simple summary statistics. We will create a 'pseudo-factor' in JMP so that summary statistics can be computed on each group.

Be sure to specify that *sex*, *species*, and the pseudo-factor *treatment* are nominal scaled variables, while *pcb* is a continuous variable.





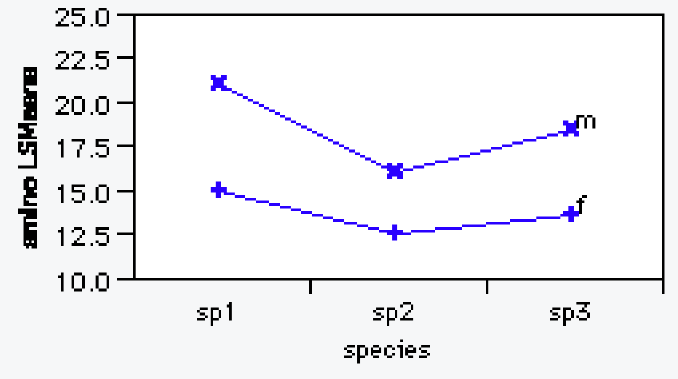
Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean
f-sp1	4	15.0750	1.23659	0.61830
f-sp2	4	12.6750	1.32759	0.66380
f-sp3	4	13.7250	1.20934	0.60467
m-sp1	4	21.2000	1.32916	0.66458
m-sp2	4	16.1750	1.66608	0.83304
m-sp3	4	18.5250	1.83734	0.91867

The standard deviations are approximately equal in all the groups. There don't appear to be any outliers or unusual points. In general the males seem to have higher levels of PCBs than the females, but there doesn't seem to be much of a difference among the mean PCB levels in the species.

The design is balanced - every treatment has the same number of replications - this makes life easier. Every treatment combination has some data - again it makes our analysis task easier.

We draw the profile plots. This can be done by hand or using Excel. If you are using JMP, you must actually fit the model first to get the interaction profile plot which seems kind of backwards.



The lines appear to be roughly parallel - so we expect that there may not be an interaction between the two factors.

Looking at the profile plot - what is the effect of gender? What is the effect of species?

The statistical model

The statistical model is written as:

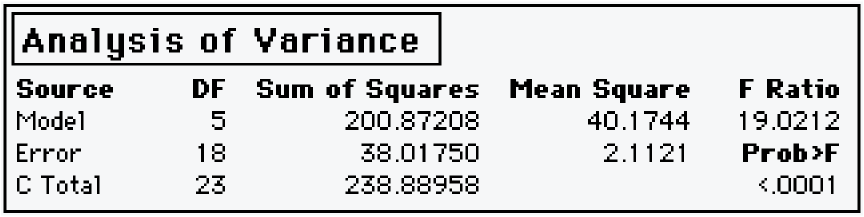
$$PCB = gender \quad species \quad gender * species$$

What does this statistical model tell us about the sources of variation in the observed data?

Fitting the model

We use the *Analyze->Fit Model* platform of JMP and complete the dialogue entries as in the previous example (we won't show it for this example).

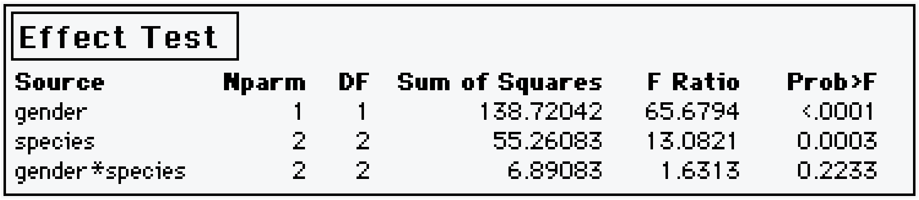
Here are the ANOVA tables from fitting the model.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	200.87208	40.1744	19.0212
Error	18	38.01750	2.1121	**Prob>F**
C Total	23	238.88958		<.0001

The **Whole Model** test simply examines if there are any statistically significant effects anywhere in the experiment. It is rarely useful.



Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
gender	1	1	138.72042	65.6794	<.0001
species	2	2	55.26083	13.0821	0.0003
gender*species	2	2	6.89083	1.6313	0.2233

The above table breaks down the whole model test according to the various terms in the model.

Start the hypothesis testing with the most complicated terms and work towards simpler terms.

The first null hypothesis is
H: no interaction between gender and species in their effect on the mean PCB levels
A: some interaction among gender and species in their effect on the mean PCB levels.

The test statistic is F=1.6313, the *p*-value (.2233) is not very small. Hence there is no evidence of an interaction in the effects of gender and species upon the mean PCB levels. This is not too surprising as the lines are fairly parallel. What does this mean in terms of the original responses, i.e., what does no interaction say about the differences in the mean PCB levels among the genders or among the species.

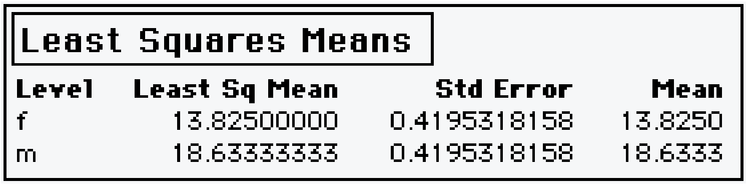
If interaction was not statistically significant, then the analysis continues along to examine the main effects. These can be examined in any order.

Examining main effects - gender.

What are the null and alternate hypotheses? The ANOVA table gives F=65.6 and the *p*-value < 0.0001 - very small. There is very strong evidence of a difference in the mean PCB levels between the two genders.

Because there are only two levels of gender, no multiple comparison procedure is needed. We would like estimates of the marginal means i.e., estimates of the mean PCB levels for each gender averaged over species, and, if possible, estimates of the mean difference in PCB levels between the two genders averaged over species:

We look at the section of the JMP output that deals with Gender effects to find the table of marginal means:

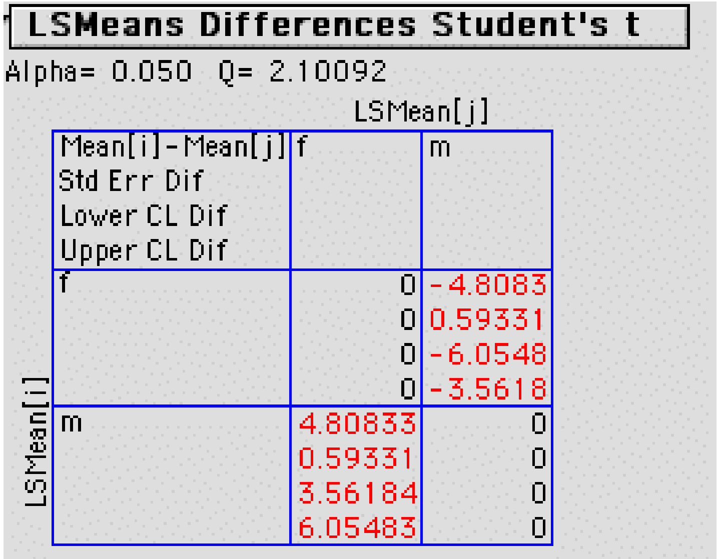


```
Least Squares Means
Level    Least Sq Mean     Std Error      Mean
f          13.82500000  0.4195318158   13.8250
m          18.63333333  0.4195318158   18.6333
```

We can compute approximate 95% confidence intervals from the mean and standard error for each marginal mean.

In multifactor designs, it may not be very useful to estimate the marginal means. Does it make sense to take an average over the three species with equal weight given to each species? If one species is more abundant than another species, perhaps it should be given a greater weight?

Estimates of the difference are always useful. These are obtained from the pop-down menu item for a multiple comparison. [As there are only two levels in this factor, it doesn't matter which multiple comparison procedure is chosen.]



```
LSMeans Differences Student's t
Alpha= 0.050  Q= 2.10092
                              LSMean[j]
            Mean[i]-Mean[j] f          m
            Std Err Dif
            Lower CL Dif
            Upper CL Dif
            f                      0   -4.8083
                                   0   0.59331
                                   0   -6.0548
                                   0   -3.5618
LSMean[i]   m               4.80833          0
                            0.59331          0
                            3.56184          0
                            6.05483          0
```

The estimated difference is -4.8 ppm (i.e., females have lower mean PCB levels on average than males) with a *se* of 0.5933. A 95% confidence interval for the difference is also given.

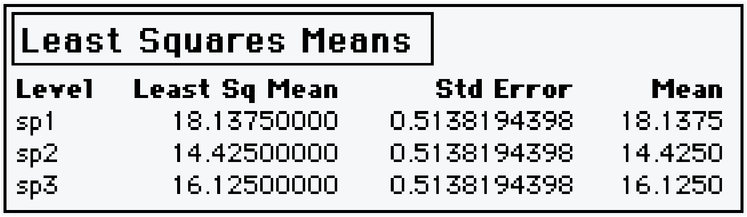
Main effects should ONLY be examined if the interaction effects are not statistically significant or if the non-parallelism is not very large.

Examining main effects - species

What are the null and alternate hypotheses?

The ANOVA table gives F=13.1 and a *p*-value of 0.0003 - very small. There is very strong evidence of a difference in the mean PCB levels among the three species.

