Experiments 8

An experiment tries to evaluate the effects of one or more treatments. Imagine you wish to evaluate the effect of some treatment. One natural strategy is: choose groups of subjects; apply this treatment at different levels to different groups; then see if the groups are different after treatment. For example, you might wish to investigate whether consuming a painkiller affected headaches. Different levels of treatment would correspond to different numbers of pills (say, none, one or two). Different subjects would be different people. We would then divide the subjects into groups, apply different levels of treatment (i.e. give them different numbers of pills), and record outcomes. Now we need to tell whether the groups are different.

Telling whether the groups are different requires care, because the subjects will differ for a variety of irrelevant reasons (bodyweight, susceptibility to medication, and so on), so there will be differences between groups—we need to be able to tell whether these differences are due to the treatment, or to the irrelevant reasons. One powerful strategy is to allocate subjects to groups before treatment at random, so that each group looks similar (i.e. each group has the same variation in bodyweight, and so on). We can then use the significance machinery of the previous chapter to tell whether the differences between groups are due to the effects of treatment.

The procedures we use can be extended to deal with multiple treatments. We deal with the case of two treatments, because it captures important phenomena. Rather than do experiments for each treatment separately, it is a good idea to look at the results for different levels of both treatments together. Doing so allows us to identify possible interactions, and to treat fewer subjects in total. The machinery doesn't change (though there is more of it) when we pass to more than two treatments.

8.1 A Simple Experiment: The Effect of a Treatment

I use the term "treatment" very broadly here. The subject of experimental design started around the question of how to evaluate fertilizers, and it's natural to think about medical treatments, but many other things can be a treatment. Examples include: the amount of RAM you put into a computer; different interface design decisions; different choices of algorithm for rendering pictures; and so on. We will evaluate the effect of some treatment by dividing subjects into groups, applying the treatment at different levels to different groups, then seeing if the groups are "different" after treatment. For this procedure to work, we need: (a) the groups to be the "same" before treatment, and (b) a sensible way to tell whether the groups are different.

Randomization is a very strong strategy for allocating subjects to groups. Randomization ensures that each group "looks like" each other group. Differences between group then result from either sampling variation or from the effects of the treatment, and we know how to assess the effects of sampling. Randomization is a strong strategy for setting up experiments, but it doesn't cover every contingency. For example, imagine one treatment involves using a particular machine at a particular setting. Changing the setting of the machine again and again might be a problem. Furthermore, the order in which you do the experiments might matter—if the machine overheats, for example, while handling a subject at one setting, you may find that the results on the next subject are affected. We will assume that the experimental equipment doesn't have a memory, and that changing settings and so on is trivial.

D. Forsyth, *Probability and Statistics for Computer Science*,
https://doi.org/10.1007/978-3-319-64410-3_8

8.1.1 Randomized Balanced Experiments

Assume you wish to use L levels of treatment (so there are L groups; not treating a subject is one level of treatment). We will evaluate the results of the experiment by comparing the outcome for each group of subjects. This means it is helpful if there are the same number of subjects in each group—we say the experiment is **balanced**—so that the error resulting from chance effects in each group is the same. We will have G subjects in each group, and so there is a total of LG subjects. We must now choose which treatment each subject gets. We allocate subjects to groups at random, ensuring that each group gets G subjects. You could do this, for example, by permuting the subjects randomly, then allocating the first G to group 1, etc.

We now perform the experiment, by treating each group of subjects at the prescribed level, and recording the results. For each subject, we will observe a measurement. Write x_{ij} for the observed value for the j'th subject in the i'th group. We want to know if the groups are different. We assume that differences in observed values in each group are purely due to noise. This means that we could model

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

where ϵ_{ij} is "noise"—unmodelled effects that have zero mean, and do not depend on the treatment. If the treatment has no effect, then each of the μ_i will be the same. We can model the treatment at level i as having an effect t_i, so that

$$\mu_i = \mu + t_i.$$

We would like μ to be the average of the μ_i, so we must have

$$\sum_u t_u = 0.$$

All this yields a model

$$x_{ij} = \mu + t_i + \epsilon_{ij}$$

We assume that the noise is independent of the treatment level. This assumption is satisfied by randomizing the allocation of subjects to treatment groups. We will assume that the noise is normally distributed, with variance σ^2. This makes it straightforward to estimate μ and μ_i with least squares. We write $\hat{\mu}$ for our estimate of μ, and so on. We then choose $\hat{\mu}$ to be the value of μ that minimizes the overall sum of squared differences

$$\hat{\mu} = \underset{\mu}{\text{argmin}} \sum_{ij} (x_{ij} - \mu)^2.$$

We choose each $\hat{\mu}_i$ to be the value of μ_i that minimizes the sum of squared differences within the i'th group

$$\hat{\mu}_i = \underset{\mu_i}{\text{argmin}} \sum_{j} (x_{ij} - \mu_i)^2.$$

All this yields

$$\hat{\mu} = \frac{\sum_{ij} x_{ij}}{GL} \text{ and } \hat{\mu}_i = \frac{\sum_j x_{ij}}{G}.$$

8.1.2 Decomposing Error in Predictions

The results are in L groups, one for each level of treatment. The overall sum of squared differences from the estimated mean $\hat{\mu}$ is

$$\text{SS}_\text{T} = \sum_{ij} (x_{ij} - \hat{\mu})^2.$$

This can be decomposed into two terms, as the exercises show. We have

$$\sum_{ij} (x_{ij} - \hat{\mu})^2 = \left[\sum_{ij} (x_{ij} - \hat{\mu}_i)^2 \right] + G \left[\sum_i (\hat{\mu} - \hat{\mu}_i)^2 \right].$$

This expression breaks the total sum of squared errors $\mathrm{SS_T}$ into two components. The first,

$$\mathrm{SS_W} = \sum_{ij} (x_{ij} - \hat{\mu}_i)^2,$$

is due to within-group variation; and the second,

$$\mathrm{SS_B} = G\left[\sum_i (\hat{\mu} - \hat{\mu}_i)^2\right]$$

is due to between-group variation. The relative size of these two terms should tell us whether the treatment has any effect or not. For example, assume there are large effects. Then $\mathrm{SS_B}$ should be "big", because the $\hat{\mu}_i$ should be different from one another, and the $\mathrm{SS_W}$ should be "small", because measurements should be rather closer to their group means than to the overall mean (and because $\mathrm{SS_T} = \mathrm{SS_W} + \mathrm{SS_B}$, so when one goes up the other must go down). Now assume there is no effect. Then the $\hat{\mu}_i$ should be quite similar to $\hat{\mu}$, and so to each other. This means $\mathrm{SS_B}$ should be "small", and so on. Using this line of reasoning requires some quantitative notion of what is "big" and what is "small".

8.1.3 Estimating the Noise Variance

The trick to knowing whether $\mathrm{SS_B}$ is "big" or not runs as follows. If there is no effect, then $\mathrm{SS_B}$ and $\mathrm{SS_W}$ can each be used to produce estimates of the noise variance. We produce these estimates, and then determine whether the difference between estimates can be explained by sampling variation. If it can't, then the treatment has some effect.

