# Comparing several means (one-way ANOVA) {#sec-Comparing-several-means-one-way-ANOVA}

```{r}

#| include: FALSE

source("header.R")

```

This chapter introduces one of the most widely used tools in psychological statistics, known as "the analysis of variance", but usually referred to as ANOVA. The basic technique was developed by Sir Ronald Fisher in the early 20th century and it is to him that we owe the rather unfortunate terminology. The term ANOVA is a little misleading, in two respects. Firstly, although the name of the technique refers to variances, ANOVA is concerned with investigating differences in means. Secondly, there are several different things out there that are all referred to as ANOVAs, some of which have only a very tenuous connection to one another. Later on in the book we'll encounter a range of different ANOVA methods that apply in quite different situations, but for the purposes of this chapter we'll only consider the simplest form of ANOVA, in which we have several different groups of observations, and we're interested in finding out whether those groups differ in terms of some outcome variable of interest. This is the question that is addressed by a one-way ANOVA.

The structure of this chapter is as follows: first I'll introduce a fictitious data set that we'll use as a running example throughout the chapter. After introducing the data, I'll describe the mechanics of how a one-way ANOVA actually works [How ANOVA works] and then focus on how you can run one in jamovi [Running an ANOVA in jamovi]. These two sections are the core of the chapter.

The remainder of the chapter discusses a range of important topics that inevitably arise when running an ANOVA, namely how to calculate effect sizes, post hoc tests and corrections for multiple comparisons and the assumptions that ANOVA relies upon. We'll also talk about how to check those assumptions and some of the things you can do if the assumptions are violated. Then we'll cover repeated measures ANOVA.

## An illustrative data set

Suppose you've become involved in a clinical trial in which you are testing a new antidepressant drug called \*Joyzepam\*. In order to construct a fair test of the drug's effectiveness, the study involves three separate drugs to be administered. One is a placebo, and the other is an existing antidepressant / anti-anxiety drug called \*Anxifree\*. A collection of 18 participants with moderate to severe depression are recruited for your initial testing. Because the drugs are sometimes administered in conjunction with psychological therapy, your study includes 9 people undergoing cognitive behavioural therapy (CBT) and 9 who are not. Participants are randomly assigned (doubly blinded, of course) a treatment, such that there are 3 CBT people and 3 no-therapy people assigned to each of the 3 drugs. A psychologist assesses the mood of each person after a 3 month run with each drug, and the overall improvement in each person's mood is assessed on a scale ranging from $-5$ to $+5$. With that as the study design, let's now load up the data file in clinicaltrial.csv. We can see that this data set contains the three variables drug, therapy and mood.gain.

For the purposes of this chapter, what we're really interested in is the effect of drug on mood.gain. The first thing to do is calculate some descriptive statistics and draw some graphs. In the @sec-Descriptive-statistics chapter we showed you how to do this, and some of the descriptive statistics we can calculate in jamovi are shown in @fig-fig13-1

```{r}

#| label: fig-fig13-1

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: Descriptives for mood gain, and box plots by drug administered

knitr::include\_graphics("images/fig13-1.png")

```

As the plot makes clear, there is a larger improvement in mood for participants in the Joyzepam group than for either the Anxifree group or the placebo group. The Anxifree group shows a larger mood gain than the control group, but the difference isn't as large. The question that we want to answer is are these difference "real", or are they just due to chance?

## How ANOVA works {#sec-How-ANOVA-works}

In order to answer the question posed by our clinical trial data we're going to run a one-way ANOVA. I'm going to start by showing you how to do it the hard way, building the statistical tool from the ground up and showing you how you could do it if you didn't have access to any of the cool built-in ANOVA functions in jamovi. And I hope you'll read it carefully, try to do it the long way once or twice to make sure you really understand how ANOVA works, and then once you've grasped the concept never ever do it this way again.