Once again, the test just tells us that there is evidence of a difference in the means, but doesn't tell us which mean appears to be different. First examine the estimates of the marginal means:

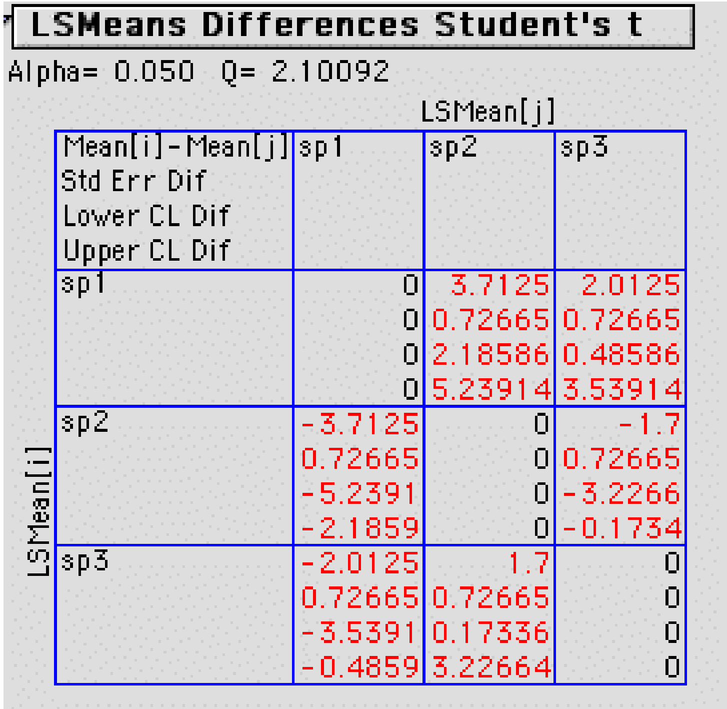


Least Squares Means

Level	Least Sq Mean	Std Error	Mean
sp1	18.13750000	0.5138194398	18.1375
sp2	14.42500000	0.5138194398	14.4250
sp3	16.12500000	0.5138194398	16.1250

We can compute approximate 95% confidence intervals from the mean and standard error for each marginal mean. What does this appear to show us? Again, is it sensible to "average" over the two sexes? As most species have an equal sex ratio, this is likely a sensible thing to do.

How about the estimate of the differences in the mean PCB levels among the species? The multiple comparisons can again be selected:



LSMeans Differences Student's t

Alpha= 0.050 Q= 2.10092

LSMean[j]

LSMean[i]

Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	sp1	sp2	sp3
sp1	0 0 0 0	3.7125 0.72665 2.18586 5.23914	2.0125 0.72665 0.48586 3.53914
sp2	-3.7125 0.72665 -5.2391 -2.1859	0 0 0 0	-1.7 0.72665 -3.2266 -0.1734
sp3	-2.0125 0.72665 -3.5391 -0.4859	1.7 0.72665 0.17336 3.22664	0 0 0 0

A profile plot of the means can also be obtained using the pop-down menu.

Analysis using SAS

Here is a link to sample SAS program[8] and its output[9]. You should find that most of what we found above also appears somewhere in the SAS output - except for power calculations which are not easily done in SAS except under INSIGHT.

9.2.3 Power and sample size

The estimation of appropriate sample sizes for an experiment proceeds in much the same way as for a single factor CRD.

Power and sample size determination for each of the main effects proceed in much the same way as for single factor designs. The *DOE->PowerSampleSize* platform is again used. If there are only two levels for a factor, you can use the

[8] http:../MyPrograms/pcb.sas

[9] http:../MyPrograms/pcb.lst

2-means option and you specify the difference to be detected. If there are two or more levels, you can use the *k-means* option and specify a set of means such that the largest and smallest values differ by the biologically significant difference and the other means are located midway between the upper and lower values. The estimate of σ can be obtained from $\sqrt{MSE}$ in the ANOVA tables, from a pilot study, or from expert opinion. The routine will then return the **TOTAL** sample size which must be split over **ALL** treatment combinations.

For example, consider a two factor experiment with Factor A at two levels and Factor B at three levels. If you find that the total sample size for detecting differences among the means for Factor A is 60, then this must be divided among the $2 \times 3 = 6$ treatment combinations which implies a sample size of 10 for each treatment combinations.

As well, you will have to compute a total sample size for each main effect. These may not be consistent with each other, e.g., the total sample size to detect the effects of Factor A may be 60, while the total sample size to detect the effects of Factor B may be 90. If possible, use the larger sample size to ensure adequate power for all factors.

For example, consider the PCB example. Here the estimate of σ is about $\sqrt{2.11} = 1.45$. The *DOE->PowerSampleSize* platform is used to estimate the power and sample size to detect a difference of about 3 ppm in the mean PCB levels for the two genders. This gives a TOTAL sample size of about 13, which when split over the 6 treatment combinations gives about 2 per species-gender combination. On the other hand, when the power-sample size platform is used to estimate the power and sample size to detect a difference of about 2 in the mean PCB levels among the three species, you find that a total sample size of about 34 is needed, which is about 6 per species-gender combination. [You will have to specify a configuration for the three means - the actual values used are unimportant as long as the difference is 2. I used 0, 3, and 2 ppm]. The two objectives are in conflict so either the larger sample size should be used, or the detectable difference should be adjusted for differences in the means among species.

It is possible to determine sample size and power to detect interactions - this is rarely done and so will not be explored further.

It is not necessary to have the sample sizes equal in all treatment groups, but it can be shown that the 'power' of the test is maximized when the sample sizes are equal for all treatment combinations.

9.2.4 Unbalanced data - Introduction

Unbalanced data can take many forms. Some of the forms are easy to analyze, some are difficult.

Here are the some illustrations of the common replication patterns that you will run into. In all cases there are 2 levels of Factor A and three levels of Factor B and an ‘x’ represents a replicate.

- **Equal replications per cell**

```
                 Factor B
               b1    b2    b3
            +-----+-----+-----+
Factor A  a1 | xx  | xx  | xx  |
            +-----+-----+-----+
          a2 | xx  | xx  | xx  |
            +-----+-----+-----+
```

  This is the easiest to deal with and two examples were given earlier in the notes.

- **Unequal replications per cell, but replicates in every cell**

```
                 Factor B
               b1    b2    b3
            +-----+-----+-----+
Factor A  a1 | xxx | xx  | xxx |
            +-----+-----+-----+
          a2 | xx  | xx  | xxxx|
            +-----+-----+-----+
```

  In this case, all cell have some data, but the number of replicates differs among cells and every cell has 2 or more replicates. The multiple replicates within a cell are needed to estimate the MSE row in the ANOVA table. An example of an analysis of this type of data will be given below. Because each cell has replicates, it is possible to check that the variation is roughly equal in all treatment groups. This type of unbalance can be analyzed “easily” if the computer package has been programmed correctly. **BEWARE: some packages (e.g. Excel) will give WRONG answers!**

- **Unequal replications per cell, with some cells having only a single observation**

```
                  Factor B
                b1    b2    b3
             +-----+-----+-----+
Factor A  a1 | x   | xx  | xxx |
             +-----+-----+-----+
          a2 | xx  | x   | xx  |
             +-----+-----+-----+
```

  In this case, all cell have some data, but the number of replicates differs among cells and some cells have only a single observation. Theoretically, there is no difference in the analysis of this experiment from the previous example. However, in this experiment, you must assume that the variability in the cells with replicates is an accurate representation of that in cells with only a single observation.

- **One observation per cell**

```
                  Factor B
                b1    b2    b3
             +-----+-----+-----+
Factor A  a1 | x   | x   | x   |
             +-----+-----+-----+
          a2 | x   | x   | x   |
             +-----+-----+-----+
```

  If you only have a single observation per cell, it is impossible to test for interaction effects. [Technically, the design has insufficient degrees of freedom for error.] We won't discuss how to analyze this type of data in this course, but basically you **MUST ASSUME** that no interaction exists, and fit a model without any interaction terms in the model. Only main effects can be tested.

- **One or more cells completely empty**

```
                 Factor B
               b1    b2    b3
             +-----+-----+-----+
Factor A  a1 | xx  | xx  |     |
             +-----+-----+-----+
          a2 | xx  | xx  | xx  |
             +-----+-----+-----+
```

  SEEK HELP! Most computer packages will give you completely WRONG results! This is tough problem. Now having said this, one simple solution to this problem (if it only occurs in one cell of the design), is to drop the column and analyze the remaining data as a two-factor, each at two levels. In the above example, level *b3* would be dropped from the experiment. However if there are many missing cells, you may find that you are dropping most of your data!

The analysis of an unbalanced design with all cells having at least one observation and some cells having at least two replicates is discussed in:

Shaw, R.G. and Mitchell-Olds, T. (1993).
ANOVA for unbalanced data:an overview.
Ecology, 74, 1638-1645.

This is available in the library or from JSTOR by following this link[10]

[We will NOT be discussing this article in class and it is not part of the course.]

9.2.5 Unbalanced data - Example - Energy consumption in pocket mice

Here is an example showing some of the problems that you may run into when analyzing unbalanced data.