Using the ith treatment group, we can estimate σ^2 by

$$\hat{\sigma}^2 = \frac{\sum_j (x_{ij} - \hat{\mu}_i)^2}{(G-1)}.$$

An even better estimate would be to average these estimates across groups. This yields

$$\hat{\sigma}^2 = \frac{1}{L}\sum_i \left[\frac{\sum_i (x_{ij} - \hat{\mu}_j)^2}{(G-1)}\right].$$

This estimate is sometimes known as the **within group variation** or **residual variation**. We write

$$\begin{aligned}\mathrm{MS_W} &= \frac{1}{L}\sum_i \left[\frac{\sum_j (x_{ij} - \hat{\mu}_i)^2}{(G-1)}\right] \\ &= \left(\frac{1}{L(G-1)}\right)\mathrm{SS_W}.\end{aligned} \tag{8.1}$$

This estimate has $L(G-1)$ degrees of freedom, because we have LG data items, but we lose one degree of freedom for each of the L means we have estimated.

Assume the treatment has no effect. Then we expect that all the $t_i = \mu_i - \mu$ are zero. In turn, each of the $\hat{\mu}_i$ is an estimate of μ. These estimates are the values of a random variable whose expected value is μ, and whose variance is σ^2/G. This means that

$$\frac{\sum_i (\hat{\mu}_i - \hat{\mu})^2}{L-1} \approx \frac{\sigma^2}{G}.$$

We can use this to produce another estimate of σ^2, where

$$\hat{\sigma}^2 = G\frac{\sum_j (\hat{\mu}_j - \hat{\mu})^2}{L-1}$$

This estimate is sometimes known as the **between group variation** or **treatment variation**. We write

$$\mathrm{MS_B} = G\frac{\sum_j(\hat{\mu}_j - \hat{\mu})^2}{(L-1)} = \left(\frac{1}{L-1}\right)\mathrm{SS_B}.$$

This estimate has $L-1$ degrees of freedom, because we have used L data items (the $\hat{\mu}_j$), but estimated one mean (the $\hat{\mu}$).

If the treatment has no effect, then any difference between these estimates would have to be the result of sampling effects. We can apply an F-test of significance (Sect. 7.3.1) to the ratio of the estimates. Now imagine that at least one of the treatment levels has an effect. In turn, this would mean that one of the $\hat{\mu}_j$ was more different from $\hat{\mu}$ than would occur if the treatment had no effect. This means that $\mathrm{MS_B}$ would be larger than would happen if the treatment had no effect. This means that, if there is no effect of treatment, the statistic

$$F = \frac{\mathrm{MS_B}}{\mathrm{MS_W}}$$

would be "about" one, with a known distribution (the F-distribution; Sect. 7.3.1). If there is a treatment effect, then we expect F to be larger than one. Finally, we need the degrees of freedom. The estimate $\mathrm{MS_W}$ has $L(G-1)$ degrees of freedom. The estimate $\mathrm{MS_B}$ has $L-1$ degrees of freedom. We can now use the F-test to obtain a p-value for F with $L-1$, $L(G-1)$ degrees of freedom. In other words, we use tables or a software environment to evaluate

$$\int_F^\infty p_f(u; L-1, L(G-1))du$$

(where p_f is the probability density for the F-statistic, Sect. 7.3.1).

8.1.4 The ANOVA Table

We now have a powerful and useful procedure that I can name and put in a box. Analyzing data by comparing variances in this way is sometimes known as **analysis of variance** or **ANOVA**. It is usual to lay out the information in a table, sometimes called an **ANOVA table**. We have only one kind of treatment we are varying in the experiment, so the experiment is known as a **one factor** experiment. The form of the table appears in Procedure 8.1.

Procedure 8.1 (Evaluating Whether a Treatment Has Significant Effects with a One-Way ANOVA for Balanced Experiments)
Perform a randomized experiment: Choose L levels of treatment. Randomize LG subjects into L treatment groups of G subjects each. Treat each subject, and record the results.

Terminology:	x_{ij}	value for the j'th subject in the i'th treatment level group	
	$\hat{\mu}$	overall mean	$\left(\sum_{ij} x_{ij}\right) / (GL)$
	$\hat{\mu}_i$	i'th group mean	$\left(\sum_j x_{ij}\right) / G$
	$\mathrm{SS_W}$	within group sum of squares	$\sum_{ij}(x_{ij} - \hat{\mu}_i)^2$
	$\mathrm{SS_B}$	between group sum of squares	$G\left[\sum_i(\hat{\mu} - \hat{\mu}_i)^2\right]$
	$\mathrm{MS_W}$	within group mean squares	$\mathrm{SS_W} / (L(G-1))$
	$\mathrm{MS_B}$	between group mean squares	$\mathrm{SS_B} / (L-1)$
	F	value of F-statistic	$\mathrm{MS_B}/\mathrm{MS_W}$
	p-value	from tables or software	$\int_F^\infty p_f(u; L-1, L(G-1))du$

(continued)

Make the ANOVA table: Form the table

	DOF	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	L-1	SS_B	MS_B	MS_B/MS_W	p-value
Residuals	L (G-1)	SS_W	MS_W		

Interpretation: If the p-value is small enough, then only an extremely unlikely set of samples could explain the difference between the levels of treatment as sampling error; it is more likely the treatment has an effect.

Worked example 8.1 (Does Depth Affect Aldrin Concentration?) Jaffe et al measured the concentration of various pollutants in the Wolf river. Assess the evidence supporting the belief that the concentration of aldrin does not depend on depth. You can find the dataset at http://www.statsci.org/data/general/wolfrive.html.

Solution The original measurements are described in Jaffe, P. R., Parker, F. L., and Wilson, D. J. (1982). Distribution of toxic substances in rivers. *Journal of the Environmental Engineering Division*, 108, 639–649. I obtained the ANOVA table below.

	DOF	Sum Sq	Mean Sq	F value	Pr(>F)
Depth	2	16.83	8.415	6.051	0.00674
Residuals	27	37.55	1.391		

The very small p-value suggests that the evidence is very strongly against the idea that concentration does not depend on depth.

8.1.5 Unbalanced Experiments

In an ideal experiment, each group has the same number of subjects. This is known as a **balanced experiment**. Such experiments can be difficult to arrange. A measurement might be impossible, or a subject might get lost, meaning that some levels of treatment may have fewer than others. The result is an **unbalanced experiment**. This doesn't affect the reasoning of the section above, though subtleties can arise. We do need to be careful about the number of degrees of freedom, though.

Assume that the i'th group has G_i items in it. Then the estimate of σ^2 based on residual variation becomes

$$\mathrm{MS_W} = \frac{1}{L}\sum_j \left[\frac{\sum_i (x_{ij}-\hat{\mu}_i)^2}{(G_i-1)}\right].$$

This estimate has $\sum_i G_i - L$ degrees of freedom, because there are $\sum_i G_i$ items and we had to estimate L means, one for each treatment level.

The estimate of σ^2 based on treatment variation is slightly more interesting. For the i'th group, the standard error of the estimate of $\hat{\mu}$ is

$$\sqrt{\frac{\sigma^2}{G_i}}$$

which means that $\sqrt{G_j}(\hat{\mu}_j - \hat{\mu})$ is the value of a normal random variable with mean 0 and variance σ^2. In turn, this means that

$$\mathrm{MS_B} = \frac{\sum_i \left[G_i(\hat{\mu}_i - \hat{\mu})^2\right]}{(L-1)}.$$

is an estimate of σ^2. This estimate has $L-1$ degrees of freedom, because we used L data items but estimated one mean.

In the balanced case, the sums of squares were meaningful because one scales them to get the estimates of σ^2. There's an established tradition of supplying them in the ANOVA table. In the unbalanced case, it's hard to turn the sums of squares into anything useful by eye. This is the result of the weighting terms in the expressions for the mean square error. I omit them from the ANOVA tables in the worked examples below.