The experimental design that I described in the previous section strongly suggests that we're interested in comparing the average mood change for the three different drugs. In that sense, we're talking about an analysis similar to the t-test (see the chapter on [Comparing two means] but involving more than two groups. If we let $\mu\_P$ denote the population mean for the mood change induced by the placebo, and let $\mu\_A$ and $\mu\_J$ denote the corresponding means for our two drugs, Anxifree and Joyzepam, then the (somewhat pessimistic) null hypothesis that we want to test is that all three population means are identical. That is, neither of the two drugs is any more effective than a placebo. We can write out this null hypothesis as:

$$H\_0: \text{ it is true that } \mu\_P=\mu\_A=\mu\_J$$

As a consequence, our alternative hypothesis is that at least one of the three different treatments is different from the others. It's a bit tricky to write this mathematically, because (as we'll discuss) there are quite a few different ways in which the null hypothesis can be false. So for now we'll just write the alternative hypothesis like this:

$$H\_1: \text{ it } \underline{ is \text{ } not } \text{ true that }

\mu\_P=\mu\_A=\mu\_J$$

This null hypothesis is a lot trickier to test than any of the ones we've seen previously. How shall we do it? A sensible guess would be to "do an ANOVA", since that's the title of the chapter, but it's not particularly clear why an "analysis of variances" will help us learn anything useful about the means. In fact, this is one of the biggest conceptual difficulties that people have when first encountering ANOVA. To see how this works, I find it most helpful to start by talking about variances, specifically between group variability and within-group variability (@fig-fig13-2).

```{r}

#| label: fig-fig13-2

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: Graphical illustration of 'between groups' variation (panel (a)) and 'within groups' variation (panel (b)). On the left the arrows show the differences in the group means. On the right the arrows highlight the variability within each group

knitr::include\_graphics("images/fig13-2.png")

```

### Two formulas for the variance of Y

First, let's start by introducing some notation. We'll use G to refer to the total number of groups. For our data set there are three drugs, so there are $G = 3$ groups. Next, we'll use $N$ to refer to the total sample size; there are a total of $N = 18$ people in our data set. Similarly, let's use $N\_k$ to denote the number of people in the k-th group. In our fake clinical trial, the sample size is $N\_k = 6$ for all three groups.[^13-comparing-several-means-one-way-anova-1] Finally, we'll use Y to denote the outcome variable. In our case, Y refers to mood change. Specifically, we'll use Yik to refer to the mood change experienced by the i-th member of the k-th group. Similarly, we'll use $\bar{Y}$ to be the average mood change, taken across all 18 people in the experiment, and $\bar{Y}\_k$ to refer to the average mood change experienced by the 6 people in group $k$.

[^13-comparing-several-means-one-way-anova-1]: When all groups have the same number of observations, the experimental design is said to be "balanced". Balance isn't such a big deal for one-way ANOVA, which is the topic of this chapter. It becomes more important when you start doing more complicated ANOVAs.

Now that we've got our notation sorted out we can start writing down formulas. To start with, let's recall the formula for the variance that we used in the section on [Measures of variability], way back in those kinder days when we were just doing descriptive statistics. The sample variance of Y is defined as follows $$Var(Y)=\frac{1}{N}\sum\_{k=1}^{G}\sum\_{i=1}^{N\_k}(Y\_{ik}-\bar{Y})^2$$ This formula looks pretty much identical to the formula for the variance in [Measures of variability]. The only difference is that this time around I've got two summations here: I'm summing over groups (i.e., values for $k$) and over the people within the groups (i.e., values for $i$). This is purely a cosmetic detail. If I'd instead used the notation $Y\_p$ to refer to the value of the outcome variable for person p in the sample, then I'd only have a single summation. The only reason that we have a double summation here is that I've classified people into groups, and then assigned numbers to people within groups.