French (1976, *Selection of high temperature for hibernation by the pocket mouse: Ecological advantages and energetic consequences* Ecology, 57, 185-191)

[10] http://links.jstor.org/sici?sici=0012-9658%28199309%2974%3A6%3C1638%3AAFUDAO%3E2.0.CO%3B2-G

collected the following data on the energy utilization of the pocket mouse (*Perognathus longimembris*) during hibernation at different temperatures:

Restricted Food		Ad libitum food	
8C	18C	8C	18C
62.69	72.60	95.73	101.19
54.07	70.97	63.95	76.88
65.73	74.32	144.30	74.08
62.98	53.02	144.30	81.40
	46.22		66.58
	59.10		84.38
	61.79		118.95
	61.89		118.95
	62.50		

All readings are in kcal/g.

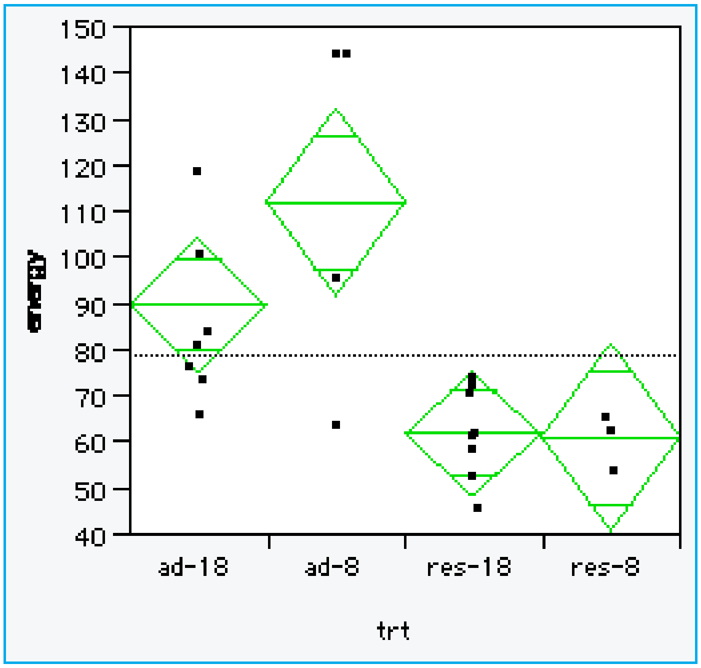
Design issues

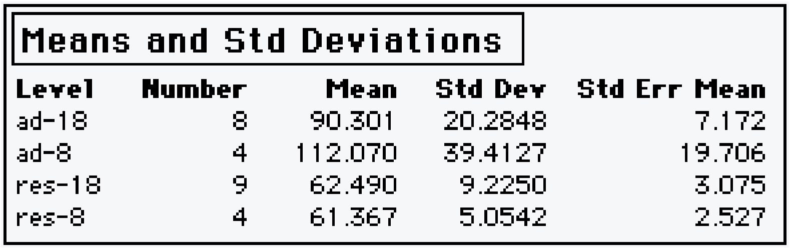
What are the factors in this experiment? Their levels? What is the response variable? Is this an "experiment" or an observational study? If the latter, what is the role of randomization in this study? What is the treatment structure in this experiment? What is the experimental unit structure? What are the experimental and observation units? Why is the design unbalanced? What unbalanced the design, or did it occur "by chance"? What is the randomization structure?

Are the factors to be considered *fixed effects*? Hence, does this experiment appear to satisfy the requirements for a two-factor fixed-effects CRD?

Preliminary summary statistics

Create some simple summary statistics. We will create a 'pseudo-factor' in JMP so that we can get the statistics on each group and make a simple plot.





Means and Std Deviations

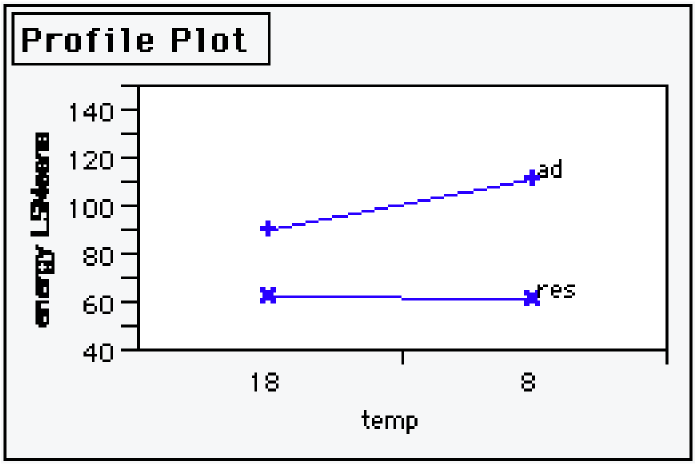
Level	Number	Mean	Std Dev	Std Err Mean
ad-18	8	90.301	20.2848	7.172
ad-8	4	112.070	39.4127	19.706
res-18	9	62.490	9.2250	3.075
res-8	4	61.367	5.0542	2.527

There doesn't appear to be any difference in the mean energy usage between the two temperature levels under the restricted food diet, but both appear to be less than the mean energy usage from the *ad libitum* groups.

The standard deviations appear to be quite different! This is very worrisome - the ANOVA method is fairly robust to unequal variances **provided the sample sizes are equal in all groups**. I would proceed with caution in the subsequent analysis!

The design is unbalanced (the sample sizes are not equal in all groups).

The profile plot is below:



The profile plot shows that some interaction may be present. Indeed, this is not unexpected given the earlier plot that showed no difference in the means between the two temperatures under the restricted diet but some apparent differences under the *ad libitum* diet. Because of the strong evidence of differential standard deviations among the groups, we may not detect this effect.

The statistical model

The model for this study is:

$$energy = food \quad temp \quad food * temp$$

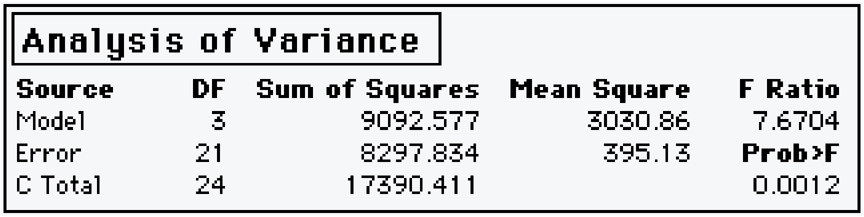
What does this statistical model tell us about the sources of variation in the observed data?

Fitting the model

We use the *Analyze->Fit Model* platform of JMP and complete the dialogue entries as in the previous example (we won't show it for this example). Again be sure that all factors have nominal scale and that the response variable has continuous scale.

Hypothesis testing and estimation

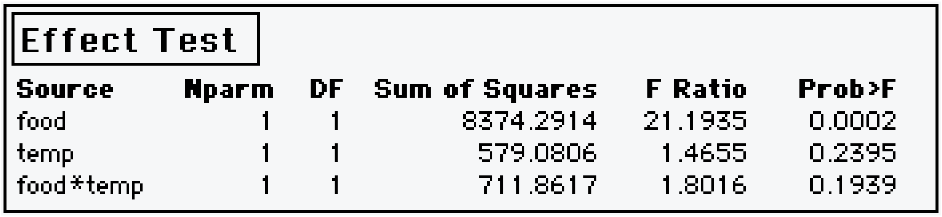
Here are the ANOVA tables from fitting the model.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	9092.577	3030.86	7.6704
Error	21	8297.834	395.13	**Prob>F**
C Total	24	17390.411		0.0012

The **Whole Model** test above simply examines if there are any statistically significant effects anywhere in the experiment. It is rarely useful but does tell us that we should proceed further and investigate the various effects.



Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
food	1	1	8374.2914	21.1935	0.0002
temp	1	1	579.0806	1.4655	0.2395
food*temp	1	1	711.8617	1.8016	0.1939

This table breaks down the whole model test according to the various terms in the model.

This is where the unbalance causes some problems. There is still controversy among statisticians on exactly how to compute the sums of squares and F-statistic s for unbalanced data. A majority of us believe that “Type III” or “marginal” sums-of-squares are appropriate (which are presented in JMP). Arguments can be made for “Type II” or “model effect” sums-of-squares (not presented in JMP but present in SAS). We won’t delve deeper into this controversy - it has been known to provoke heated-debates and “knuckle-sandwiches” among statisticians (and you thought that we were a dull lot!).

As in the balanced case, start the hypothesis testing with the most complicated terms and work towards simpler terms. In this case, start with the interaction effects.

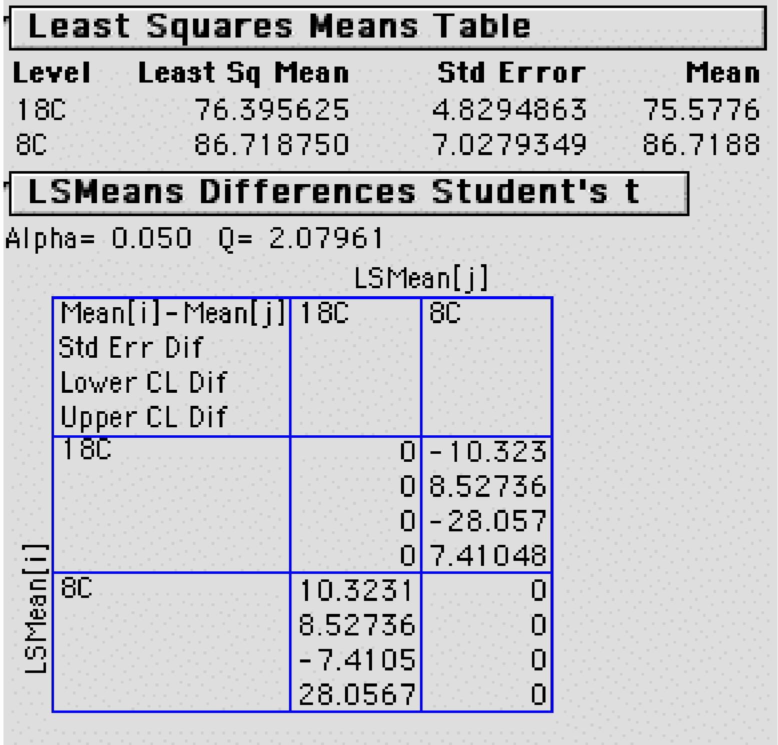
Our null hypothesis is
H: no interaction among the effects of food and temperature on the mean response
A: some interaction among the effects of food and temperature on the mean response.

