Chapter 12: Statistics 2

## Learning objectives

* Carrying out an ANOVA in R/RStudio
* Reporting the ANOVA results in APA format, extracting key numbers from the output table
* Making figures for factorial designs
* Modern reproducibility theory: effect sizes, power, sensitivity

# Manual for running a 2x2 ANOVA in R

Review the data in Excel using the .xlsx file. Calculate descriptive statistics as we did with our earlier analysis of Experiment 1 with the simpler two group design (these may also be provided). Note that there are now four conditions to calculate condition means, SD and SE. You should also examine marginal means, where two conditions are combined, so that you can observe the magnitude of the main effects.

To run the analysis, use the RStudio program to start an analysis session

* Launch RStudio.
* Use File -> Open and navigate to the folder on your computer where you’ve installed the 205 files and associated data from our experiments
* Open the file “205\_Exp2\_ANOVA.R”
* Set the “working directory” to where your data are stored on your computer. If you have put the “exp2\_data\_Fall2022.csv” in the same folder as the “205\_Exp2\_ANOVA.R” file, navigate to the Session menu, then to Set Working Directory and select the top option “To Source File location.”
* To run a single step of the analysis press the “-> Run” button that is in the upper right part of the top-left panel. This carries out the step in the script on which the cursor is currently.  If you didn’t do the installation of the ‘psych’ and ‘ez’ packages above, put the cursor on line 2 and Run.  Then put the cursor on line 3 and Run.
* The installation process will also download and install a series of other packages needed (called dependencies). The process should only take a few minutes to run.  You may get red warning messages (e.g., if you have already installed these) but it's working if you get a 'successfully unpacked' message as well.
* Now move down to line 6, “library(psych)” and press Run. This loads a set of routines for data analysis for psychology experiment data that are helpful.
* The cursor moves down to the next line after each Run. Press it again to load the library on line 7, 8, and 9 (‘psychTools’, ‘tidyr’, and ‘ez’).  You may again get warning messages about prior R versions, but everything will still run.
* The next step, line 12 will start loading our actual data. If everything is working you should see: “Data from the .csv file exp2\_data\_Fall2022.csv has been loaded.” In red in bottom left panel.
* Run on line 14 to see the output of the describeBy function, which provides descriptive statistics for our data.  Check that these numbers are identical to the descriptive statistics you calculated the Excel version of the data, Exp2\_data\_Fall2022.xlsx.
* Run on line 16 to carry out the ANOVA on these data.  There will be some warning messages even if it runs successfully.
* Run on line 30 to print the ANOVA table

# How to format the report of a 2x2 ANOVA

The output of the ezANOVA program is a table of information containing the inferential statistics that need to be reported as the results of the analysis. The statistical parameter resulting from an ANOVA analysis is an F-ratio, typically written as F. The F statistic is reported with two degrees of freedom. First the numerator df (df-n) which is related to the number of levels within the condition being reported on. This number is number-of-levels minus 1, which is always 1 for a 2x2 design. The second df is the denominator (df-d), which is related to the number of participants in the study across all conditions. For a 2x2 design, this number is the total N minus 4 (the df is reduced by 1 and then 1 more for each of the 3 hypotheses being tested). In a written description of the results the format is F(df-n,df-d) = X.XX, p<.YY.

As with all other inferential statistics, we also obtain a p-value which means the probability of having observed the difference occurring in the data under the null hypothesis. We use the same standard criterion for this, p<.05. In many cases, the ANOVA output table must be read carefully to make sure that you are using the correct p-value. Other information might be included that is also on a 0.0-1.0 scale that can look visually like a p-value.

There will be 3 lines reporting results for a 2x2 ANOVA. The first line will report the first main effect, i.e., the difference between levels of that condition ignoring the other factor. The second line will be the other main effect. The third line is the interaction term, typically listed as something like “Factor1:Factor2” and will tell you whether there is a reliable influence across factors. Importantly, the statistical report of the interaction term does not tell you anything about what the interaction is. It could be a reliable super-additive interaction, a 3:1 interaction (or sub-additive) or a cross-over interaction. Just looking at the ANOVA table cannot tell you which. It is necessary to review the descriptive statistics to interpret the interaction.