A concrete example might be useful here. Let's consider @tbl-tab13-1, in which we have a total of $N = 5$ people sorted into $G = 2$ groups. Arbitrarily, let's say that the "cool" people are group 1 and the "uncool" people are group 2. It turns out that we have three cool people ($N\_1 = 3$) and two uncool people ($N\_2 = 2$)

```{r}

#| label: tbl-tab13-1

#| tbl-cap: Grumpiness for people in cool and uncool groups

huxtabs[[13]][[1]]

```

Notice that I've constructed two different labelling schemes here. We have a "person" variable p so it would be perfectly sensible to refer to Yp as the grumpiness of the p-th person in the sample. For instance, the table shows that Tim is the fourth so we'd say $p = 4$. So, when talking about the grumpiness $Y$ of this "Tim" person, whoever he might be, we could refer to his grumpiness by saying that $Y\_p = 91$, for person $p = 4$ that is. However, that's not the only way we could refer to Tim. As an alternative we could note that Tim belongs to the "uncool" group ($k = 2$), and is in fact the first person listed in the uncool group ($i = 1$). So it's equally valid to refer to Tim's grumpiness by saying that $Y\_{ik} = 91$, where $k = 2$ and $i = 1$.

In other words, each person p corresponds to a unique ik combination, and so the formula that I gave above is actually identical to our original formula for the variance, which would be $$Var(Y)=\frac{1}{N}\sum\_{p=1}^{N}(Y\_p-\bar{Y})^2$$ In both formulas, all we're doing is summing over all of the observations in the sample. Most of the time we would just use the simpler Yp notation; the equation using $Y\_p$ is clearly the simpler of the two. However, when doing an ANOVA it's important to keep track of which participants belong in which groups, and we need to use the Yik notation to do this.

### From variances to sums of squares

Okay, now that we've got a good grasp on how the variance is calculated, let's define something called the \*\*total sum of squares\*\*, which is denoted SStot. This is very simple. Instead of averaging the squared deviations, which is what we do when calculating the variance, we just add them up.[^13-comparing-several-means-one-way-anova-2]

[^13-comparing-several-means-one-way-anova-2]: So the formula for the total sum of squares is almost identical to the formula for the variance $$SS\_{tot}=\sum\_{k=1}^{G} \sum\_{i=1}^{N\_k} (Y\_{ik} - \bar{Y})^2$$

When we talk about analysing variances in the context of ANOVA, what we're really doing is working with the total sums of squares rather than the actual variance. [^13-comparing-several-means-one-way-anova-3]

[^13-comparing-several-means-one-way-anova-3]: One very nice thing about the total sum of squares is that we can break it up into two different kinds of variation First, we can talk about the within-group sum of squares, in which we look to see how different each individual person is from their own group mean $$SS\_{w}= \sum\_{k=1}^{G} \sum\_{i=1}^{N\_k} (Y\_{ik} - \bar{Y}\_k)^2$$ where $\bar{Y}\_k$ is a group mean. In our example, $\bar{Y}\_k$ would be the average mood change experienced by those people given the k-th drug. So, instead of comparing individuals to the average of all people in the experiment, we're only comparing them to those people in the the same group. As a consequence, you'd expect the value of $SS\_w$ to be smaller than the total sum of squares, because it's completely ignoring any group differences, i.e., whether the drugs will have different effects on people's moods

Next, we can define a third notion of variation which captures only the differences between groups. We do this by looking at the differences between the group means $\bar{Y}\_k$ and grand mean $\bar{Y}$. [^13-comparing-several-means-one-way-anova-4]

[^13-comparing-several-means-one-way-anova-4]: In order to quantify the extent of this variation, what we do is calculate the between-group sum of squares $$ \begin{aligned} SS\_{b} &= \sum\_{k=1}^{G} \sum\_{i=1}^{N\_k} ( \bar{Y}\_{k} - \bar{Y} )^2 \\ &= \sum\_{k=1}^{G} N\_k ( \bar{Y}\_{k} - \bar{Y} )^2 \end{aligned} $$

It's not too difficult to show that the total variation among people in the experiment ($SS\_{tot}$ is actually the sum of the differences between the groups $SS\_b$ and the variation inside the groups $SS\_w$. That is,

$$SS\_w+SS\_b=SS\_{tot}$$ Yay.

Okay, so what have we found out? We've discovered that the total variability associated with the outcome variable ($SS\_{tot}$) can be mathematically carved up into the sum of "the variation due to the differences in the sample means for the different groups" ($SS\_b$) plus "all the rest of the variation" ($SS\_w$) [^13-comparing-several-means-one-way-anova-5].