Worked example 8.2 (Does Olestra in Potato Chips Produce Gastro-Intestinal Symptoms?) Olestra is a fat that is edible, but cannot be digested. This means the effective calorie content of chips containing olestra is reduced, but there might be gastro-intestinal consequences. The paper "Gastrointestinal Symptoms Following Consumption of Olestra or Regular Triglyceride Potato Chips", JAMA, 279: 150–152, L. Cheskin, R. Miday, N. Zorich, and T. Filloon (1998). ran a double blind randomized trial. They observed 89 people eating chips with olestra who had a GI outcome; 474 who did not; 93 eating chips without olestra who had a GI outcome; and 436 eating chips without olestra who did not. Subjects ate chips *ad libitem* (i.e. until they felt like stopping). Assess the evidence that olestra causes a GI outcome under these circumstances using an ANOVA.

Solution Each subject gets a 1 if there is a GI effect, and a 0 otherwise. For this data, I found

	DOF	Mean Sq	F value	Pr(>F)
Fat	1	0.086	0.63	0.43
Residuals	1090	0.14		

suggesting that there is no reason to feel that consumption of olestra in potato chips eaten *ad libitem* causes GI effects.

Worked example 8.2 should give you pause. You should think carefully about what the experiment actually shows. It should not be interpreted as saying that eating quantities of indigestible fat has no GI consequences, and the authors didn't make this claim. The claim might be true, but the experiment is silent on the point. Interpreting an experiment often requires scrupulousl precision. What this experiment says is there is no reason to conclude that eating a particular indigestible fat in potato chips eaten *ad libitem* causes GI consequences. One would really like to know how much of the relevant fats each group of subjects ate. As an extreme example, think about chips cooked in axle grease eaten *ad libitem*. These would very likely not cause GI symptoms because you wouldn't eat any at all.

Worked example 8.2 is not really all that interesting, because there are only two levels of treatment. You could analyze these experiments with the methods of Sect. 7.2. Here is a more interesting example.

Worked example 8.3 (Does Hair Color Affect Pain Threshold?) An experiment at the University of Melbourne tested the dependency of the pain threshold of human subjects on their hair color. You can find the dataset at http://www.statsci.org/data/oz/blonds.html. Assess the evidence that pain threshold depends on hair color.

Solution If you download the dataset, you'll find that there are four hair colors and nineteen data items. For each subject, there is a number. Larger values of this number represent higher tolerable pain thresholds. I obtained the ANOVA table below.

	DOF	Mean Sq	F value	Pr(>F)
HairColour	3	454.0	7.04	0.0035
Residuals	15	64.5		

This dataset appears in various places on the web, but its status as a "real" dataset is a bit marginal. I wasn't able to find out who conducted this experiment, where the results were published, and whether it conformed to human subjects requirements when conducted.

8.1.6 Significant Differences

We know how to reject the hypothesis that the treatment has no effect by computing a p-value from an ANOVA. We'd like to do more. Generally, it is dangerous to fish around in the data produced by an experiment to see if anything fits; if there is enough data, something should have a low p-value. It is important to avoid this danger if you want to interpret the experimental results correctly (sometimes, it is profitable, though shameful, not to). A natural first step here is a boxplot of what happens at each treatment level (Fig. 8.1). With luck, the effect of the treatment will be so strong that significance testing is a formality.

One way to avoid this danger is to declare a set of hypotheses we want to evaluate in advance of the experiment, then investigate those. The procedure is particularly straightforward for hypotheses known as **contrasts**. These are linear functions of the treatment means. Recall I wrote μ_i for the treatment means. A contrast takes the form

$$\sum_i c_i \mu_i$$

for some set of constants c_i. There are two relatively straightforward procedures. We could produce a confidence interval for a contrast. Alternatively, we can evaluate the significance of the evidence against a contrast being zero. Each of these procedures relies on the fact that we know the distribution of the estimate $\hat{\mu}_i$ for the treatment means. In particular, $\hat{\mu}_i$ is the value of a random variable M_j which is normally distributed (assuming that ϵ_{ij} are normally distributed); has mean μ_i; and, writing G_i for the number of subjects in the i'th group, has variance

$$\frac{\sigma^2}{G_i}.$$

In turn, this means that

$$\sum_i c_i \hat{\mu}_i$$

is the value of a random variable C which is normally distributed; has mean $\sum_i c_i \mu_i$; and has variance

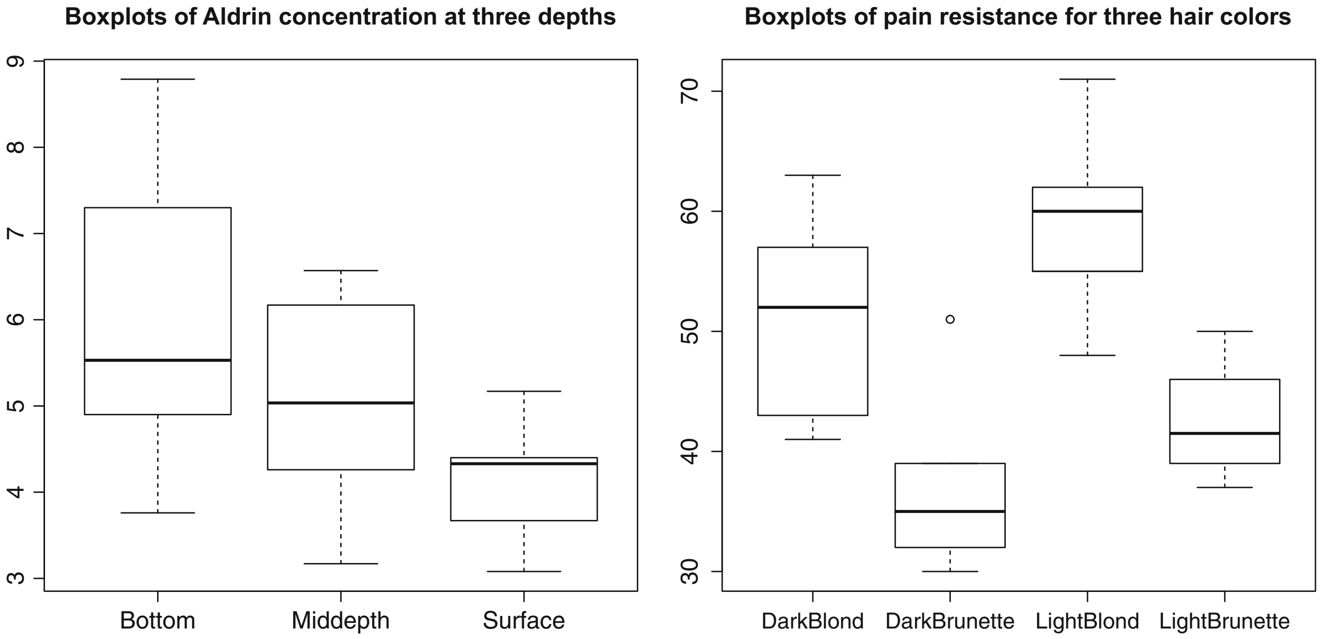

Fig. 8.1 On the *left*, a boxplot of concentration of Aldrin at three different depths (see Worked example 8.1). There's a strong suggestion here that the concentration is affected by depth level, which is the treatment. Notice how the box for the "Surface" measurements does not intersect the box for the "Deep" measurements. On the *right*, a boxplot of the tolerable pain level by hair color for the data of Worked example 8.3. There's a strong suggestion here that hair color affects pain threshold, though you should be careful relying on this conclusion. As the worked example points out, it isn't clear where the data came from, or what it means

$$\sigma^2 \sum_i \frac{c_i^2}{G_i}$$

Now producing a confidence interval for a contrast follows the lines of Sect. 6.2. Evaluating the significance of the evidence against a contrast being zero follows the lines of Sect. 7.1. Notice that we do not *know* σ^2, but must estimate it. This means we should use a T-test. We estimate σ^2 as MS_W (Sect. 8.1.3), so the number of degrees of freedom is $\sum_i (G_i - 1)$.

