Research Methods II, Spring Term 2003

Logic of repeated measures designs

Imagine you want to test the effect of drinking Fosters on the time it takes to respond to a light turning red. The independent variable, amount of Fosters drunk, has two levels: zero pints and one pint of Fosters. The dependent variable is reaction time. Let me indicate how well each subject did with a dot, where the height of the dot indicates their reaction time:

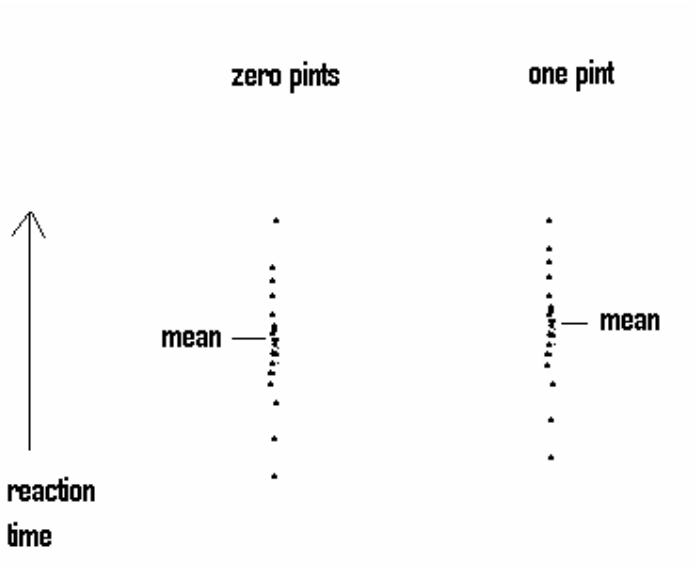

Figure 1.

Imagine this is a between subjects design, i.e. each subject contributes just one dot. Looking at Figure 1, you can see there is a small difference in the group means. Is this a real difference? I have tried to draw a considerable spread of scores round the mean (within-group variance) so as to make you doubt whether such sample mean difference as does exist should be taken to reflect a real population difference.

Now consider the following hypothetical results in Figure 2, with the same group means, the same number of subjects, but different spread about the mean:

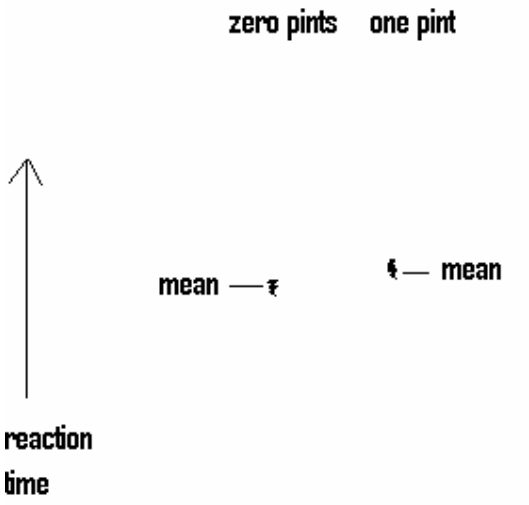

Figure 2

In Figure 2 the spread of scores about the mean is very small. I hope your intuition now is that if you obtained these results in an experiment, rather than those in Figure 1, you would be much more likely to conclude, just looking at the data, that Fosters does affect reaction time.

In ANOVA you are trying to see a signal through noise. The signal is the population difference. We estimate the signal with our sample mean difference. For between-subjects, the noise is the within-group variability. The bigger the within-group variability the harder it is for us to discern a signal. In Figure 1, there is a lot of noise, so we doubt we are really seeing any signal; in Figure 2 the noise is small, and we pick up the same signal easily.

Now imagine we obtained the same data as in Figure 1, but this time the design was a within-subjects design. That is, each subject contributes two dots, one in each condition. Figure 3 presents the same data as Figure 1, but with the corresponding dots for the same subject joined up with a line.