How does that help me find out whether the groups have different population means? Um. Wait. Hold on a second. Now that I think about it, this is exactly what we were looking for. If the null hypothesis is true then you'd expect all the sample means to be pretty similar to each other, right? And that would imply that you'd expect $SS\_b$ to be really small, or at least you'd expect it to be a lot smaller than "the variation associated with everything else", $SS\_w$. Hmm. I detect a hypothesis test coming on.

[^13-comparing-several-means-one-way-anova-5]: SS\_w is also referred to in an independent ANOVA as the error variance, or $SS\_{error}$

### From sums of squares to the F-test

As we saw in the last section, the qualitative idea behind ANOVA is to compare the two sums of squares values $SS\_b$ and $SS\_w$ to each other. If the between-group variation $SS\_b$ is large relative to the within-group variation $SS\_w$ then we have reason to suspect that the population means for the different groups aren't identical to each other. In order to convert this into a workable hypothesis test, there's a little bit of "fiddling around" needed. What I'll do is first show you what we do to calculate our test statistic, the \*\*F ratio\*\*, and then try to give you a feel for why we do it this way.

In order to convert our SS values into an F-ratio the first thing we need to calculate is the \*\*degrees of freedom\*\* associated with the $SS\_b$ and $SS\_w$ values. As usual, the degrees of freedom corresponds to the number of unique "data points" that contribute to a particular calculation, minus the number of "constraints" that they need to satisfy. For the within-groups variability what we're calculating is the variation of the individual observations ($N$ data points) around the group means ($G$ constraints). In contrast, for the between groups variability we're interested in the variation of the group means (G data points) around the grand mean (1 constraint). Therefore, the degrees of freedom here are:

$$df\_b=G-1$$ $$df\_w=N-G$$

Okay, that seems simple enough. What we do next is convert our summed squares value into a "mean squares" value, which we do by dividing by the degrees of freedom:

$$MS\_b=\frac{SS\_b}{df\_b}$$ $$MS\_w=\frac{SS\_w}{df\_w}$$

Finally, we calculate the F-ratio by dividing the between-groups MS by the within-groups MS:

$$F=\frac{MS\_b}{MS\_w}$$

At a very general level, the intuition behind the F statistic is straightforward. Bigger values of F means that the between-groups variation is large relative to the within-groups variation. As a consequence, the larger the value of F the more evidence we have against the null hypothesis. But how large does $F$ have to be in order to actually reject $H\_0$? In order to understand this, you need a slightly deeper understanding of what ANOVA is and what the mean squares values actually are.

The next section discusses that in a bit of detail, but for readers that aren't interested in the details of what the test is actually measuring I'll cut to the chase. In order to complete our hypothesis test we need to know the sampling distribution for F if the null hypothesis is true. Not surprisingly, the sampling distribution for the F <u>statistic</u> under the null hypothesis is an $F$ distribution. If you recall our discussion of the F distribution in @sec-Introduction-to-probability, the $F$ <u>distribution</u> has two parameters, corresponding to the two degrees of freedom involved. The first one $df\_1$ is the between groups degrees of freedom $df\_b$, and the second one $df\_2$ is the within groups degrees of freedom $df\_w$.

```{r}

#| label: tbl-tab13-2

#| tbl-cap: All of the key quantities involved in an ANOVA organised into a 'standard' ANOVA table. The formulas for all quantities (except the p-value which has a very ugly formula and would be nightmarishly hard to calculate without a computer) are shown

huxtabs[[13]][[2]]

```

A summary of all the key quantities involved in a one-way ANOVA, including the formulas showing how they are calculated, is shown in @tbl-tab13-2.