One natural set of contrasts are the differences between treatment means. You should notice that if there are L treatment levels there are $L(L-1)/2$ differences, which will get inconvenient if there are many treatment means. With that said, it is quite usual to look for significant difference between treatment means. The simplest procedure for doing so follows the recipe for contrasts, above.

Worked example 8.4 (How Significant are the Differences in Aldrin Concentration at Different Depths?) Jaffe et al measured the concentration of various pollutants in the Wolf river. Assess how different the mean concentration is at each depth. You can find the dataset at http://www.statsci.org/data/general/wolfrive.html.

Solution There are three depths: Surface, Middepth, and Bottom. I found

$$\begin{aligned}\hat{\mu}_\text{Middepth} - \hat{\mu}_\text{Surface} &= 0.83\\ \hat{\mu}_\text{Bottom} - \hat{\mu}_\text{Middepth} &= 1.00\\ \hat{\mu}_\text{Bottom} - \hat{\mu}_\text{Surface} &= 1.83\end{aligned}$$

and the standard error for each was 0.56. This means that the difference between Middepth and Surface is about one and half standard errors, and is quite possibly due to sampling effects (using a one sided t-test with 9 degrees of freedom, I got a p-value of 0.086). The difference between Bottom and Middepth is unlikely to be due to sampling effects (p-value: 0.053); and the difference between Bottom and Surface is extremely unlikely to be due to sampling effects (p-value: 0.0049). I used a one-sided test here, because I wanted to assess the fraction of samples likely to have even *larger* differences than the one observed.

You need to approach this procedure with moderate caution. If L is large, it is possible that you will think more differences are significant than is really the case, because you are looking at multiple differences and samples. The significance test hasn't taken this into account. More advanced methods—beyond the scope of this account—are required to do deal with this.

8.2 Two Factor Experiments

Now imagine there are two factors that likely affect the outcome of an experiment. For example, we might be testing the effect of having more RAM and having a higher clock speed on the speed of some program. The treatments here are obvious: different levels of treatment correspond to different amounts of RAM and different clock speeds (say, small, medium and large RAM; low, medium and high clock speeds). One possibility is to set up one experiment to test the effect of RAM and another to test the effect of the clock speed. But this is not a good idea.

First, the treatments might interact in some way, and the experiment we set up should be able to identify this if it happens. Second, it turns out that a sensible design that investigates both factors together will involve fewer treatments than if we investigate the factors separately.

This will become clear as we set up the experiment for two factors. We will deal only with balanced experiments, because they're easier to interpret. We will assume that the first factor has L_1 levels, and the second factor has L_2 levels. It's natural to visualize the experiment as an $L_1 \times L_2$ table of cells, where each cell contains the subjects for the experiment at those levels (Fig. 8.2). Each cell will contain G subjects. Performing this experiment will involve measuring $L_1 \times L_2 \times G$ measurements.

Now assume the factors do not interact. Then we can use the rows to estimate the effect of factor one at L_1 different levels using $G \times L_2$ measurements per cell. Similarly, we can use the columns to estimate the effect of factor two at L_2 different levels using $G \times L_1$ measurements per cell. To obtain the same number of measurements per cell for each factor

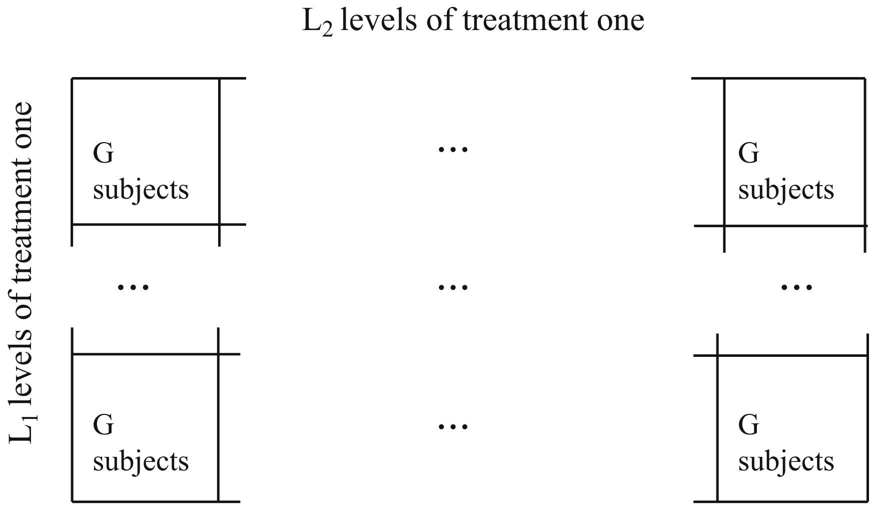

Fig. 8.2 Think about a two factor experiment as an $L_1 \times L_2$ table of cells, each containing G subjects chosen at random. The subjects in each cell get the level of treatments one and two chosen by the cell's indices in the table. We write x_{ijk} for the response observed for the k'th subject in cell i, j

with independent experiments would take more experiments. You would have to use $L_1 \times (G \times L_2)$ experiments for factor one and another $L_2 \times (G \times L_1)$ experiments for factor two.

Randomization is still a strong strategy for assigning groups to cells. Again, we will assume that the experimental equipment doesn't have a memory, and that changing settings and so on is trivial. We allocate subjects to groups at random, ensuring that each group gets G subjects. You could do this, for example, by permuting the subjects randomly, then allocating the first G to group 1, 1, etc.

We then perform the experiment, by treating each group of subjects at the prescribed level, and recording the results. For each subject, we will observe a measurement. We want to know if there is an interaction between the treatments, and if either treatment has any effect. We can investigate this question using methods like those for one-factor experiments.

We assume that differences in observed values in each group are purely due to noise. Write x_{ijk} for the observed value for the k'th subject in the treatment group that has the i'th level of the first treatment and the j'th level of the second treatment. This means that we could model

$$x_{ijk} = \mu_{ij} + \epsilon_{ij}$$

where ϵ_{ij} is "noise"—unmodelled effects that have zero mean. There are three effects that could cause the groups to have different m_{ij}. The first treatment might have an effect; the second treatment might have an effect; or there could be an interaction between the treatments that could have an effect. Write a_i for the effects caused by the first treatment, b_j for the effects caused by the second treatment, and c_{ij} for interaction effects. We can model

$$\mu_{ij} = \mu + a_i + b_j + c_{ij}.$$

As in the case for single factor experiments, we constrain a_i, b_j and c_{ij} so that the mean of the μ_{ij} terms is μ. This means that $\sum_u a_u = 0$, $\sum_v b_v = 0$, $\sum_u c_{uv} = 0$, and $\sum_v c_{uv} = 0$ (notice this means that $\sum_{uv} c_{uv} = 0$, too).