## Example

In the example below, we have simulated data from a hypothetical experiment in which participants were given anagrams to solve which were either hard or easy (difficulty factor) and written in either red or black ink (color factor). After solving the puzzles participants were offered candy and the number of pieces of chocolate taken was scored as the dependent variable (a potentially questionable operational definition of stress, but this is not a real experiment).

The output of analysis using R/Rstudio is shown in the table below. First the descriptive statistics are shown for the four conditions. Although it is somewhat more challenging to read this format of output table than the traditional means table shown previously, these data as they would be provided by general use statistical tools are provided for practice. When preparing a formal report of the results of an experiment, it is highly recommended to make either a figure or means table report of the condition means before writing out the results of an analysis. That will simplify the process of identifying the direction of the observed effects so that they can be communicated clearly in the Results section of a written manuscript.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Descriptive statistics by group  : Black  : Easy  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 20 3 0.79 3 3 1.48 2 4 2 0 -1.5 0.18 | | | | | | | |
| : Red  : Easy  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 20 2.75 1.16 3 2.81 1.48 1 4 3 -0.3 -1.48 0.26 | | | | | | | |
| : Black  : Hard  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 20 3.2 0.77 3 3.19 0 2 5 3 0.34 -0.3 0.17 | | | | | | | |
| : Red  : Hard  vars n mean sd median trimmed mad min max range skew kurtosis se  X1 1 20 3.8 0.89 4 3.81 1.48 2 5 3 -0.05 -1.14 0.2 | | | | | | | |
|  | | | | | | | |
| > print(anova\_result)  $ANOVA | | |  |  |  |  |  |
|  | Effect | DFn | DFd | F | p | p<.05 | ges |
| 2 | Color | 1 | 76 | 0.725643 | 0.396975415 |  | 0.009457634 |
| 3 | Difficulty | 1 | 76 | 9.255651 | 0.003222014 | \* | 0.108563488 |
| 4 | Color:Difficulty | 1 | 76 | 4.279813 | 0.041968548 | \* | 0.053311197 |
|  |  |  |  |  |  |  |  |
| $`Levene's Test for Homogeneity of Variance`  DFn DFd SSn SSd F p p<.05  1 3 76 2.2375 26.95 2.103278 0.1068009 | | | | | | | |

The key statistical information is in the table following the heading “$ANOVA.” The first effect reported is the main effect of Color (line following “2”). The F column contains the F-ratio and the two columns to the left indicate the degrees of freedom. This would be written as F(1,76) = 0.72. The p-value for this main effect is 0.39, which is >0.05 and not a statistically reliable effect. The next column headed “p<.05” is blank for non-reliable effects and contains a “\*” for reliable effects. The last column ‘ges’ is the generalized effect size related to *partial eta-squared*. This is a standardized effect size for which a ‘small’ effect is around 0.01, a ‘medium’ sized effect is around 0.06 and a ‘large’ effect is greater than 0.14. The scale for this effect size measure is somewhat different than other common standardized effect sizes (e.g., Cohen’s d), which is a challenge in using/reading these in published research. In all effect size systems, the null hypothesis is an effect size of exactly zero. Here we can see that the estimated Color effect size is smaller than ‘small,’ which is why it is not reliable. The Difficulty effect is medium to large, and the interaction between factors is medium-sized.

## Making a 2x2 figure

To make a figure with 2x2 data in MS Excel, start by creating a labeled means table with just the mean performance in each condition. For the data above, that would look like this.

Table

Description automatically generated

Select all 9 cells (including labels) and Insert either a line or bar chart from Excel options. Then you can format the resulting figure (called ‘chart’ in Excel) the same way as we did for the earlier Experiment 1 data. Make sure the axes are visible and y-axis labeled and add error bars for each of the two series of data. The resulting line graph will look something like this:

Chart, line chart

Description automatically generated

The most challenging part of editing the graph is adding the error bars to reflect our calculated SE of the mean. Do this by creating another table of SE values. Then select the data series, Add Chart Element -> Error Bars and select the bottom option Custom Error bars and set both the positive and negative values to the SE values of that series.