[Additional technical detail [^13-comparing-several-means-one-way-anova-6]]

[^13-comparing-several-means-one-way-anova-6]: At a fundamental level ANOVA is a competition between two different statistical models, $H\_0$ and $H\_1$. When I described the null and alternative hypotheses at the start of the section, I was a little imprecise about what these models actually are. I'll remedy that now, though you probably won't like me for doing so. If you recall, our null hypothesis was that all of the group means are identical to one another. If so, then a natural way to think about the outcome variable $Y\_{ik}$ is to describe individual scores in terms of a single population mean µ, plus the deviation from that population mean. This deviation is usually denoted $\epsilon\_{ik}$ and is traditionally called the error or residual associated with that observation. Be careful though. Just like we saw with the word "significant", the word "error" has a technical meaning in statistics that isn't quite the same as its everyday English definition. In everyday language, "error" implies a mistake of some kind, but in statistics it doesn't (or at least, not necessarily). With that in mind, the word "residual" is a better term than the word "error". In statistics both words mean "leftover variability", that is "stuff" that the model can't explain.

In any case, here's what the null hypothesis looks like when we write it as a statistical model $$Y\_{ik}=\mu+\epsilon\_{ik}$$ where we make the assumption (discussed later) that the residual values $\epsilon\_{ik}$ are normally distributed, with mean $0$ and a standard deviation $\sigma$ that is the same for all groups. To use the notation that we introduced in the [Introduction to probability] we would write this assumption like this $$\epsilon\_{ik} \sim Normal(0,\sigma^2)$$ What about the alternative hypothesis, $H\_1$? The only difference between the null hypothesis and the alternative hypothesis is that we allow each group to have a different population mean. So, if we let $\mu\_k$ denote the population mean for the k-th group in our experiment, then the statistical model corresponding to $H\_1$ is $$Y\_{ik}=\mu\_k+\epsilon\_{ik}$$ where, once again, we assume that the error terms are normally distributed with mean 0 and standard deviation $\sigma$. That is, the alternative hypothesis also assumes that $\epsilon \sim Normal(0,\sigma^2)$ Okay, now that we've described the statistical models underpinning $H\_0$ and $H\_1$ in more detail, it's now pretty straightforward to say what the mean square values are measuring, and what this means for the interpretation of $F$. I won't bore you with the proof of this but it turns out that the within-groups mean square, $MS\_w$, can be viewed as an estimator of the error variance $\sigma^2$ . The between-groups mean square $MS\_b$ is also an estimator, but what it estimates is the error variance plus a quantity that depends on the true differences among the group means. If we call this quantity $Q$, then we can see that the F-statistic is basically$^a$ $$F=\frac{\hat{Q}+\hat{\sigma}^2}{\hat{\sigma}^2}$$ where the true value $Q = 0$ if the null hypothesis is true, and Q < 0 if the alternative hypothesis is true (e.g., @Hays1994, ch. 10). Therefore, at a bare minimum the $F$ \*value must be larger than\* 1 to have any chance of rejecting the null hypothesis. Note that this doesn't mean that it's impossible to get an F-value less than 1. What it means is that if the null hypothesis is true the sampling distribution of the F ratio has a mean of 1,[\^b] and so we need to see F-values larger than 1 in order to safely reject the null. To be a bit more precise about the sampling distribution, notice that if the null hypothesis is true, both MSb and MSw are estimators of the variance of the residuals $\epsilon\_{ik}$. If those residuals are normally distributed, then you might suspect that the estimate of the variance of $\epsilon\_{ik}$ is chi-square distributed, because (as discussed in the @sec-Other-useful-distributions) that's what a chi-square distribution is: it's what you get when you square a bunch of normally-distributed things and add them up. And since the F distribution is (again, by definition) what you get when you take the ratio between two things that are $\chi^2$ distributed, we have our sampling distribution. Obviously, I'm glossing over a whole lot of stuff when I say this, but in broad terms, this really is where our sampling distribution comes from. <br /> --- <br /> $^a$ If you read ahead to @sec-Factorial-ANOVA and look at how the "treatment effect" at level k of a factor is defined in terms of the $\alpha\_k$ values (see section on Factorial ANOVA 2: balanced designs, interactions allowed]), it turns out that $Q$ refers to a weighted mean of the squared treatment effects, $Q = \frac{(\sum\_{k=1}^{G}N\_k \alpha\_k^2)}{(G-1)}$ <br /> $^b$ Or, if we want to be sticklers for accuracy, $1+ \frac{2}{df\_2-2}$