Our test statistic is F=1.80, the p-value (.1939) is not very small. Hence there is no evidence of an interaction among the effects of food and temperature on the mean response. This is somewhat surprising given the profile plots constructed earlier - it may be an artifact of the unequal standard deviations among the groups or because our sample sizes are not very large.

Examining main effects - temperature.

The F-statistic is 1.47, the p-value is 0.2395. There is no evidence of a difference among the mean energy requirements at the two temperature levels.

This would be confirmed by looking at the estimated marginal means and the estimated difference in the means:



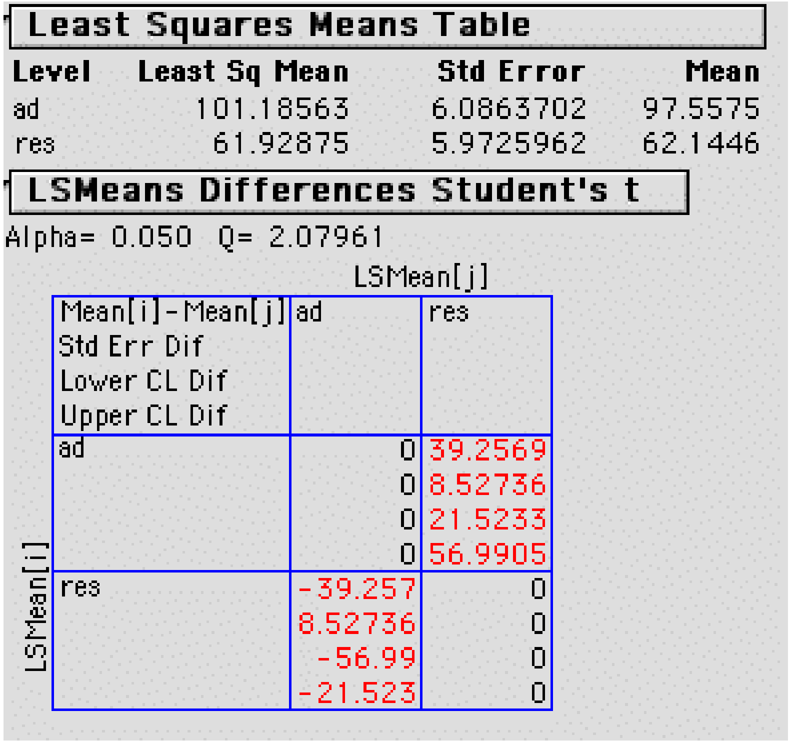
```
Least Squares Means Table
Level    Least Sq Mean     Std Error      Mean
18C          76.395625     4.8294863   75.5776
8C           86.718750     7.0279349   86.7188

LSMeans Differences Student's t
Alpha= 0.050  Q= 2.07961
                          LSMean[j]
         Mean[i]-Mean[j] 18C       8C
         Std Err Dif
         Lower CL Dif
         Upper CL Dif
         18C                     0 -10.323
                                 0 8.52736
                                 0 -28.057
                                 0 7.41048
         8C              10.3231         0
                         8.52736         0
                         -7.4105         0
                         28.0567         0
LSMean[i]
```

Here is where the unbalance again causes some difficulty in the analysis. Notice that the LSMeans no longer equal the raw means. In unbalanced data, simple means across the factors may be affected by unequal sample sizes. The simple mean for the 8C levels would include 4 mice at the restricted diet and 4 mice at the *ad libitum* diet - an equal split. However, the simple mean for the 18C diet would include 9 mice at the restricted diet and 10 mice under the *ad libitum* diet - no longer an equal weighting among the two diets. The LSMeans are computed by giving equal weights to each of the two means from the two diets. There is no universal agreement upon this (and you thought that Statistics was so cut and dried) but, in most situations, the least square means seem preferable.

Examining main effects - food level.

The ANOVA table gives F=21.2 and the *p*-value is 0.0002 - very small. There is very strong evidence of a difference in the mean energy requirements between the two food levels. Because we only have two levels, it is obvious where the difference lies. But we find the Least Square Means and estimated differences for the food levels:



Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
ad	101.18563	6.0863702	97.5575
res	61.92875	5.9725962	62.1446

LSMeans Differences Student's t

Alpha= 0.050 Q= 2.07961

LSMean[j]

LSMean[i]	Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	ad	res
	ad	0 0 0 0	39.2569 8.52736 21.5233 56.9905
	res	-39.257 8.52736 -56.99 -21.523	0 0 0 0

Again, notice that the least square means are **different** than the raw means which are shown in the last column. This will usually happen when the design is unbalanced. The least square means give equal weight to the two temperature levels while the raw means weight the two temperature levels according to the observed sample sizes.

Power and sample size

We did detect an effect in the mean between the two food levels but not between the means for the two temperature levels.

The *DOE->PowerSampleSize* can be used to assess the power of this design and determine what sample size would be needed to be 80% confident of detecting such a difference in the future.

Analysis using SAS

Here are links to a sample SAS program[11] and its output[12] You should find that most of what we found above also appears somewhere in the SAS output - except for power calculations which are not easily done in SAS except under INSIGHT.

Adjusting for unequal variances?

This is beyond the scope of this course, but a formal test for unequal variances showed clear evidence of a problem. A more exact test, fortunately, came to similar conclusions, but the estimated standard error of the difference is slightly different.

9.3 Completely randomized design random and fixed effects

As noted below, the analysis of these **mixed models** is fraught with difficulty and great care must be taken. In many cases, computer packages may give wrong or mis-leading results without any warning!

Recall the criteria for determining if a factor is a fixed or random effect. A factor is a **fixed effect** if :

- the same levels would be used if the experiment were to be repeated;
- inference will be limited ONLY to the levels used in the experiment;

A factor is a **random effect** if:

- the levels were chosen at random from a larger set of levels

[11] http:../MyPrograms/mouse.sas

[12] http:../MyPrograms/mouse.lst

- new levels would be chosen if the experiment were to be repeated
- inference is about the entire set of potential levels - not just the levels chosen in the experiment.

9.3.1 Example - Rancid fat - Fixed and random effects

A study was conducted to investigate the effects of irradiating fat with gamma radiation to prevent it from going rancid. [This is a proposed treatment for many foods to kill many of the bacteria which cause foods to spoil. The promoters of this treatment claim that it doesn't affect taste or nutrition, and is perfectly safe.]

This experiment was a collaboration between two laboratories. In this experiment, 12 batches of fat were obtained and split between the 2 laboratories. Six of the samples were irradiated at each laboratory. In each lab, 6 rats (all aged 30 to 34 days) were obtained, and six rats were assigned at random to each of the fat groups.

The rats were allowed to feed *ad libitum* and the total consumption of fat (grams) was noted over 73 days.

Here are the raw data:

Consumption	Fat type	Laboratory
709	control	1
679	control	1
699	control	1
562	treated	1
518	treated	1
496	treated	1
657	control	2
594	control	2
677	control	2
508	treated	2
505	treated	2
539	treated	2

Design issues

What are the factors in this experiment? Their levels? What is the response variable? What is the treatment structure in this experiment?

Now consider the two factors in more detail. In this case, the treatment is clearly of interest - we really want to know if the mean amount consumed is the same - presumably if it is, then the two types of fat are equally palatable (at least to rats). This will be a fixed effect.

However, the laboratory factor is quite different. We likely are not interested in a specific comparison between these two laboratories - rather we can visualize that these two laboratories were selected at random from all possible laboratories and we really want to know if there is any evidence that the results would change if new laboratories were chosen. As well, we may decide to repeat the experiment using a different set of laboratories in the future, but we would keep the treatment of the fats fixed. For this reason, we would consider the laboratories factor as a **random effect**.

What is the experimental unit structure? What are the experimental and observation units? In this case, the experimental unit is a rat and the observation unit is a rat.

