One-Way Repeated Measures ANOVA by Hand

Repeated Measures vs. Between-group designs

When we collect data, we choose between two methods of data collection. The first is to manipulate the independent variable using different participants: so, different groups of people take part in each experimental condition (a *between-group* or *between-subjects* design). The second method is to manipulate the independent variable using the same participants. Simplistically, this method means that we test the same group of participants in all experimental conditions (a *within-subject* or *repeated measures* design). The way in which the data are collected determines the type of test that is used to analyze the data.

Just like the *t*-test, there are different versions of ANOVA designed to deal with repeated measures designs. The rationale for ANOVA is the same regardless of whether the data are collected using a between-group or a repeated measures design. That is, we carry out some experimental manipulation. We expect this manipulation to create variation in scores across conditions (*systematic variation*). However, we also expect there to be variation caused by individual difference in performance and other random factors that we cannot account for (*unsystematic variation*). In the ANOVA we want to work out how much variation there is in total and then calculate how much of this variation was caused by our experimental manipulation and how much is due to other factors.

Repeated Measures ANOVA

If we stick to a simple example in which there are only two experimental conditions and a repeated measures design has been used, the same participants participate in both conditions. So, we measure subject's behaviour in condition 1 *and* in condition 2. If there is no experimental manipulation then we expect a person's behaviour to be roughly the same across conditions. We expect this because external factors such as age, gender, IQ, motivation and arousal will be the same in every condition (a person's gender etc. will not change from when they are tested in condition 1 to when they are tested in condition 2). If the performance measure is reliable, and the variable or characteristic that we are measuring remains stable over time, then it is probable that a subject's performance in condition 1 will be very highly related to their performance in condition 2. So, participants who score highly in condition 1 will also score highly in condition 2.

When we introduce an experimental manipulation, we do something different to participants in condition 1, to what we do to them in conditions 2. So, the only difference between conditions 1 and 2 is the manipulation that the experimenter has made. Therefore, any differences between the means of the two conditions are probably due to the experimental manipulation.

Of course humans are not robots (well, apart from statistics lecturers) and so a person's performance would not be identical across conditions if there were no experimental manipulation. As such, when we look at the variation caused by our experimental manipulation, we still need to be aware that some of that variation will be caused by unknown factors. As such, in repeated measures ANOVA we have to look at the variation across conditions for a given participant (as a measure of the effect of our experiment) but also gauge the error within that variance.

Benefits of Repeated Measures Designs

Sensitivity: The effect of our experimental manipulation is likely to be more apparent in a repeated measures design than in a between-group design because in the former unsystematic

variation can be caused only by differences in the way in which someone behaves at different times. In between-group designs we have differences in innate ability contributing to the unsystematic variation. Therefore, this error variation will almost always be much larger than if the same participants had been used. When we look at the effect of our experimental manipulation, it is always against a background of 'noise' caused by random, uncontrollable differences between our conditions. In a repeated measures design this 'noise' is kept to a minimum and so the effect of the experiment is more likely to show up. This means that repeated measures designs have more power to detect effects that genuinely exist than independent designs.

Economy: Repeated measures designs make more efficient use of participants and so save time and money. However, although in theory you could have a participant take part in many different conditions, they do tend to get very bored and frustrated in long experiments. Therefore, it's always worth trying to bear in mind what your participants will have to endure before designing an experiment with 250 different experimental conditions☺

Disadvantages of Repeated measures Designs

If the same people participate in more than one experimental condition, although they are naive during the first experimental condition they come to the second experimental condition with prior experience of what is expected of them. At the very least they will be familiar with the dependent measure. The two most important sources of systematic variation in this type of design are:

- **Practice effects**: Participants may perform differently in the second condition because of familiarity with the experimental situation and/or the measures being used.
- **Boredom**: Participants may perform differently in the second condition because they are tired or bored from having completed the first condition.

Although these effects are impossible to eliminate completely, we can ensure that they produce no systematic variation between our conditions. We can do this by *counterbalancing* the order in which a person participates in a condition. We can randomly decide that a subject either has condition 1 before condition 2, or that they have condition 2 before condition 1 (see Field, 2000, chapter 6 for more detail). One type of counterbalancing is the Latin Square, in which all conditions appear in all locations within the design for a subgroup of participants. Below is an example of a Latin Square for 3 and 4 repeated measures conditions:

	Time 1	Time 2	Time 3
Order 1	**A**	**B**	**C**
Order 2	**B**	**C**	**A**
Order 3	**C**	**A**	**B**

Latin Square for 3 conditions (A, B & C)