Table

Description automatically generated

# Modern Reproducibility Theory

You may have heard that psychology, as well as a variety of other scientific domains, is currently experiencing a “replicability crisis.” This has been inspired by a series of attempts to replicate well-known findings that have not produced reliable differences among conditions that were originally observed as reliable. There are substantial issues with the replication methodology that has been used that likely indicate that the term “crisis” is more extreme than warranted. However, the concern has usefully drawn attention to some aspects of how we carry out statistical inference in psychological science that we can use to improve our overall scientific progress.

## Replicability outside of Psychological Science

The most substantial and worrisome concerns about replicability have come from areas in which research findings have the most direct impact on people’s choices, behavior and policy. Many of these arise form research in which there is an issue of “conflict of interest” between the science and the researchers carrying out the science. Scientific questions that impact things people buy (e.g., nutritional supplements, pharmacological agents, educational tools) is sometimes carried out by researchers who have financial interests affected by the outcome of research. It can be very difficult to maintain fully unbiased research practices when there are substantial financial incentives for specific outcomes. Medical research is usually done with very thorough external oversight to correct for this issue (but not always, sadly). Research in subfields such as nutrition or educational tools, however, is often not given as much attention and as a result, the replicability crisis in some other fields is a bigger issue.

The statistical model we have used so far reflects the approach used in the bulk of psychological research aimed at rejecting the null with a criterion of p<.05 (less that 5% chance of the data appearing as observed if the null hypothesis were true). This leads to is reporting results with a binary outcome: either the effect was reliable or not. There are several difficulties created by trying to make the outcome as simple as yes/no.

**“Marginal” effects**. It is not uncommon for research to be carefully carried out, analyzed properly and find that the probability of rejecting the null does not meet the .05 threshold but is instead in the range of .051 to 0.10. This poses some challenges for drawing interpretations of the results. We cannot claim that the results are reliable because they are not. However, the null hypothesis has actually been found to be somewhat improbable so simply saying that the effects are not reliable seems to miss an important aspect of the data. The simple binary model does not provide guidance for how to deal with these kinds of results.

**Miniscule effects**. It is also possible to have a statistically reliable effect that is actually extremely small. For example, if we found that an alternate studying method led to an reliable increase in memory performance of 1% accuracy, we would have a significant but somewhat uninteresting effect. This problem is fairly uncommon in experimental work as even small effects can have theoretical implications, but comes up in more applied research or in some large-scale non-experimental studies. Here the simple binary model does not help us explain a reliable but not particularly useful effect.

**Null findings**. Sometimes our experimental hypothesis depends on providing evidence for a null effect. For example, we might want to show that sugar does not lead to hyperactivity in children. The simple binary model does not provide a method for evaluating this hypothesis since a “non-significant” findings could reflect a marginal effect or a true absence of an effect.

## Effect sizes

Increasingly, the way researchers have sought to improve communication of results is to focus more on measures of the **effect size**. This changes our inference from “did the IV affect the DV?” into, “how much does the IV affect the DV?” In this approach, note that the null hypothesis is now the same as saying the effect size equals zero. Whenever we carry out an analysis, we are estimating the effect size based on our sample, which is a subgroup from a broader population. Unless we measure the entire population, we can never assert that the effect size is exactly one specific value. This is the difficulty of arguing for the null hypothesis. Our estimates can provide evidence that the effect size is not very different from zero, but not that it is exactly zero. When we fail to reject the null, we can only say that we are not sure that our current effect size estimate is different from zero.

As we reviewed earlier, an **unstandardized effect size** is simply the difference in the DV between conditions of interest across the average (mean) scores. In some cases, this can help communicate the results of an experiment, but it has the weakness of not incorporating any information about the variability of performance that was observed. **Standardized effect sizes** all incorporate a measure of variance to rescale the difference in means with the intention of providing a common scale for denoting effects across a scale something like ‘small,’ ‘medium,’ and ‘large.’ Unfortunately, the field of psychological science has not yet converged on a standard methodology analogous to the reporting of p-values. Instead, there are several different forms of standardized effect sizes that are used depending on methodology and analysis type. Here we will briefly review two of these.