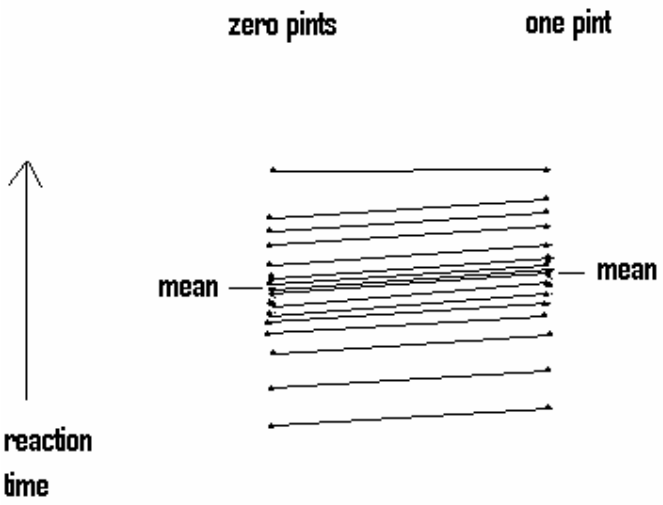

Figure 3.
If you saw these data, would you be inclined to conclude that Fosters does really affect reaction time? Notice that for virtually every subject, the effect of Fosters is the same: To slow down reaction time by the same amount, e.g. 100 ms. The effect might be small but it is very consistent. I hope you feel that now the evidence supports the conclusion that Fosters really does affect reaction time.

The within-group variability is now irrelevant to how willing we to draw the conclusion that Fosters affects reaction time. In Figure 4, I have added a subject far from the mean; but rather than making us think the data are more noisy, the consistency with which all subjects are affected by Fosters allows us to see the signal very clearly.

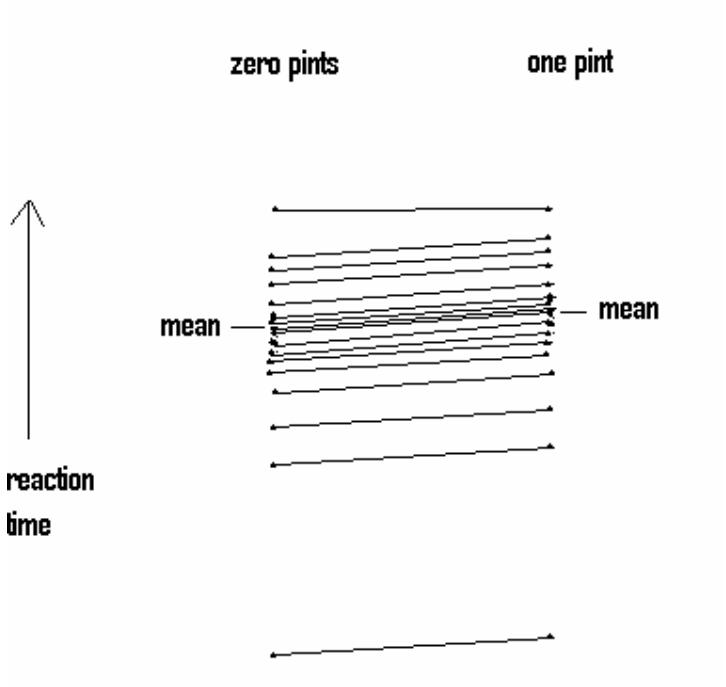

Figure 4.

In summary, in within-subjects (repeated measures) designs, within-group variability is no longer the measure of noise through which we are trying to see the signal. Now what is relevant is the consistency with which the manipulation affects each subject. Consider Figure 5, showing the same data as Figures 1 and 3, but this time the dots belonging to the same subject are paired differently:

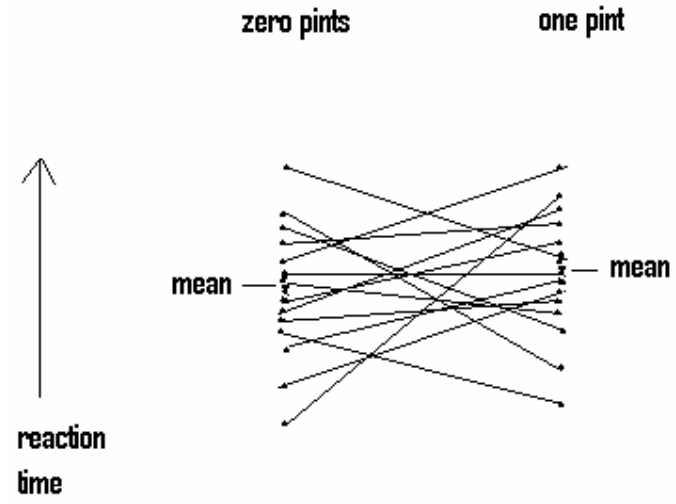

Figure 5.

If these were your data, would you think that Fosters affects reaction time? Same mean difference, same estimate of signal. But now the data seem very noisy, and we are inclined to think there isn't really a signal there at all. We wouldn't say with any confidence that Fosters affects reaction time.

In summary, in the within-subjects case, the relevant measure of noise is the inconsistency with which the manipulation affects subject. The more inconsistent the effect over subjects, the more the noise.

Notice geometrically how this inconsistency manifests itself. In Figure 3, the inconsistency is low and the lines are parallel. In Figure 5, the inconsistency is large and the lines are very much non-parallel. When I say "parallel lines" what do you think of? Remember non-parallel lines means interaction.