### A worked example

The previous discussion was fairly abstract and a little on the technical side, so I think that at this point it might be useful to see a worked example. For that, let's go back to the clinical trial data that I introduced at the start of the chapter. The descriptive statistics that we calculated at the beginning tell us our group means: an average mood gain of $0.45$ for the placebo, $0.72$ for Anxifree, and $1.48$ for Joyzepam. With that in mind, let's party like it's 1899 [^13-comparing-several-means-one-way-anova-7] and start doing some pencil and paper calculations. I'll only do this for the first $5$ observations because it's not bloody $1899$ and I'm very lazy. Let's start by calculating $SS\_w$, the within-group sums of squares. First, let's draw up a nice table to help us with our calculations (@tbl-tab13-3)

[^13-comparing-several-means-one-way-anova-7]: Or, to be precise, party like "it's 1899 and we've got no friends and nothing better to do with our time than do some calculations that wouldn't have made any sense in 1899 because ANOVA didn't exist until about the 1920s".

```{r}

#| label: tbl-tab13-3

#| tbl-cap: A worked example...1

huxtabs[[13]][[3]]

```

At this stage, the only thing I've included in the table is the raw data itself. That is, the grouping variable (i.e., drug) and outcome variable (i.e. mood.gain) for each person. Note that the outcome variable here corresponds to the $\bar{Y}\_{ik}$ value in our equation previously. The next step in the calculation is to write down, for each person in the study, the corresponding group mean, $\bar{Y}\_k$. This is slightly repetitive but not particularly difficult since we already calculated those group means when doing our descriptive statistics, see @tbl-tab13-4.

```{r}

#| label: tbl-tab13-4

#| tbl-cap: A worked example...2

huxtabs[[13]][[4]]

```

Now that we've written those down, we need to calculate, again for every person, the deviation from the corresponding group mean. That is, we want to subtract $Y\_{ik} - \bar{Y}\_k$. After we've done that, we need to square everything. When we do that, here's what we get (@tbl-tab13-5)

```{r}

#| label: tbl-tab13-5

#| tbl-cap: A worked example...3

huxtabs[[13]][[5]]

```

The last step is equally straightforward. In order to calculate the within-group sum of squares we just add up the squared deviations across all observations:

$$

\begin{split}

SS\_w & = 0.0025 + 0.0225 + 0.1225 + 0.0136 + 0.1003 \\

& = 0.2614

\end{split}

$$

Of course, if we actually wanted to get the right answer we'd need to do this for all 18 observations in the data set, not just the first five. We could continue with the pencil and paper calculations if we wanted to, but it's pretty tedious. Alternatively, it's not too hard to do this in a dedicated spreadsheet programme such as OpenOffice or Excel. Try and do it yourself. The one that I did, in Excel, is in the file clinicaltrial\_anova.xls. When you do it you should end up with a within-group sum of squares value of $1.39$.

Okay. Now that we've calculated the within groups variation, $SS\_w$, it's time to turn our attention to the between-group sum of squares, $SS\_b$. The calculations for this case are very similar. The main difference is that instead of calculating the differences between an observation Yik and a group mean $\bar{Y}\_k$ for all of the observations, we calculate the differences between the group means $\bar{Y}\_k$ and the grand mean $\bar{Y}$ (in this case $0.88$) for all of the groups (@tbl-tab13-6).

```{r}

#| label: tbl-tab13-6

#| tbl-cap: A worked example...4

huxtabs[[13]][[6]]

```

However, for the between group calculations we need to multiply each of these squared deviations by $N\_k$, the number of observations in the group. We do this because every observation in the group (all $N\_k$ of them) is associated with a between group difference. So if there are six people in the placebo group and the placebo group mean differs from the grand mean by $0.19$, then the total between group variation associated with these six people is $6 \times 0.19 = 1.14$. So we have to extend our little table of calculations (@tbl-tab13-7).

```{r}

#| label: tbl-tab13-7

#| tbl-cap: A worked example...5

huxtabs[[13]][[7]]