As in the single factor case, we assume that the noise is independent of the treatment level. This assumption is satisfied by randomizing the allocation of subjects to treatment groups. We will assume that the noise is normally distributed, with variance σ^2. We will use traditional notation, and write $\mu_{i\cdot} = \mu + a_i$ and $\mu_{\cdot j} = \mu + b_j$. Again, it is straightforward to estimate μ, $\mu_{i\cdot}$, $\mu_{\cdot j}$, and μ_{ij} with least squares. As before, we write $\hat{\mu}$ for our estimate of μ, and so on. As before, we then choose $\hat{\mu}$ to be the value of μ that minimizes the overall sum of squared differences

$$\sum_{ijk} (x_{ijk} - \mu)^2.$$

We choose each $\hat{\mu}_{ij}$ to be the value of μ_{ij} that minimizes the sum of squared differences within the i, j'th group

$$\hat{\mu}_{ij} = \underset{\mu_{ij}}{\text{argmin}} \sum_k (x_{ijk} - \mu_{ij})^2.$$

Now consider $\hat{\mu}_{i\cdot}$. There are L_1 different values, one for each level of the first factor. These account for the effects of the first factor alone, so we choose $\hat{\mu}_{i\cdot}$ to minimize the sum of squared differences to *every* measurement at the i'th level of the first factor. This is

$$\hat{\mu}_{i\cdot} = \underset{\mu_{i\cdot}}{\text{argmin}} \sum_{jk} (x_{ijk} - \mu_{i\cdot})^2.$$

A similar argument works for $\hat{\mu}_{\cdot j}$. There are L_2 different values, one for each level of the first factor. These account for the effects of the second factor alone, so we choose $\hat{\mu}_{\cdot j}$ to minimize the sum of squared differences to *every* measurement at the j'th level of the second factor. This is

$$\hat{\mu}_{\cdot j} = \underset{\mu_{\cdot j}}{\text{argmin}} \sum_{ik} (x_{ijk} - \mu_{\cdot j})^2.$$

All this yields

$$\begin{aligned}
\hat{\mu} &= \frac{\sum_{ijk} x_{ijk}}{GL_1L_2} \\
\hat{\mu}_{ij} &= \frac{\sum_{k} x_{ijk}}{G} \\
\hat{\mu}_{i\cdot} &= \frac{\sum_{jk} x_{ijk}}{GL_2} \\
\hat{\mu}_{\cdot j} &= \frac{\sum_{ik} x_{ijk}}{GL_1}.
\end{aligned}$$

8.2.1 Decomposing the Error

Think of the results as sitting in a table of $L_1 \times L_2$ cells, one for each pair of treatment levels. The overall sum of squared differences from the estimated mean $\hat{\mu}$ is

$$\text{SS}_{\text{T}} = \sum_{ijk} (x_{ijk} - \hat{\mu})^2.$$

This can be decomposed into four terms, rather like the pattern in Sect. 8.1.2. We have that

$$\sum_{ijk} (x_{ijk} - \hat{\mu})^2 = \sum_{ijk} \begin{bmatrix} (x_{ijk} - \hat{\mu}_{ij}) + \\ (\hat{\mu}_{i\cdot} - \hat{\mu}) + \\ (\hat{\mu}_{\cdot j} - \hat{\mu}) + \\ (\hat{\mu}_{ij} - \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot j} + \hat{\mu}) \end{bmatrix}^2$$

$$= \begin{bmatrix} \sum_{ijk} (x_{ijk} - \hat{\mu}_{ij})^2 + \\ GL_2 \sum_{i} (\hat{\mu}_{i\cdot} - \hat{\mu})^2 + \\ GL_1 \sum_{j} (\hat{\mu}_{\cdot j} - \hat{\mu})^2 + \\ G \sum_{ij} (\hat{\mu}_{ij} - \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot j} + \hat{\mu})^2 \end{bmatrix},$$

which you can establish using (a lot more of) the same reasoning as in Sect. 8.1.2. We label these terms

$$\begin{aligned}
\mathrm{SS}_{\mathrm{T}} &= \sum_{ijk}(x_{ijk} - \hat{\mu})^2 \\
\mathrm{SS}_{\mathrm{W}} &= \sum_{ijk}(x_{ijk} - \hat{\mu}_{ij})^2 \\
\mathrm{SS}_{\mathrm{Tr1}} &= GL_2 \sum_{i}(\hat{\mu}_{i\cdot} - \hat{\mu})^2 \\
\mathrm{SS}_{\mathrm{Tr2}} &= GL_1 \sum_{j}(\hat{\mu}_{\cdot j} - \hat{\mu})^2 \\
\mathrm{SS}_{\mathrm{I}} &= G \sum_{ij}(\hat{\mu}_{ij} - \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot j} + \hat{\mu})^2.
\end{aligned}$$

The same line of reasoning as in Sect. 8.1.2 applies to these terms. Imagine neither treatment has any effect, and there is no interaction. Then each of $\hat{\mu}_{ij}$, $\hat{\mu}_{i\cdot}$ and $\hat{\mu}_{\cdot j}$ should be similar, meaning that SS_{W} should be large compared to the other terms. Now imagine there is a strong interaction between the treatments. This means that c_{ij} is "large"; but $(\hat{\mu}_{ij} - \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot j} + \hat{\mu})$ is an estimate of c_{ij}, so we expect that SS_{I} will be "large". Again, we need to be crisp about what it means to be "large", and we can do this using ideas of significance.

As before, we will use estimates of the noise variance to compute measures of significance. We can estimate the variance of the noise using any of the cells. A better estimate is obtained by averaging over the cells, so we have as an estimate of the noise variance, sometimes called the **within group mean squares** and written MS_{W}. We have

$$\begin{aligned}
\hat{\sigma}^2_{\text{cellave}} &= \left(\frac{1}{L_1L_2}\right)\sum_{ijk}\frac{(x_{ijk} - \hat{\mu}_{ij})^2}{G-1} \\
&= \mathrm{SS}_{\mathrm{W}}\left(\frac{1}{L_1L_2(G-1)}\right) = \mathrm{MS}_{\mathrm{W}}.
\end{aligned}$$

This estimate has $L_1L_2(G-1)$ degrees of freedom, because we used L_1L_2G data items, and estimated L_1L_2 means.

8.2.2 Interaction Between Effects

Assume there is no interaction between the treatments. Then the true c_{ij} should be zero. Notice that we have the following estimates

$$\begin{aligned}
&\mu + a_i + b_j + c_{ij} \approx \hat{\mu}_{ij} \\
&\mu + a_i \approx \hat{\mu}_{i\cdot} \\
&\mu + b_j \approx \hat{\mu}_{\cdot j} \\
&\mu \approx \hat{\mu}
\end{aligned}$$

so that we can estimate c_{ij} as $\hat{\mu}_{ij} - \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot j} + \hat{\mu}$. Now each subject's response has some noise in it, with variance σ^2 which is unknown, but is the same for all subjects and all treatment levels. This means that $\hat{\mu}_{ij} - \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot j} + \hat{\mu}$ is the value of a random variable whose mean is zero. This estimate is averaged over G subjects, so the variance of the random variable is σ^2/G. So we can estimate

$$\begin{aligned}
\hat{\sigma}^2_{\text{inter}} &= G\frac{\sum_{ij}(\hat{\mu}_{ij} - \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot j} + \hat{\mu})^2}{(L_1-1)(L_2-1)} \\
&= \left(\frac{1}{(L_1-1)(L_2-1)}\right)\mathrm{SS}_{\mathrm{I}} = \mathrm{MS}_{\mathrm{I}}.
\end{aligned}$$

This estimate is sometimes called the **interaction mean squares**, and written MS_{I}.