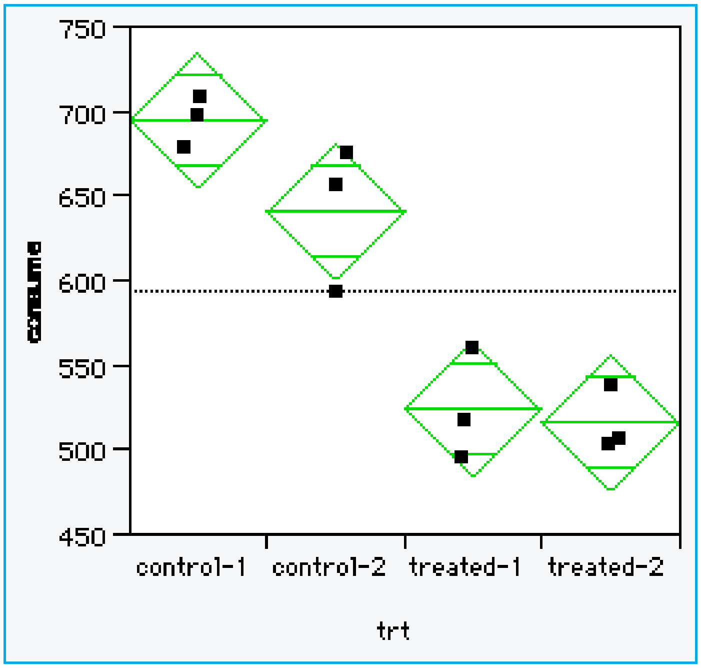
What is the randomization structure? Here there are 12 separate batches of fat, randomly assigned to the two treatment groups and to the two laboratories and to the rats.

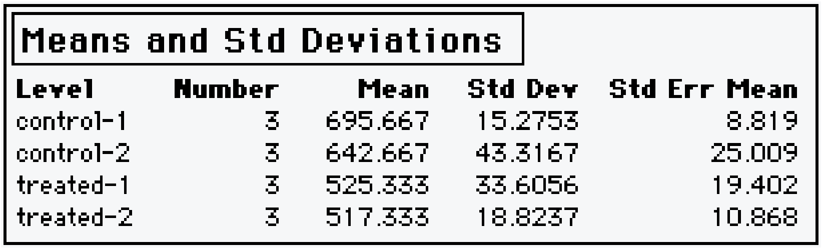
Notice how the experiment would differ if there were only 2 batches of fat, one of which was assigned to lab 1 and the other to lab 2. In this case, there would be complete confounding of batches of fat with laboratories. As well, there would not be a complete randomization, and the design would be more akin to a relative of a blocked design. **I can't emphasize too strongly the importance of carefully specifying how an experiment is done before trying to analyze the results of the experiment.**

Hence, this experiment appears to satisfy the requirements for a two-factor CRD, albeit with one factor fixed and the other factor random.

Preliminary summary statistics

Create some simple summary statistics. We will create a 'pseudo-factor' in JMP so that we can get the statistics on each group and make a simple plot.





Means and Std Deviations

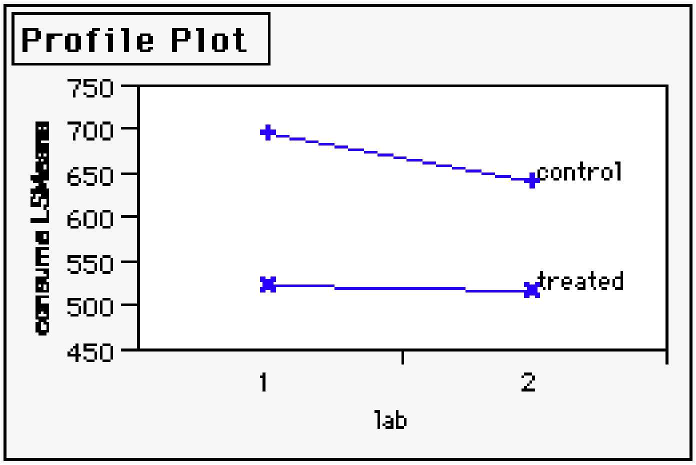
Level	Number	Mean	Std Dev	Std Err Mean
control-1	3	695.667	15.2753	8.819
control-2	3	642.667	43.3167	25.009
treated-1	3	525.333	33.6056	19.402
treated-2	3	517.333	18.8237	10.868

The rats eating treated fat appear to consume less, on average, than the rats eating untreated fat. There doesn't appear to be much of a difference in the means between the two laboratories.

The standard deviations appear to be roughly equal in all groups. [Given that the total sample size is 3 in each group, it is very difficult to detect any real differences in the standard deviations.]

The design is balanced (the sample sizes are equal in all groups).

The profile plot is below:



The lines don't appear to be parallel so there may be some evidence of an interaction - however, given that there are only 3 observations per group the precision of each mean is terrible and the lines could, in fact, be parallel.

The statistical model

The model for this study is constructed in a similar fashion as before. There will be terms for each main effect of the factors and their interaction. Because there is only one size of experimental unit, it is not necessary to specify anything. Similarly, complete randomization is implicitly assumed. The overall model is:

$$consume = radiation \quad lab(R) \quad radiation * lab(R)$$

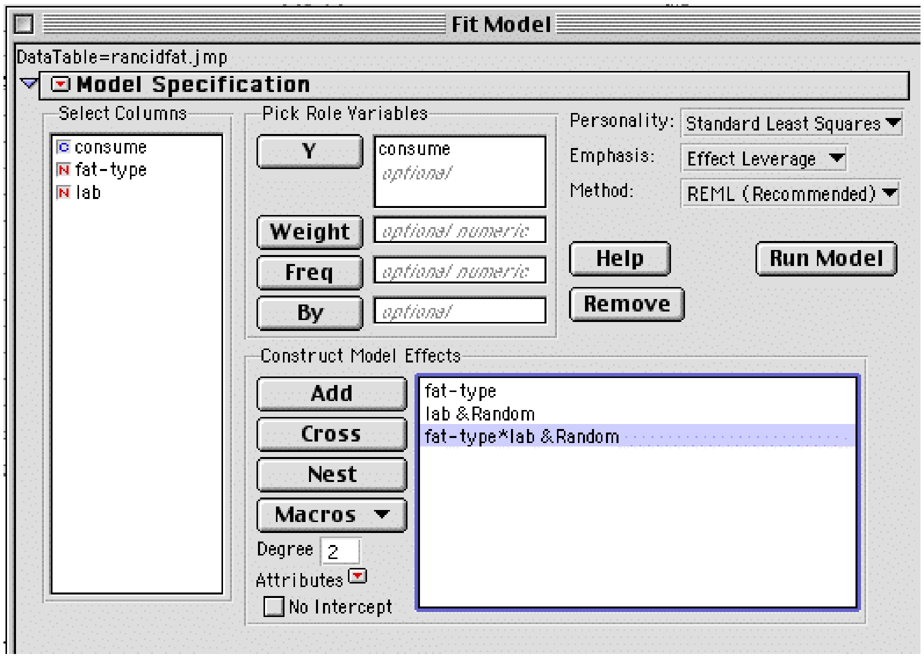
where the (R) indicates a random effect.

Because we believe that one of the factors is a random effect we add a further two assumptions to the model. For each random effect and any interaction involving a random effect, we assume that these are normally distributed with mean 0 and an associated variance component.

So in this model, there are three sources of random variation - experimental units (the rats), the laboratories, and the interaction between the laboratories and the treatments.

Fitting the model

We use the *Analyze->Fit Model* platform of JMP and complete the dialogue entries as shown below. The factors must be specified as nominal scale; the response variable as a continuous scale variable. The effects and their interaction are entered as before using the *ADD* and *CROSS* buttons. The *& random* attribute is added by clicking on a term in the model effects box and using the *Attribute* pop-down menu to select a random effect. **DON'T FORGET TO SPECIFY THE RANDOM EFFECTS!** using the **Effect Attributes** pop-up menu. **All interactions that include a random effect must also be specified as random effects.**



DataTable=rancidfat.jmp

Y: consume

Personality: Standard Least Squares

Emphasis: Effect Leverage

Method: REML (Recommended)

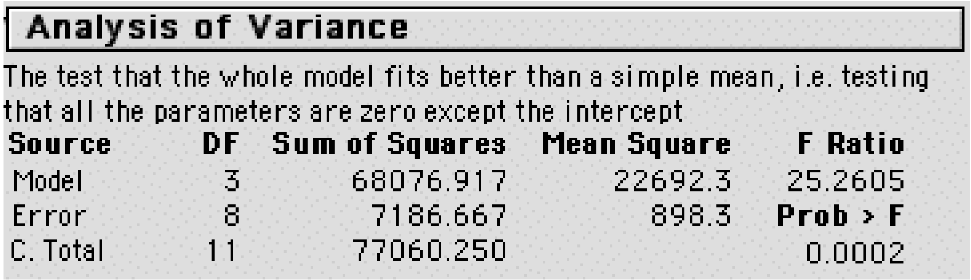
Degree 2

fat-type
lab &Random
fat-type*lab &Random

Notice that the *METHOD* selector has two options: REML (Restricted maximum likelihood estimation) which is a newly developed method for random and mixed models; and EMS - which is likely the method shown in textbooks. The two methods will give similar results in many cases, but the REML method is more comprehensive, more flexible, and less prone to error on the part of the analyst.

Hypothesis testing and estimation

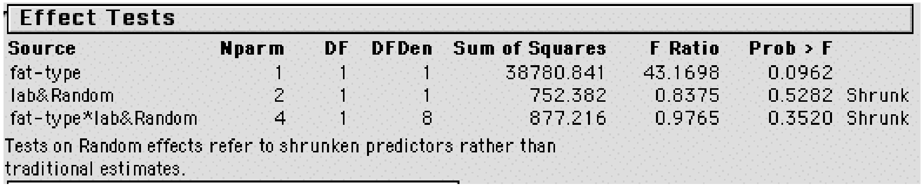
The summary ANOVA table looks very similar to what we've seen before



```
Analysis of Variance
The test that the whole model fits better than a simple mean, i.e. testing
that all the parameters are zero except the intercept
Source      DF   Sum of Squares   Mean Square    F Ratio
Model        3        68076.917       22692.3    25.2605
Error        8         7186.667         898.3   Prob > F
C. Total    11        77060.250                   0.0002
```

The **Whole Model** test simply examines if there are any statistically significant effects anywhere in the experiment. It is rarely useful but does tell us that we should proceed further and investigate the various effects.