```

And so now our between group sum of squares is obtained by summing these "weighted squared deviations" over all three groups in the study:

$$\begin{aligned} SS\_b & = 1.14 + 0.18 + 2.16 \\ &= 3.48 \end{aligned}$$

As you can see, the between group calculations are a lot shorter[^13-comparing-several-means-one-way-anova-8]. Now that we've calculated our sums of squares values, $SS\_b$ and $SS\_w$, the rest of the ANOVA is pretty painless. The next step is to calculate the degrees of freedom. Since we have $G = 3$ groups and $N = 18$ observations in total our degrees of freedom can be calculated by simple subtraction:

[^13-comparing-several-means-one-way-anova-8]: In the Excel clinicaltrial anova.xls the value for SSb worked out to be very slightly different, $3.45$, than that shown in the text above (rounding errors!)

$$

\begin{split}

df\_b & = G-1 = 2 \\

df\_w & = N-G = 15

\end{split}

$$

Next, since we've now calculated the values for the sums of squares and the degrees of freedom, for both the within-groups variability and the between-groups variability, we can obtain the mean square values by dividing one by the other:

$$

\begin{split}

MS\_b & = \frac{SS\_b}{df\_b} = \frac{3.48}{2} = 1.74 \\

MS\_w & = \frac{SS\_w}{df\_w} = \frac{1.39}{15} = 0.09

\end{split}

$$

We're almost done. The mean square values can be used to calculate the F-value, which is the test statistic that we're interested in. We do this by dividing the between-groups MS value by the within-groups MS value.

$$

\begin{split}

F & = \frac{MS\_b}{MS\_w} = \frac{1.74}{0.09} \\

& = 19.3

\end{split}

$$

Woohooo! This is terribly exciting, yes? Now that we have our test statistic, the last step is to find out whether the test itself gives us a significant result. As discussed in [Hypothesis testing] back in the "old days" what we'd do is open up a statistics textbook or flick to the back section which would actually have a huge lookup table and we would find the threshold $F$ value corresponding to a particular value of alpha (the null hypothesis rejection region), e.g. $0.05$, $0.01$ or $0.001$, for 2 and 15 degrees of freedom. Doing it this way would give us a threshold F value for an alpha of $0.001$ of $11.34$. As this is less than our calculated $F$ value we say that $p < 0.001$. But those were the old days, and nowadays fancy stats software calculates the exact p-value for you. In fact, the exact p-value is $0.000071$. So, unless we're being \*extremely\* conservative about our Type I error rate, we're pretty much guaranteed to reject the null hypothesis.

At this point, we're basically done. Having completed our calculations, it's traditional to organise all these numbers into an ANOVA table like the one in Table 13.1. For our clinical trial data, the ANOVA table would look like @tbl-tab13-8.

```{r}

#| label: tbl-tab13-8

#| tbl-cap: The ANOVA results table

huxtabs[[13]][[8]]

```

These days, you'll probably never have much reason to want to construct one of these tables yourself, but you will find that almost all statistical software (jamovi included) tends to organise the output of an ANOVA into a table like this, so it's a good idea to get used to reading them. However, although the software will output a full ANOVA table, there's almost never a good reason to include the whole table in your write up. A pretty standard way of reporting the stats block for this result would be to write something like this:

> \*One-way ANOVA showed a significant effect of drug on mood gain (F(2,15) = 19.3, p < .001).\*

Sigh. So much work for one short sentence.

## Running an ANOVA in jamovi

I'm pretty sure I know what you're thinking after reading the last section, especially if you followed my advice and did all of that by pencil and paper (i.e., in a spreadsheet) yourself. Doing the ANOVA calculations yourself sucks. There's quite a lot of calculations that we needed to do along the way, and it would be tedious to have to do this over and over again every time you wanted to do an ANOVA.