If there is no interaction, then any difference between $\mathrm{MS_I}$ and $\mathrm{MS_W}$ would have to be the result of sampling effects. We can apply an F-test of significance (Sect. 7.3.1), applied to the ratio of the estimates. In turn, this would mean that

$$F = \frac{\mathrm{MS_I}}{\mathrm{MS_W}}$$

would be "about" one, with a known distribution (the F-distribution; Sect. 7.3.1). Now imagine there is some interaction. Then we expect that $\mathrm{MS_I}$ is larger than $\mathrm{MS_W}$, because the $\hat{\mu}_{ij}$ differ from predictions made assuming no interaction, so we expect F is larger than one. Finally, we need the degrees of freedom. The estimate $\mathrm{MS_I}$ has $(L_1L_2 - 1)$ degrees of freedom. The estimate $\mathrm{MS_W}$ has $L_1L_2(G-1)$ degrees of freedom. We can now use the F-test to obtain a p-value for F with (L_1L_2-1), $L_1L_2(G-1)$ degrees of freedom. In other words, we use tables or a software environment to evaluate

$$\int_F^\infty p_f(u; (L_1L_2 - 1), L_1L_2(G-1))du$$

(where p_f is the probability density for the F-statistic, Sect. 7.3.1).

8.2.3 The Effects of a Treatment

Assume we find no interaction between the treatments. Now we investigate the effect of treatment one. If there is no effect, then the true a_i should be zero. Recall that a_i can be estimated by $\hat{\mu}_{i\cdot} - \hat{\mu}$. Each subject's response has some noise in it, with variance σ^2 which is unknown, but is the same for all subjects and all treatment levels. This means that, for each i, we have $\hat{\mu}_{i\cdot} - \hat{\mu}$ is the value of a random variable whose mean is zero. The estimate of a_i has been obtained by averaging GL_2 subjects (all those who have level i of treatment one) so the variance of this random variable is $\sigma^2/(GL_2)$. We can estimate σ^2 using treatment one means. This estimate is sometimes called the **treatment one mean squares**, and written $\mathrm{MS_{T1}}$. We have

$$\hat{\sigma}^2_{\mathrm{T1}} = GL_2\frac{\sum_i(\hat{\mu}_{i\cdot} - \hat{\mu})^2}{L_1 - 1} = \mathrm{SS_{Tr1}}\left(\frac{1}{L_1 - 1}\right) = \mathrm{MS_{T1}}.$$

If there is no effect, then any difference between $\mathrm{MS_{T1}}$ and $\mathrm{MS_W}$ would have to be the result of sampling effects. We can apply an F-test of significance (Sect. 7.3.1), applied to the ratio of the estimates. Now imagine there is some effect of treatment. This would mean that

$$F = \frac{\mathrm{MS_{T1}}}{\mathrm{MS_W}}$$

would be "about" one, with a known distribution (the F-distribution; Sect. 7.3.1). If there is a treatment effect, then we expect F to be larger than one. Finally, we need the degrees of freedom. The estimate $\mathrm{MS_{T1}}$ has $(L_1 - 1)$ degrees of freedom. The estimate $\mathrm{MS_W}$ has $L_1L_2(G-1)$ degrees of freedom. We can now use the F-test to obtain a p-value for F with $(L_1 - 1)$, $L_1L_2(G-1)$ degrees of freedom. If the p-value is small enough, then only an extremely unlikely set of samples could explain the difference between the levels of treatment one as sampling error; it is more likely the treatment has an effect.

Treatment two works like treatment one. The estimate of σ^2 from the treatment two means is called the **treatment two mean squares**, and written $\mathrm{MS_{T2}}$. We have

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{T2}} &= GL_1\frac{\sum_j(\hat{\mu}_{\cdot j} - \hat{\mu})^2}{L_2 - 1} \\ &= \mathrm{SS_{Tr2}}\left(\frac{1}{L_2 - 1}\right) = \mathrm{MS_{T2}}.\end{aligned}$$

We end up using the F-test to obtain a p-value for the statistic

$$F = \frac{\mathrm{MS_{T2}}}{\mathrm{MS_W}}$$

with $(L_2 - 1)$, $L_1L_2(G-1)$ degrees of freedom.

8.2.4 Setting Up An ANOVA Table

We now have a powerful and useful procedure that I can name. The analysis is referred to as a **two-factor ANOVA** (or **two-way ANOVA**. Because there is so much of it, it is hard to put in one box, and I have used two. Again, we compute a set of terms, put them in a highly informative table, and then draw conclusions by inspecting the table.

Procedure 8.2 (Setting up a Two-Way ANOVA) Perform a randomized experiment: Choose L_1 levels of treatment for the first treatment, and L_2 levels for the second. Randomize L_1L_2G subjects into $L_1 \times L_2$ treatment groups of G subjects each. Treat each subject, and record the results.

Terminology:

x_{ijk}	value for the k'th subjectfor level i of treatment one and level j of treatment two	
$\hat{\mu}$	overall mean	$\sum_{ijk} x_{ijk}/(GL_1L_2)$
$\hat{\mu}_{ij}$	group means	$\sum_k x_{ijk}/(G)$
$\hat{\mu}_{i\cdot}$	treatmentone means	$\sum_{jk} x_{ijk}/(GL_2)$
$\hat{\mu}_{\cdot j}$	treatmenttwo means	$\sum_{ik} x_{ijk}/(GL_1)$
SS_W	within group sum of squares	$\sum_{ijk}(x_{ijk} - \hat{\mu}_{ij})^2$
SS_I	interaction sum of squares	$G\sum_{ij}(\hat{\mu}_{ij} - \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot j} + \hat{\mu})^2$
$\text{SS}_{\text{Tr}1}$	treatment one sum of squares	$GL_2\sum_i(\hat{\mu}_{i\cdot} - \hat{\mu})^2$
$\text{SS}_{\text{Tr}2}$	treatment two sum of squares	$GL_1\sum_j(\hat{\mu}_{\cdot j} - \hat{\mu})^2$
MS_W	within group mean squares	$(1/(L_1L_2(G-1)))\,\text{SS}_\text{W}$
MS_I	interaction mean squares	$(1/[(L_1-1)(L_2-1)])\,\text{SS}_\text{I}$
$\text{MS}_{\text{T}1}$	treatment one mean squares	$(1/(L_1-1))\,\text{SS}_{\text{Tr}1}$
$\text{MS}_{\text{T}2}$	treatment two mean squares	$(1/(L_2-1))\,\text{SS}_{\text{Tr}2}$
F_i	interaction F-statistic	$\text{MS}_\text{I}/\text{MS}_\text{W}$
F_1	treatment F-statistic	$\text{MS}_{\text{T}1}/\text{MS}_\text{W}$
F_2	treatment two F-statistic	$\text{MS}_{\text{T}2}/\text{MS}_\text{W}$
p-values	Statistic	DOF
interaction	F_i	$[L_1L_2-1]$, $[L_1L_2(G-1)]$
treatment one	F_1	$[L_1-1]$ $[L_1L_2(G-1)]$
treatment two	F_2	$[L_2-1]$, $[L_1L_2(G-1)]$

Procedure 8.3 (Forming and Interpreting a Two-Way ANOVA Table) Make the ANOVA table: Form the table

	DOF	Sum Sq	Mean Sq	F value	Pr(>F)
Tr 1	$L_1 - 1$	$\text{SS}_{\text{Tr}1}$	$\text{MS}_{\text{T}1}$	$\text{MS}_{\text{T}1}/\text{MS}_\text{W}$	tr 1 p-value
Tr 2	$L_2 - 1$	$\text{SS}_{\text{Tr}2}$	$\text{MS}_{\text{T}2}$	$\text{MS}_{\text{T}2}/\text{MS}_\text{W}$	tr 2 p-value
Tr 1:Tr 2	$L_1\,L_2 - 1$	SS_I	MS_I	$\text{MS}_\text{I}/\text{MS}_\text{W}$	int p-value
Res	$L_1L_2\,(G-1)$	SS_W	MS_W		

Interpretation: If the interaction p-value is small enough, it is likely the treatments interact. If the treatments interact, they must have an effect. If the treatments appear not to interact, then a small value of either of the treatment p-values implies that only an extremely unlikely set of samples could explain the difference between groups.