The effect tests for the individual effects in the model are constructed differently and the output looks different:



```
Effect Tests
Source                  Nparm   DF  DFDen  Sum of Squares   F Ratio   Prob > F
fat-type                    1    1      1       38780.841   43.1698     0.0962
lab&Random                  2    1      1         752.382    0.8375     0.5282  Shrunk
fat-type*lab&Random         4    1      8         877.216    0.9765     0.3520  Shrunk
Tests on Random effects refer to shrunken predictors rather than
traditional estimates.
```

CAUTION- This is where some packages make mistakes. For example, PROC GLM in SAS always tests everything against MSE unless you explicitly state otherwise. This is not correct. JMP also includes adjustments (called shrinking) for random effects that are just starting to make their way into the mainstream statistical packages. What do they test? These tests are valid tests based on the Henderson (1984) model framework, but for a statistical hypothesis, that is not interesting: that the effect sizes of the levels you randomly drew from the population happened to all be zero. In this case, these test if these particular laboratories have any effect upon the mean response. However, because you are interested in the population of levels, rather than just in the levels you happened to draw, you should be looking at the variance component instead, with its confidence interval. This is beyond the scope of this course. These shrunken effect tests on random effects will always be smaller and less significant than the old style tests often found in textbooks.

We again start hypothesis testing with the most complicated terms and work towards simpler terms.

In this example, we start with the interaction effects.

Our null hypothesis is
H: no interaction among the effects of treatment and these particular laboratories in the mean amount eaten
A: some interaction among the effects of treatment and these particular laboratories in the mean amount eaten.

Our test statistic is F=0.976, the *p*-value (.352) is not very small. Hence there is no evidence of an interaction among the effects of treatment and these particular laboratories on the mean amount eaten.

This is somewhat surprising given the profile plots constructed earlier - however, I suspect that the poor precision of each mean (after all there are only three observations in each group) means it is difficult to detect any interaction.

Examining main effects - laboratories

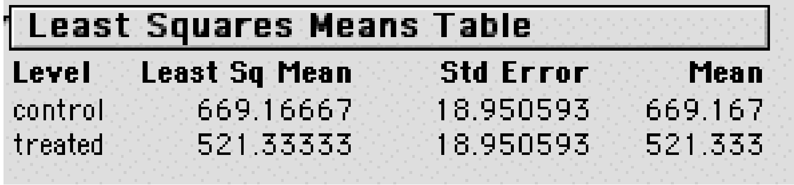
We find the F-statistic to be 0.838 with a *p*-value of 0.528. Now because this is a random effect, we are saying that there is no evidence of a difference in the mean amount consumed among these particular laboratories.

It is not customary to examine the actual LSMeans for random effects. Rather, interest lies in an estimate of the variation among labs - this is not covered in this course. [As noted earlier, this information is available from the REML Variance Component Estimates.]

Examining main effects - treatment

The ANOVA table gives F=43.1 and the *p*-value 0.0962 - some evidence of a difference in the mean consumption between the two treatment of the fat.

As this was a fixed effect, it makes sense to examine the LSMeans:

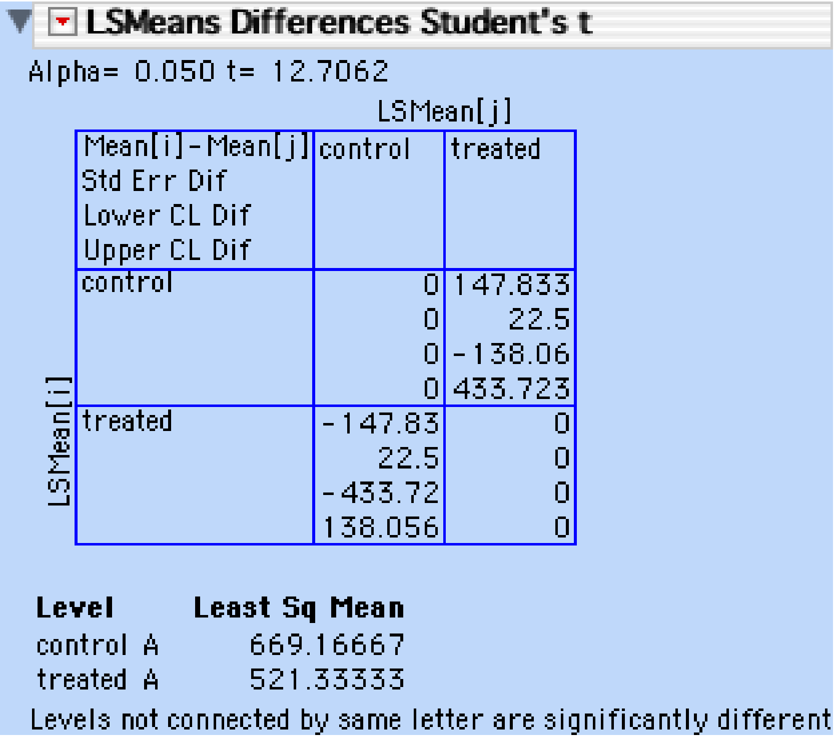


Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
control	669.16667	18.950593	669.167
treated	521.33333	18.950593	521.333

MANY PACKAGES give wrong SEs for marginal means when random effects are present. JMP v3.0 had problems; these were corrected in v4.0 and later. SAS's PROC GLM computes the *se* incorrectly; but computes the proper *se* in PROC Mixed.

Estimates of the difference, its *se*, and confidence intervals are found in the usual fashion.



```
LSMeans Differences Student's t
Alpha= 0.050 t= 12.7062
                                    LSMean[j]
            Mean[i]-Mean[j]  control    treated
            Std Err Dif
            Lower CL Dif
            Upper CL Dif
            control                0    147.833
                                   0       22.5
                                   0    -138.06
                                   0    433.723
LSMean[i]   treated          -147.83          0
                                22.5          0
                             -433.72          0
                             138.056          0

Level       Least Sq Mean
control  A      669.16667
treated  A      521.33333
Levels not connected by same letter are significantly different
```

Notice that here the results are a bit counterintuitive. The estimated difference is about 150 g with a standard error of about 23 g. The usual rule of estimate ± 2 standard errors would seem to give a confidence interval that does not cover zero which is contradictory to the p-value of 0.09 seen earlier. This is a case where the usual "rule-of-thumb" doesn't hold - the problem is that you really only have a sample size of 2 - the two laboratories! The reported confidence interval is correct and shows that the estimated effect may be zero. [13]

MANY PACKAGES BOTCH THIS COMPUTATION. You must be very, very careful that your package does this computation correctly - even SAS's PROC GLM does this wrong (but does it correctly in PROC MIXED).

This all goes to prove that 'To Err is Human, but it really takes a computer to foul things up.'

[13] Earlier versions of JMP computed incorrect confidence intervals. The above output is from JMP 4.0.5 or higher 5.0 and is correct.

Analysis using SAS

Here is a link to a sample SAS program[14] and output[15]. You should find that most of what we found above also appears somewhere in the SAS output. Note that because some of the output from GLM is NOT CORRECT when you have multiple error terms, you should be using PROC MIXED to obtain correct output.

9.3.2 Example - Mosquito repellent - Fixed and random effects

[Based upon a real study conducted at the University of Manitoba but using simulated data.]. Biting insects can be a real pest and a health hazard - e.g. malaria and equine encephilitus are serious diseases transmitted by mosquitoes. What is the best method of deterring these pesky critters from biting?

There are a number of insect repellents available on the market place. Some use the chemical DEET which is quite effective but many people are reluctant to use these sprays because they are quite 'strong' - e.g., many sprays containing DEET will soften paint. As an alternative, there is a strong 'urban legend' about an Avon (a perfume and toiletry company) product called 'Skin so soft' that many people claim is also an effective repellent.

To investigate these claims, twenty four volunteers were recruited. These were randomly assigned to 3 groups of 8 people which then went to 3 locations on the University Campus. At each location, half of the volunteers spread a DEET product on their right arm; the other half used the Avon product on their right arm. Each subject stood at least 10 m from any other subject. Then the subjects let mosquitos bite their exposed arm, and after 15 minutes, the total number and severity of the bites was scored using a standard scale for such studies (how some one came up this scale I can only hazard a guess!). The higher the score, the worse the biting experience.

Here are the raw data:

[14] ../MyPrograms/rancidfat.sas

[15] ../MyPrograms/rancidfat.lst

score	product	location
21	deet	1
19	deet	1
20	deet	1
22	deet	1
14	sss	1
15	sss	1
13	sss	1
16	sss	1
14	deet	2
17	deet	2
15	deet	2
17	deet	2
12	sss	2
11	sss	2
12	sss	2
14	sss	2
16	deet	3
20	deet	3
18	deet	3
19	deet	3
14	sss	3
14	sss	3
14	sss	3
12	sss	3

Design issues

What are the factors in this experiment? Their levels? What is the response variable? What is the treatment structure in this experiment?