	Time 1	Time 2	Time 3	Time 4
Order 1	**A**	**B**	**C**	**D**
Order 2	**B**	**C**	**D**	**A**
Order 3	**C**	**D**	**A**	**B**
Order 4	**D**	**A**	**B**	**C**

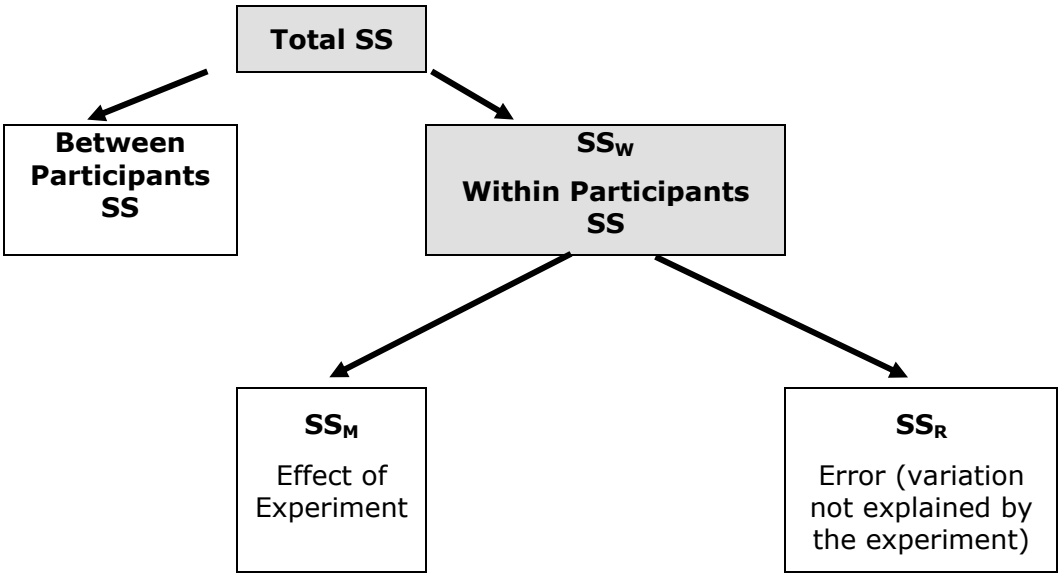
Latin Square for 4 conditions (A, B, C & D)

Another alternative is to randomise the order of conditions for a given participant and hope that across all participants each condition will occur at a variety of positions within the experiment.

Partitioning Variance

In a repeated measures ANOVA the effect of our experiment is shown up in the within-subject variance (rather than in the between-group variance). Some of the within-participants variation comes from the effects of our experimental manipulation: we did different things in each experimental condition to the participants, and so variation in an individual's scores will partly be due to these manipulations. For example, if everyone scores higher in one condition than another, it's reasonable to assume that this happened not by chance, but because we did

something different to the participants in one of the conditions compared to what we did to them in the other one. *Because* we did the *same* thing to everyone within a particular condition, any random (unsystematic) variation across participants is probably due to random factors outside our control, unrelated to our experimental manipulations (we could call this 'error'). As in independent ANOVA, we use an *F*-ratio that compares the size of the variation due to our experimental manipulations, to the size of the variation due to random factors. If the former is big relative to the latter, we get a big value of *F*, and we can conclude that the observed results are unlikely to have arisen by chance.



Calculating Repeated Measures ANOVA by Hand

Example: There is a lot of concern among students as to the consistency of marking between lecturers. It is pretty common that lecturers obtain reputations for being 'hard markers' or 'light markers' but there is often little to substantiate these reputations. So, a group of students investigated the consistency of marking by submitting the same essay to four different lecturers. The mark given by each lecturer was recorded for each of the 8 essays. It was important that the same essays were used for all lecturers because this eliminated any individual differences in the standard of work that each lecturer was marking. This design is, therefore, repeated measures because every lecturer marked each essay. The independent variable (IV) was the lecturer who marked the report and the dependent variable (DV) was the percentage mark given (see Table 1).