One common standard effect size measure is **Cohen’s d**, which is often reported with t-tests to help communicate the findings. It is calculated as a ratio of the mean difference to variance and scales such that when **d** is greater than or equal to 0.8, the difference is described as a ‘large’ effect. For values greater than 1.5 or 2.0, sometimes more extreme adjectives are used. Values in the range of 0.5 are described as ‘medium’ size and near 0.2 are described as ‘small.’ Understanding the effect size when it is in the ‘small’ range is critical for planning experimental research. Small effect sizes may be real but require very large numbers of participants in order to have a good chance of observing the effect within a single study (see Power Analysis below). This is an important idea to understand in conditions where research “fails to replicate” prior work as a research study can fail to reject the null hypothesis due to a Type 2 error with small effect sizes. Unfortunately, there is no objective way to tell in these cases if the experiment was insufficiently sensitive to the real effect size or if the real effect size was actually zero (the null hypothesis is true).

Another common effect size measure is **eta-squared** or **η2** which is calculated as a standardized mean difference arising from ANOVA (sometimes called **partial eta-squared**). This can be treated as an effect magnitude estimate like Cohen’s d, but the scale is different. A ‘small’ effect on an **eta-squared** measure is 0.01, with 0.06 being ‘medium’ and 0.14 or larger being ‘large.’ Be careful with the different scales as a ‘large’ eta-squared effect looks numerically smaller than a small Cohen’s d.

## Power Analysis and Sensitivity in design

When planning a research study, particularly a rigorous Randomized Clinical Trial (RCT), it is important to be able to specify in advance exactly how many participants are expected to be in the research study. This is done by carrying out a p**ower analysis**, which is based on an a priori estimate of the effect size to be observed in the study. A power analysis takes a standardized effect size and with a specific number of participants expected to be recruited, provides a probability estimate of the chance of obtaining a reliable statistical difference between conditions. The math of carrying out this analysis is beyond our scope here, but the underlying idea is that even where there is a real, true difference between conditions, data can still be variable enough that our statistics do not work (we fail to reject the null, a Type 2 error). In many formal research proposals, studies are designed around a power analysis based estimate of 80% or 90% likelihood of success.

In many experimental research studies, the researchers do not start with a strongly held numerical estimate of the expected effect size. In this case, it is impossible to carry out a formal power analysis before starting research. However, if the data indicate no reliable statistical differences, it may lead the researchers to consider that their design lacks **sensitivity** to the observed effect size. That is, the effect size is smaller than could be detected with the sample size available. This is often the case in results termed “marginal” above. The best practice in this case is to estimate the effect size from the “failed” study and use this to design a better follow-up study with larger n and/or a more powerful manipulation.

A consideration of power and sensitivity points out the difficulty of interpreting findings that “fail to replicate” prior studies that have been commonly reported as inspiring a “replicability crisis.” We should actually expect studies to fail to replicate some of the time, even with real effect sizes when the effect is subtle, as many interesting effects are. Power analysis with effect sizes in the ‘small’ range can indicate that it may take several hundred participants to have a high probability of obtaining a reliable effect. There are certainly publications that have found reliable effects with smaller sample sizes, suggesting the researchers may have been lucky. We will consider the implications of this later in Chapter 19 (Responsible Conduct of Research).

## Bayesian analysis

An entirely alternate approach to statistical inference exists based on Bayesian analysis. This approach focuses on the probability that the experimental hypothesis is correct and how this is influenced and updated as data becomes available. The probability of truth of the hypothesis acts like a quantitative effect size measure and follows a very robust mathematical tradition. This approach has a major limitation in that the impact of any set of data can only be considered in the context of the beliefs of the researchers before the experiment was carried out. Since researchers often do not agree, it is very difficult to objectively quantify the effect of data on everybody’s beliefs. The Bayesian model is very intuitive as a model for how each individual scientist approaches carrying out research. The math is also a useful exercise to review, however, it is unlikely that we will have time to consider this alternate approach in any depth in this class.