There are two factors - the product used and the location where the subjects were ‘subjected’ to mosquito bites. As in the previous example, the level of interest in the two factors is quite different. The product is clearly a fixed effect - we are interested in these specific levels. The location effect is a random effect - these levels are chosen, in some sense, from all possible sites in Winnipeg, and we would like to make inferences to all the possible locations in Winnipeg.

What is the experimental unit structure? What are the experimental and observation units?

In this case, the experimental unit is a person and the observational unit is a person.

What is the randomization structure?

Here there are 24 separate people that could be randomized completely to the locations and products.

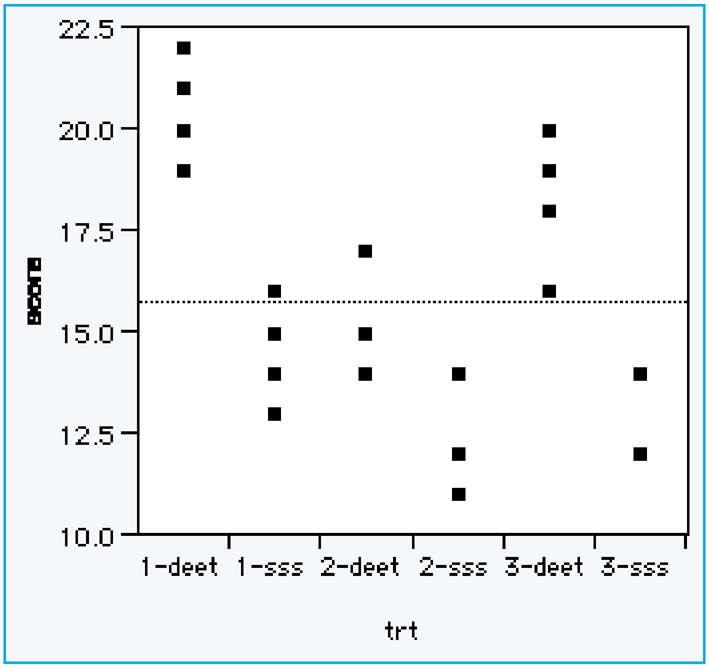
Hence, this experiment appears to satisfy the requirements for a two-factor CRD, albeit with one factor fixed and the other factor random.

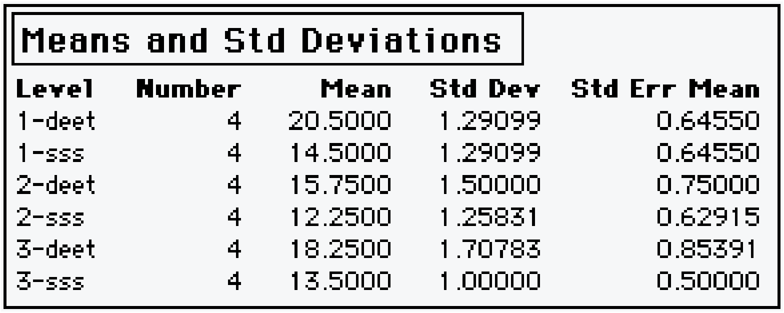
In general, factors like 'location', or 'site' are usually random effects. However, in many cases, the designs are not really CRD's as in this example. Here, the experimental units could be randomly assigned to the locations. This is not always true and it is very easy to fall into a common trap. Consider an experiment where locations along the the west coast of Vancouver Island are to be surveyed, and at each location, 4 patches of beach are found, covered with oil to simulate an oil spill, and then two different clean up methods applied. On the surface, this looks very similar to the previous experiment - locations, multiple patches of beach at each location (which is like multiple subjects), and two different treatments applied. Yet, you **COULD NOT RANDOMIZE** patches of beach to locations - they are permanently affixed to each location. This latter example is NOT a CRD! It is a relative of a blocked design where locations are blocks, and experimental units within blocks are randomly assigned to the treatment combinations (in this case each treatment would appear twice in each block).

I can't emphasize too strongly that you must be very careful when dealing with 'location' or 'time' as a factor - in many cases the design is NOT a CRD.

Preliminary summary statistics

We will create a 'pseudo-factor' in JMP so that we can get the statistics on each group and make a simple plot.





Means and Std Deviations

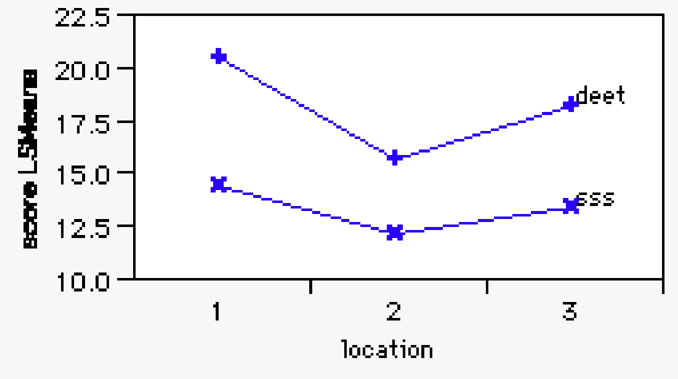
Level	Number	Mean	Std Dev	Std Err Mean
1-deet	4	20.5000	1.29099	0.64550
1-sss	4	14.5000	1.29099	0.64550
2-deet	4	15.7500	1.50000	0.75000
2-sss	4	12.2500	1.25831	0.62915
3-deet	4	18.2500	1.70783	0.85391
3-sss	4	13.5000	1.00000	0.50000

The SSS appears to work at least (or perhaps even better) than the DEET formulations.

The standard deviations appear to be roughly equal in all groups.

The design is balanced (the sample sizes are equal in all groups).

The profile plot is below:



The lines appears to be parallel, so we won't be too surprised if we don't detect any interaction.

The statistical model

The statistical model for this study will contain terms corresponding to the main effects and interactions of the factors (some of which will be random effects). Because there is only size of experimental unit and because randomization was complete, no terms will be explicitly given and these contributions are implicit. The model is:

$$score = product \;\; location(R) \;\; product * location(R)$$

Because we believe that one of the factor is a random effect we add a further two assumptions to the model. For each random effect and any interaction involving a random effect, we assume that these are normally distributed with mean 0 and an associated variance component.

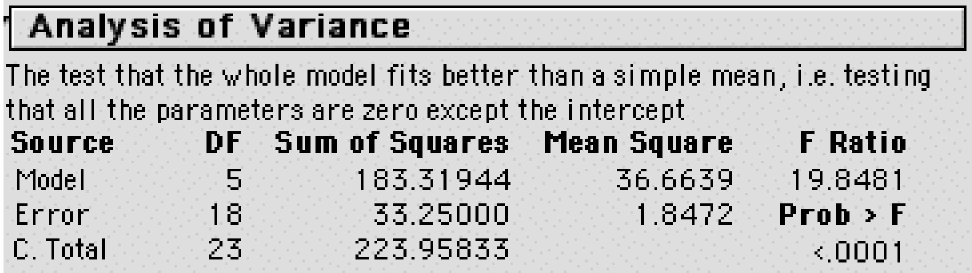
So in this model, there are three sources of variation - residual error (corresponding to the experimental units of people), the locations, and the interaction between the locations and the products.

Fitting the model

We use the *Analyze->Fit Model* platform of JMP and specify the effects as we did in past examples. **DON'T FORGET TO SPECIFY THE RANDOM EFFECTS!** using the **Effect Attributes** pop-up menu, and don't forget to specify that both the location and the location by product interaction terms are random effects.

Hypothesis testing and estimation

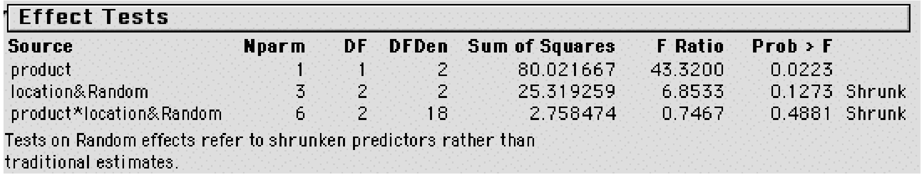
The overall ANOVA table is below:



```
Analysis of Variance
The test that the whole model fits better than a simple mean, i.e. testing
that all the parameters are zero except the intercept
Source      DF   Sum of Squares   Mean Square     F Ratio
Model        5       183.31944        36.6639     19.8481
Error       18        33.25000         1.8472    Prob > F
C. Total    23       223.95833                    <.0001
```

As before, this simply tells us that either products or locations or both has a statistically significant effect upon the mean score.

The individual effect tests must be examined:



```
Effect Tests
Source                      Nparm   DF  DFDen  Sum of Squares   F Ratio   Prob > F
product                         1    1      2       80.021667   43.3200     0.0223
location&Random                 3    2      2       25.319259    6.8533     0.1273  Shrunk
product*location&Random         6    2     18        2.758474    0.7467     0.4881  Shrunk
Tests on Random effects refer to shrunken predictors rather than
traditional estimates.
```

CAUTION- some packages will give you results that are WRONG!

We start the hypothesis testing with the most complicated terms and work towards simpler terms. Again note that many of the tests are not of particular interest.

We start with the interaction effects.

Our null hypothesis is
H: no interaction among the effects of product and these particular locations upon the mean score
A: some interaction among the effects of product and these particular locations upon the mean score.