Table 1: Essay marks

Essay	Tutor 1 (Dr. Field)	Tutor 2 (Dr. Smith)	Tutor 3 (Dr. Scrote)	Tutor 4 (Dr. Death)	Mean	Variance
1	62	58	63	64	61.75	6.92
2	63	60	68	65	64.00	11.33
3	65	61	72	65	65.75	20.92
4	68	64	58	61	62.75	18.25
5	69	65	54	59	61.75	43.58
6	71	67	65	50	63.25	84.25
7	78	66	67	50	65.25	132.92
8	75	73	75	45	67.00	216.00
Mean:	**68.875**	**64.25**	**65.25**	**57.375**		

Step 1: Calculate SS_T

Remember from one-way independent ANOVA that SS_T is calculated using the following equation:

$$SS_T = s^2_{grand}(N-1)$$

The grand variance is simply the variance of all scores when we ignore the group to which they belong. So if we treated the data as one big group it would look as follows:

62	58	63	64
63	60	68	65
65	61	72	65
68	64	58	61
69	65	54	59
71	67	65	50
78	66	67	50
75	73	75	45

Grand Mean = 63.9375

The variance of these scores is **55.028** (try this on your calculators). We used 32 scores to generate this value, and so *N* is 32. As such the equation becomes:

$$\begin{aligned} SS_T &= s^2_{grand}(N-1) \\ &= 55.028\,(32-1) \\ &= 1705.868 \end{aligned}$$

The degrees of freedom for this SS will be N–1, or 31.

Step 2: Calculate the Within-Participants SS

The crucial difference with a repeated measures design is that the variance in which we're interested occurs within participants (because we've manipulated our independent variable within each participant). When we calculated sums of squares we look at the squared difference between the mean and individual scores. This can be expressed in terms of the variance across a number of scores and the number of scores on which the variance is based. For example, when we calculate the residual sum of squares in independent ANOVA (SS_R) we used the following equation:

$$\begin{aligned} SS_R &= \sum (x_i - \bar{x}_i)^2 \\ SS_R &= s^2(n-1) \end{aligned} \qquad (1)$$

For each group. To get the SS_R across all groups we calculate this for each group and then add them up:

$$SS_R = s^2_{group1}(n_1 - 1) + s^2_{group2}(n_2 - 1) + s^2_{group3}(n_3 - 1)$$

This equation can be adapted to show how we work out how much variability within participants has to be explained. If we call this sum of squares SS_W (for Within-participants SS), and apply the same equation, get the following.

$$SS_W = s^2_{person1}(n_1 - 1) + s^2_{person2}(n_2 - 1) + s^2_{person3}(n_3 - 1) \ldots + s^2_{person\,n}(n_n - 1) \qquad (2)$$

This equation simply means that were looking at the variation in an individuals score and then adding these variances for all the people in the study. The *n*s simply represent the number of

scores on which the variances are based (i.e. the number of experimental conditions). We can use the variances in Table 1 to give us:

$$\begin{aligned}SS_W &= s^2_{subject1}(n_1 - 1) + s^2_{subject2}(n_2 - 1) + s^2_{subject3}(n_3 - 1)\ldots \\ &= (6.92)(4-1) + (11.33)(4-1) + (20.92)(4-1) + (18.25)(4-1) \\ &\quad + (43.58)(4-1) + (84.25)(4-1) + (132.92)(4-1) + (216)(4-1) \\ &= 20.76 + 34 + 62.75 + 54.75 + 130.75 + 252.75 + 398.75 + 648 \\ &= 1602.5\end{aligned}$$

The degrees of freedom for each person are n–1 (i.e. the number of conditions minus 1). To get the total degrees of freedom we add the df for all participants. So, with 8 participants (essays) and 4 conditions (i.e. $n = 4$) we get $8 \times 3 = 24$ degrees of freedom.

Step 3: Calculate the Model Sum of Squares (SS_M)

So far, we know that the total amount of variation within the data is 1705.868 units. We also know that 1602.5 of those units are explained by the variance created by individuals' performances under different conditions. Now some of this variation is the result of our experimental manipulation and some of this variation is simply random fluctuation. The next step is to work out how much variance is explained by our manipulation and how much is not.

In independent ANOVA, we worked out how much variation could be explained by our experiment (the Model SS) by looking at the means for each group and comparing these to the overall mean. So, we measured the variance resulting from the differences between group means and the overall mean. We do the same thing with a repeated measures design. First we calculate the mean for each level of the independent variable (in this case the mean mark given by each lecturer), and compare these values with the overall mean of all marks.