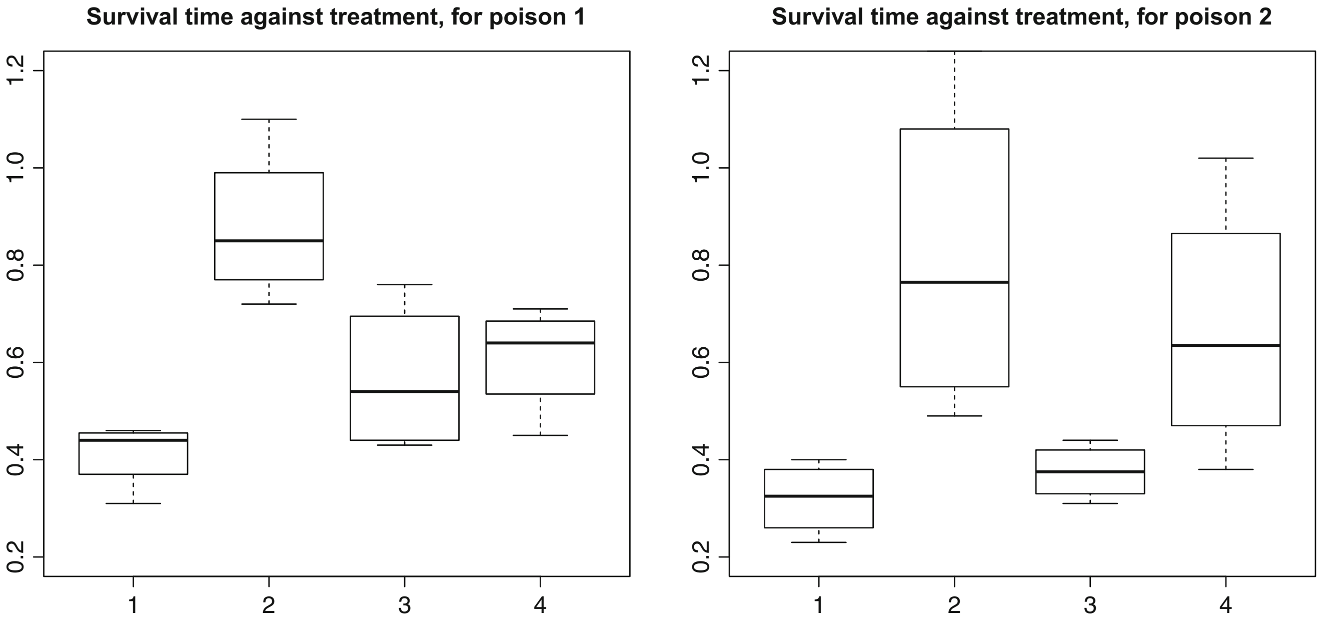

Fig. 8.3 On the *left*, boxplots of the survival time of subjects (animals of unrecorded species) poisoned with poison 1, by antidote type, for the experiment of Worked example 8.6. On the *right*, for poison 2. Generally, if there is no effect of a treatment the boxes in a graph should look the same; if there is no interaction, the pattern formed by the boxes should look the same, with perhaps a shift up or down of all boxes. The poisons clearly have an effect, the antidotes clearly have an effect, and (though the shape of the box patterns looks a bit different), there really isn't evidence that there is an interaction

Worked example 8.5 (Poison and Treatment) Investigate the effects of poison and treatment, using the data at http://www.statsci.org/data/general/poison.html

Solution This dataset records survival times of animals poisoned with one of three poisons and supplied with one of four antidotes. I obtained the following ANOVA table

	DOF	Sum Sq	Mean Sq	F value	Pr(>F)
Poison	2	1.03301	0.51651	23.2217	3.331e-07
Antidote	3	0.92121	0.30707	13.8056	3.777e-06
Poison: Antidote	6	0.25014	0.04169	1.8743	0.1123
Residuals	36	0.80073	0.02224		

From which it seems safe to conclude that there isn't any interaction, but the antidotes work, and the poison does have an effect on survival time. Figure 8.3 drives home these conclusions.

Worked example 8.6 (Memory and Older People) Use Eysenck's data from http://www.statsci.org/data/general/eysenck.html to determine whether Age interacts with Processing, and whether either Age or Processing have significant effects on the number of words memorized

Solution In 1974, Eysenck investigated memory in people. There were two effects of interest. First, the age of the subjects (Age). Eysenck used two groups, one of subjects between 55 and 65 years old, and another of younger subjects. Second, the extent to which material to be memorized was processed (Process). Eysenck used five groups. Each was given a list of words, and asked to perform a task. Afterward, each was asked to write down all the words

(continued)

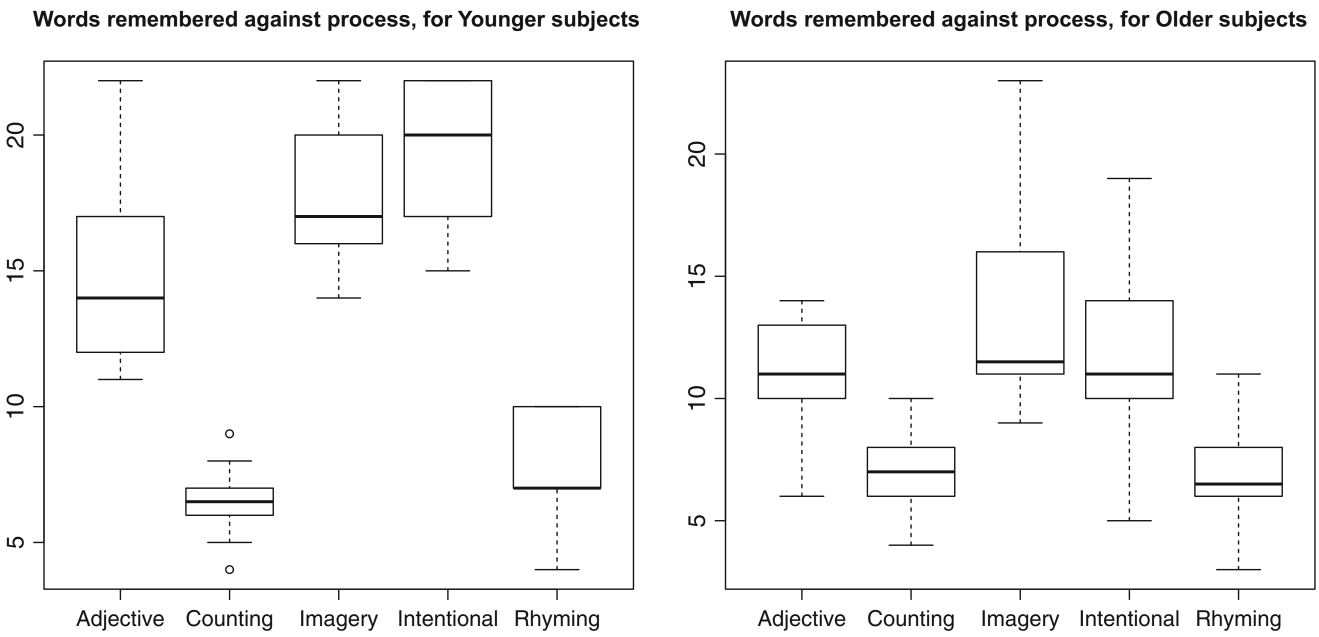

Fig. 8.4 On the *left*, boxplots of the number of words remembered using different processes for younger subjects in the experiment of Worked example 8.6. On the *right*, for older subjects. Generally, if there is no effect of a treatment the boxes in a graph should look the same; if there is no interaction, the pattern formed by the boxes should look the same, with perhaps a shift up or down of all boxes. The interaction between age and process should be clear. Both age groups find Counting and Rhyming hard, but the effect is much more pronounced for younger subjects