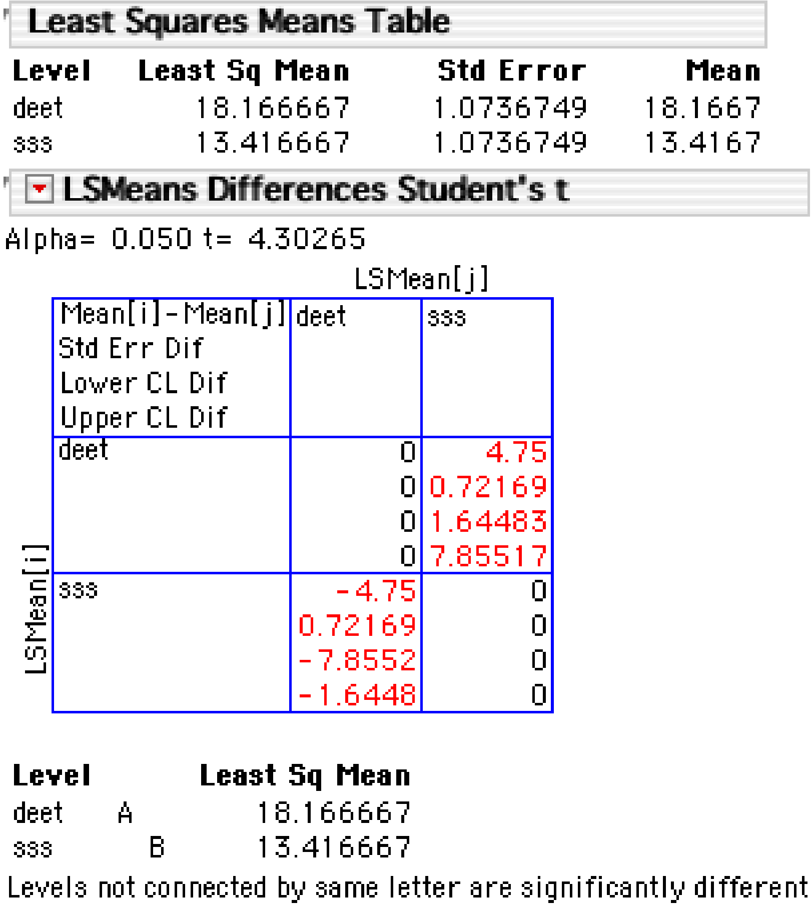
Our test statistic is F=.747, the *p*-value (.48) is not very small. Hence there is no evidence of an interaction among the effects of product and these particular locations on the mean score.

Examining main effects - locations

We find the F-statistic to be 6.85 with a p-value of 0.13 Now because this is a random effect, we are saying that there is no evidence of a difference in the mean score among these particular locations in Winnipeg. Note that if we wish to generalize to all locations in Winnipeg, we would look at the variance components which is beyond the scope of this course.

Examining main effects - product

The ANOVA table gives F=43.3 and the p-value 0.0223 - strong evidence of a difference in the mean score between the two products. We also estimate the marginal means, the estimated difference between the mean, and the se of the difference.



Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
deet	18.166667	1.0736749	18.1667
sss	13.416667	1.0736749	13.4167

LSMeans Differences Student's t

Alpha= 0.050 t= 4.30265

LSMean[j]

Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	deet	sss
deet	0 0 0 0	4.75 0.72169 1.64483 7.85517
sss	-4.75 0.72169 -7.8552 -1.6448	0 0 0 0

LSMean[i]

Level		Least Sq Mean
deet	A	18.166667
sss	B	13.416667

Levels not connected by same letter are significantly different

MANY PACKAGES may give incorrect *se*.

Analysis using SAS

Here is a link to a sample SAS program[16] and output[17]. You should find that most of what we found above also appears somewhere in the SAS output. Note that GLM should not be used when there are random and mixed effects.

9.3.3 Power and sample size

CAUTION - computing power and sample size when there is more than one random term in a model (i.e. a mixed model) is difficult. You CANNOT use the fixed effect tables directly even for the fixed effects in the model, and the power tables for the random effects are completely different. You will need to seek some professional help in these cases.

JMP v4.0 sample size/power determinations are for models with fixed effects ONLY and are not appropriate.

9.3.4 All effects random

It is possible to consider models where all effects are random.

The analysis proceeds in a similar fashion, i.e., you need to specify that all effects are random, then have the computer compute the test statistics. Many packages botch these computations.

However, such studies are very rare in ecology and we won't cover them in this course.

9.3.5 Unbalanced data with random effects

THESE ARE VERY DIFFICULT STUDIES TO ANALYZE PROPERLY

[16] ../MyPrograms/mosquito.sas

[17] ../MyPrograms/mosquito.lst

If you have such a study, seek very good help - even seasoned statisticians who do not have experience with these types of studies will get them wrong.

Even worse, most computer packages are hopeless with these types of studies - only recently has PROC MIXED and JMP from SAS been developed to analyze unbalanced data sets reasonably well.

9.3.6 Scheffe vs. Searle formulations

It turns out that there are two alternate formulations for a model that involves fixed and random effects. These are generically known as the Scheffe and Searle formulations and are a source of controversy among statisticians even today. Unfortunately, these two formulations compute the test statistics in slightly different ways leading to potentially different test-statistics which is, to say the least, annoying. **BE VERY, VERY, VERY CAREFUL** with mixed effect models! These are hard things to fit even for seasoned statisticians.

An introduction to the problems encountered with mixed models is found in Schwarz, C.J. (1993). The Mixed-Model ANOVA: The Truth, the Computer Packages, and the Books. Part I: Balanced Data. The American Statistician, 47, 48-59.

9.4 Blocking in two-factor CRD designs

This twist on the design poses no real problems. Things to look out for:

- the randomization within each block is done independently of every other block
- the blocks must be complete, i.e., every treatment combination must appear in every block.
- when analyzing the data, specify factors as fixed or random as well as blocks as fixed or random.
- as before, random blocks or random factors imply that some packages may give you incorrect results.
- random blocks or random factors imply that some of the packages will have difficult in estimating the *se* of the marginal means and of the contrasts among the means.

An example of such a design will be done in an assignment.

9.5 FAQ

9.5.1 How to determine sample size in 2 factor designs

How are sample sizes determined in 2 factor experiments?

In a later chapter, you will analyze in detail an experiment to Investigate ways of warming people suffering from hypothermia.

Suppose that literature reviews have shown that in past experiments, the standard deviation in the time needed to rewarm bodies was around 10 minutes. We are interested in detecting differences of about 10 minutes in the mean time needed to rewarm bodies among the three methods and between the two genders. What sample sizes would be needed for a CRD.

First, compute the sample size required for each of the two factors.

When examining gender effects, we find that the standard deviation is about 10, the difference to detect is about 10, and a total sample size of about 34 is required for an 80% power at $\alpha = 0.05$.

When examining method effects, we find that again the standard deviation is about 10, the difference between the smallest and largest mean is set to 10 while the third mean is placed in the middle, and a total sample size of about 60 is needed for an 80% power at $\alpha = 0.05$.

The two results must be reconciled. If you must obtain the desired power for both factors, then you must use the larger sample size, i.e., about 10 subjects per treatment combination.

If this is too costly, you will have to make compromises. For example, you could choose a sample size of 50 which would give you the desired power for detecting gender effects, but not method effects.

9.5.2 What is the difference between a 'block' and a 'factor'?

Blocks are typically not manipulated but rather are a collection of experimental units that are similar. You usually assume that blocking effects don't interact with factor effects. Usually, blocks are not assigned to experimental units - the experimental units are conveniently grouped into blocks. Blocks are "passive".

A factor has levels that are manipulated and randomized over the experimental units. Factors are usually assigned to experimental units. [Of course in analytical surveys, units are randomly selected from each level rather than being assigned to levels.] There is usually no natural grouping of experimental units to factor levels. Factors are "active".

In the case of the rancid fat, laboratories were assigned at random to irradiate the experimental units - the batches of fat. There is no natural grouping of batches of fat with the laboratories. Consequently, laboratories were treated as a factor rather than a block.

In the case of the mosquito repellent, people were randomly assigned to locations. There were no natural groups of people at each location. Location was again treated as a factor rather than a block.

In the case of seedling growth, the blocks were locations around the province. At each location, there were several 1 ha plots. The plots were grouped naturally by location. Hence, location should be treated as a block rather than as a factor.

In some cases, the distinction is not as clear cut. The computer packages will give the same results if a term is factor or a block so the final results will be the same. The only real reason to distinguish carefully among blocks or factors is for interpreting the results. Usually, blocks are not randomized so tests for "block" effects don't make much sense.

9.5.3 If there is evidence of an interaction, does the analysis stop there?

In the case of an interaction being detected, first examine if a transformation would remove the interaction. For example, if the factor operates multiplicatively (it reduces yield by 1/2) rather than additively (it reduced yield by 50 kg/ha) a log-transform would remove interaction effects.

If evidence of interaction is still present, then it really doesn't make much sense to test main effects. An interaction between two factors indicates that the effect of a factor changes depending on the level of the other factor. The test for the main effect in the presence of an interaction would examine if the average effect exists - this may not bear any relevance to the individual effects.

At this point, you should examine the individual treatment combinations to see which treatments appear to differ from other treatments.