The interaction we are interested in is an unusual one, because one of the variables is "subjects". It has levels of Tom, Maria, Jane, and so on. In Figure 5 there is an interaction between subjects and amount of Fosters. In Figure 3 there is virtually no interaction at all.

In a within-subjects design, noise is measured by the interaction of subjects and the independent variable. The higher this interaction, the more the inconsistency, the greater the noise.

In the within-subjects case, ANOVA proceeds like the between-subjects case, but this time mean square error (the estimate of the noise through which you are trying to see the signal) is not the within-group variability but the interaction between subjects and treatment (treatment is another word for the independent variable). Mean square treatment, the between condition variability, is the same in the within-subjects case as in the between-subjects case. (Mean square treatment is the estimate of signal, as seen though noise). The F ratio is the ratio of mean square treatment to mean square error. If the null hypothesis is true, F is expected to be about 1.

If subjects are affected by the independent variable consistently, than a within-subjects design brings with it gains of sensitivity compared to a between-subjects design. You no longer care that some subjects are overall e.g. very fast or very slow; each subject is compared with themselves.

<u>Assumptions of within-subjects designs.</u>

Remember in the between-subjects case, ANOVA required two assumptions: Homogeneity of variance (the variances within each group are about the same size as each other), and a normal distribution of scores within each group. There are exactly equivalent assumptions in the repeated measures case.

Consider first the assumption of homogeneity of variance. In the between-subjects case, we assume the variances within each group (which are our separate measures of noise) are the same (thereby allowing us to combine them into one overall measure of noise, mean square error). In the repeated measures case, noise is measured by the inconsistency with which the treatment affects the subjects. Inconsistency can be measured by first taking the difference between the two conditions for each subject. If this difference is the same for each subject, the treatment has had a consistent effect. Conversely, if there is a large variance in the difference scores, the treatment has had an inconsistent effect. Variance between difference scores means inconsistency (means subject*treatment interaction).

Imagine you had three conditions: (0) (0 pints of Fosters), (1) (1 pint of Fosters), and (2) (2 pints of Fosters). Now there are three measures of inconsistency. There is the inconsistency of the effect of having 0 vs 1 pint (measured by the variance of the (0) - (1) difference scores, call this variance 1); the inconsistency of the effect of having 0 vs 2 pints (measured by the variance of the (0) - (2) difference scores, call this variance 2); and the inconsistency of the effect of having 1 vs 2 pints (the variance of the (1) - (2) difference scores, call this variance 3). Between-subjects designs assume homogeneity of variance; within-subjects designs have an equivalent assumption, namely that variance 1, variance 2, and variance 3 (our three measures of

noise) are equal (thereby allowing us to combine them into one measure of noise, mean square error). This is called the sphericity assumption.

In addition, we will assume that the population scores in each condition are normally distributed, just as in the between-subjects case. (In fact, the repeated measures case only assumes that the population difference scores for each pair-wise comparison, e.g. treatment 1 with treatment 2, are normally distributed. But if the scores themselves are normally distributed, the differences between the scores necessarily are normally distributed as well.)

Carry-over effects

Repeated measures designs can give you a gain in sensitivity over between-subjects designs. But they can give you problems as well. Because you are testing subjects two (or more) times, there can be "carry over" effects. Practice effects mean the subjects get better at the task just because they are doing the same task repeatedly. Fatigue and boredom effects mean the subjects get worse on the second lot of testing.

You should counterbalance the order of the tasks to at least partially address these problems. Run as many subjects with condition 1 first as condition 2 first. (If you have more than two conditions, there are various ways of balancing order: Look up "Latin square" in a stats book if your project involves repeated measures with more than 2 levels.)

But counterbalancing does not solve "differential carry over effects." One can be hung over after being in the 4 pints condition, impairing performance in the subsequent 0 pints condition, but there is not such a carry over effect going from the 0 pints to 4 pints condition. If you learn the Method of Loci first in a memory experiment, a subject may feel tempted to use it surreptitiously, to great benefit, in a subsequent rote rehearsal condition; but there won't be such a carry over effect when the tasks are in the opposite order. Sometimes doing a task once contaminates the subject; they are not naive and cannot be tested again.

So within-subjects designs are inherently problematic. You must make a judgement call as to whether there are problematic carry-over effects for any experiment you design; if you think there could be, use the statistically purer between-subjects design. On the other hand, if you think there should not be differential carry over effects, you might gain in sensitivity by using the repeated measures design. If you use the repeated measures design and it looks messy you can always disregard all but the first lot of testing and treat it as a between-subjects design.