they could remember, and the number of words remembered is the response. The Intentional group was told they were to remember the words. The other four groups were asked to perform a task with the words. These four were: Adjective (give an adjective for each word); Counting (count the number of letters in each word); Imagery (form vivid images of each word); and Rhyming (think of a word that rhymed with each word). There are 10 subjects in each group. The original study was published as: Eysenck, M. W. (1974). Age differences in incidental learning. Developmental Psychology, 10, 936–941.

I obtained the ANOVA table

	DOF	Sum Sq	Mean Sq	F value	Pr(>F)
Age	1	240.25	240.25	29.9356	3.981e-07
Process	4	1514.94	378.74	47.1911	< 2.2e-16
Age: Process	4	190.30	47.58	5.9279	0.0002793
Residuals	90	722.30	8.03		

(where the "Age:Process" line is the interaction term). Here there is an interaction, and each treatment has an effect, i.e. Age affects how you remember things, the process by which you interact with them affects it too, and the two factors affect one another. Figure 8.4 shows boxplots which make the interactions pretty clear.

8.3 You Should

8.3.1 Remember These Definitions

8.3.2 Remember These Terms

Randomization	179
balanced	180
within group variation	181
residual variation	181
between group variation	182
treatment variation	182
analysis of variance	182
ANOVA	182
ANOVA table	182
one factor	182
balanced experiment	183
unbalanced experiment	183
contrasts	185
within group mean squares	189
interaction mean squares	189
treatment one mean squares	190
treatment two mean squares	190
two-factor ANOVA	191
two-way ANOVA	191

8.3.3 Remember These Facts

8.3.4 Use These Procedures

Evaluate treatment effects with a one-way ANOVA	182
Set up a two-way ANOVA	191
Evaluate two treatments with a two-way ANOVA	191

8.3.5 Be Able to

- Set up a simple randomized balanced one factor experiment.
- Construct an ANOVA table for the resulting data, and use it to tell whether the treatment has an effect or not.
- Estimate the significance of differences in results for different treatment levels.
- Set up a simple randomized balanced two factor experiment.
- Construct an ANOVA table for the resulting data, and use it to tell whether there are interactions, and whether either treatment has an effect or not.
- Estimate the significance of differences in results for different treatment levels.
- Interpret data from an unbalanced one-factor experiment.

Problems

Decomposing the Squared Error

8.1 You will show that the squared error of a one-way experiment decomposes into two terms, as in Sect. 8.1.2. Write

$$\sum_{ij}(x_{ij}-\hat{\mu})^2 = \sum_{ij}((x_{ij}-\hat{\mu}_i)+(\hat{\mu}_i-\hat{\mu}))^2.$$

(a) Show that the square can be expanded and rearranged to obtain

$$\left[\sum_{ij}(x_{ij}-\hat{\mu}_i)^2\right]+G\left[\sum_{i}(\hat{\mu}-\hat{\mu}_i)^2\right]+2\sum_i\left[\left(\sum_j(x_{ij}-\hat{\mu}_i)\right)(\hat{\mu}-\hat{\mu}_i)\right].$$

(b) Show that, because $\hat{\mu}_i$ minimizes $\sum_j(x_{ij}-\mu_i)^2$,

$$\left(\sum_j(x_{ij}-\hat{\mu}_i)\right)=0$$

(c) Now show

$$\sum_{ij}(x_{ij}-\hat{\mu})^2=\left[\sum_{ij}(x_{ij}-\hat{\mu}_i)^2\right]+G\left[\sum_i(\hat{\mu}-\hat{\mu}_i)^2\right].$$

8.2 You will show that the squared error of a two-way experiment decomposes into four terms, after the pattern of the previous exercise.

(a) Show that

$$\sum_{ijk}(x_{ijk}-\hat{\mu})^2=\sum_{ijk}\begin{bmatrix}(x_{ijk}-\hat{\mu}_{ij})+\\(\hat{\mu}_{i\cdot}-\hat{\mu})+\\(\hat{\mu}_{\cdot j}-\hat{\mu})+\\(\hat{\mu}_{ij}-\hat{\mu}_{i\cdot}-\hat{\mu}_{\cdot j}+\hat{\mu})\end{bmatrix}^2.$$

(b) Show that

$$\sum_k(x_{ijk}-\hat{\mu}_{ij})=0$$

for any i, j, by recalling the definition of $\hat{\mu}_{ij}$.

(c) Show that

$$\sum_i(\hat{\mu}_{i\cdot}-\hat{\mu})=0$$

for any i, by recalling the definition of $\hat{\mu}_{i\cdot}$ and $\hat{\mu}$.

(d) Show that

$$\sum_j(\hat{\mu}_{j\cdot}-\hat{\mu})=0$$

for any j, by recalling the definition of $\hat{\mu}_{j\cdot}$ and $\hat{\mu}$.

(e) Now show that

$$\sum_{ijk}(x_{ijk}-\hat{\mu})^2 = \begin{bmatrix} \sum_{ijk}(x_{ijk}-\hat{\mu}_{ij})^2+ \\ GL_2\sum_i(\hat{\mu}_{i\cdot}-\hat{\mu})^2+ \\ GL_1\sum_j(\hat{\mu}_{\cdot j}-\hat{\mu})^2+ \\ G\sum_{ij}(\hat{\mu}_{ij}-\hat{\mu}_{i\cdot}-\hat{\mu}_{\cdot j}+\hat{\mu})^2 \end{bmatrix}.$$

Unbalanced One-Way Experiments

8.3 The Rules of Rugby You will find a dataset giving the times of passages of play in games of rugby at http://www.statsci.org/data/oz/rugby.html. The data was collected by Hollings and Triggs in 1993. The first five games were played under the old rules, and the second five under the new rules. Use an unbalanced one-way ANOVA to determine whether the change of rules made a difference in the times of passages of play.

8.4 Eye Color and Flicker Frequency You will find a dataset recording the critical flicker frequency and eye color for 19 individuals at http://www.statsci.org/data/general/flicker.html. The data was collected by Devore and Peck in 1973. Use an unbalanced one-way ANOVA to determine whether individuals with different eye colors have different critical flicker frequencies.

8.5 Survival on the Titanic You will find a dataset recording the survival status of passengers on the Titanic at http://www.statsci.org/data/general/titanic.html. This data comes originally from *Encyclopedia Titanica*, by Philip Hinde. Use an unbalanced one-way ANOVA to determine whether the class of the passenger's ticket had an effect on their survival.

Two-Way Experiments

8.6 Paper Planes You will find a dataset recording the performance of paper planes at http://www.statsci.org/data/oz/planes.html. This data comes originally from the paper *What is the use of experiments conducted by statistics students?*, by M.S. Mackisack, which appeared in the Journal of Statistics Education in 1994. Use a two-way ANOVA to analyze the effects of paper and angle on the distance covered by the plane. Is there an interaction between these variables?