So, we calculate this SS in the same way a for independent ANOVA:

1. *Calculate the difference between the mean of each group and the grand mean.*
2. *Square each of these differences.*
3. *Multiply each result by the number of participants within that group (*n_i*).*
4. *Add the values for each group together.*

$$SS_M = \sum n_i(\bar{x}_i - \bar{x}_{grand})^2 \qquad (3)$$

Using the means from the essay data, we can calculate SS_M as follows:

$$\begin{aligned}SS_M &= 8(68.875 - 63.9375)^2 + 8(64.25 - 63.9375)^2 + 8(65.25 - 63.9375)^2 + 8(57.375 - 63.9375)^2 \\ &= 8(4.9375)^2 + 8(0.3125)^2 + 8(1.3125)^2 + 8(-6.5625)^2 \\ &= 554.125\end{aligned}$$

For SS_M, the degrees of freedom (df_M) are again one less than the number of things used to calculate the sum of squares. For the model sums of squares we calculated the sum of squared errors between the **four** means and the grand mean. Hence, we used four things (the means at each level of the IV) to calculate these sums of squares. So, the degrees of freedom will be 3. So, as with independent ANOVA the model degrees of freedom is always the number of *groups* (*k*) minus 1:

$$df_M = k - 1 = 3$$

Step 4: Calculate the Residual Sum of Squares (SS_R)

We now know that there are 1706 units of variation to be explained in our data, and that the variation across our conditions accounts for 1602 units. Of these 1602 units, our experimental manipulation can explain 554 units. The final sum of squares is the residual sum of squares (SS_R), which tells us how much of the variation cannot be explained by the model. This value is the amount of variation caused by extraneous factors outside of experimental control. Knowing SS_W and SS_M already, the simplest way to calculate SS_R is to subtract SS_M from SS_W ($SS_R = SS_W - SS_M$).

$$\begin{aligned} SS_R &= SS_W - SS_M \\ &= 1602.5 - 554.125 \\ &= 1048.375 \end{aligned}$$

The degrees of freedom are calculated in a similar way:

$$\begin{aligned} df_R &= df_W - df_M \\ &= 24 - 3 \\ &= 21 \end{aligned}$$

Step 5: Calculate the Mean Squares

SS_M tells us how much variation the model (e.g. the experimental manipulation) explains and SS_R tells us how much variation is due to extraneous factors. However, because both of these values are summed values the number of scores that were summed influences them. As with independent ANOVA we eliminate this bias by calculating the average sum of squares (known as the *mean squares*, MS), which is simply the sum of squares divided by the degrees of freedom:

$$MS_M = \frac{SS_M}{df_M} = \frac{554.125}{3} = 184.708$$

$$MS_R = \frac{SS_R}{df_R} = \frac{1048.375}{21} = 49.923$$

MS_M represents the average amount of variation explained by the model (e.g. the systematic variation), whereas MS_R is a gauge of the average amount of variation explained by extraneous variables (the unsystematic variation).

Step 5: Calculate The F-Ratio

The *F*-ratio is a measure of the ratio of the variation explained by the model and the variation explained by unsystematic factors. It can be calculated by dividing the model mean squares by the residual mean squares.

$$F = \frac{MS_M}{MS_R} \qquad (4)$$

As with the independent ANOVA, the *F*-ratio is still the ratio of systematic variation to unsystematic variation. As such, it is the ratio of the experimental effect to the effect on performance of unexplained factors. For the marking data, the *F*-ratio is:

$$F = \frac{MS_M}{MS_R} = \frac{184.708}{49.923} = 3.70$$

This value is greater than 1, which indicates that the experimental manipulation had some effect above and beyond the effect of extraneous factors. As with independent ANOVA this value can be compared against a critical value based on its degrees of freedom (in this case 3 and 21).

Using the Table from Field 2004 (see Handout on One-Way Independent ANOVA) we need to read across to 3 degrees of freedom (model degrees of freedom) and then read down to 21

(residual degrees of freedom). In fact Wright reports only 21 and 25 degrees of freedom so, to be safe, we should use 20 because this will give us a more conservative estimate (i.e. we are less likely to accept a result that is false). The value for a 0.05 probability level is 3.10 and the value for a 0.01 probability level is 4.94. If our experimental manipulation has been successful (in other words, if our experimental manipulation has caused some groups to behave differently to others) we expect to find an F-ratio greater than these values. The value of *F* we calculated was 3.7, and because this is bigger than the critical value of 3.10, we can say that the person marking the essay did have a significant effect on the mark it received. However, note that because the observed value is smaller than 4.94 (the critical value for a 0.01 significance value) we know that this *F*-value is not significant at the stricter 0.01 level. We can report this by saying 'The lecturer marking the essay had a significant effect on the mark that essay received [F (3, 21) = 3.70, $p < 0.05$]'. Note that we quote the degrees of freedom associated with the *F*-ratio and the probability (p) of obtaining that *F*-value by chance.

This handout is based on material from:

Field, A. P. (2004). *Discovering statistics using SPSS: advanced techniques for the beginner (second edition)*. London: Sage.

It is © Andy Field, and if you want the new chapter then drop me an email (although the old version of the chapter is perfectly good!).