### Using jamovi to specify your ANOVA

To make life easier for you, jamovi can do ANOVA...hurrah! Go to the 'ANOVA' - 'ANOVA' analysis, and move the mood.gain variable across so it is in the 'Dependent Variable' box, and then move the drug variable across so it is in the 'Fixed Factors' box. This should give the results as shown in @fig-fig13-3. [^13-comparing-several-means-one-way-anova-9] Note I have also checked the $\eta^2$ checkbox, pronounced "eta" squared, under the 'Effect Size' option and this is also shown on the results table. We will come back to effect sizes a bit later.

```{r}

#| label: fig-fig13-3

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: jamovi results table for ANOVA of mood gain by drug administered

knitr::include\_graphics("images/fig13-3.png")

```

[^13-comparing-several-means-one-way-anova-9]: The jamovi results are more accurate than the ones in the text above, due to rounding errors.

The jamovi results table shows you the sums of squares values, the degrees of freedom, and a couple of other quantities that we're not really interested in right now. Notice, however, that jamovi doesn't use the names "between-group" and "within-group". Instead, it tries to assign more meaningful names. In our particular example, the between groups variance corresponds to the effect that the drug has on the outcome variable, and the within groups variance corresponds to the "leftover" variability so it calls that the residuals. If we compare these numbers to the numbers that I calculated by hand in [A worked example], you can see that they're more or less the same, apart from rounding errors. The between groups sums of squares is $SS\_b = 3.45$, the within groups sums of squares is $SS\_w = 1.39$, and the degrees of freedom are $2$ and $15$ respectively. We also get the F-value and the p-value and, again, these are more or less the same, give or take rounding errors, to the numbers that we calculated ourselves when doing it the long and tedious way.

## Effect size

There's a few different ways you could measure the effect size in an ANOVA, but the most commonly used measures are $\eta^2$ (eta squared) and partial $\eta^2$. For a one way analysis of variance they're identical to each other, so for the moment I'll just explain $\eta^2$ . The definition of $\eta^2$ is actually really simple

$$\eta^2=\frac{SS\_b}{SS\_{tot}}$$

That's all it is. So when I look at the ANOVA table in @fig-fig13-3, I see that $SS\_b = 3.45$ and $SS\_tot = 3.45 + 1.39 = 4.84$. Thus we get an $\eta^2$ value of

$$\eta^2=\frac{3.45}{4.84}=0.71$$

The interpretation of $\eta^2$ is equally straightforward. It refers to the proportion of the variability in the outcome variable (mood.gain) that can be explained in terms of the predictor (drug). A value of $\eta^2=0$ means that there is no relationship at all between the two, whereas a value of $\eta^2=1$ means that the relationship is perfect. Better yet, the $\eta^2$ value is very closely related to $R^2$, as discussed previously in the [The $R^2$ value] section, and has an equivalent interpretation. Although many statistics text books suggest $\eta^2$ as the default effect size measure in ANOVA, there's an interesting <a href="https://daniellakens.blogspot.com/2015/06/why-you-should-use-omega-squared.html" target="\_blank">blog post</a> by Daniel Lakens suggesting that eta-squared is perhaps not the best measure of effect size in real world data analysis, because it can be a biased estimator. Usefully, there is also an option in jamovi to specify omega-squared ($\omega^2$), which is less biased, alongside eta-squared.

## Multiple comparisons and post hoc tests

Any time you run an ANOVA with more than two groups and you end up with a significant effect, the first thing you'll probably want to ask is which groups are actually different from one another. In our drugs example, our null hypothesis was that all three drugs (placebo, Anxifree and Joyzepam) have the exact same effect on mood. But if you think about it, the null hypothesis is actually claiming three different things all at once here. Specifically, it claims that:

- Your competitor's drug (Anxifree) is no better than a placebo (i.e., $\mu\_A = \mu\_P$ )

- Your drug (Joyzepam) is no better than a placebo (i.e., $\mu\_J = \mu\_P$ )

- Anxifree and Joyzepam are equally effective (i.e., $\mu\_J = \mu\_A$)

If any one of those three claims is false, then the null hypothesis is also false. So, now that we've rejected our null hypothesis, we're thinking that at least one of those things isn't true. But which ones? All three of these propositions are of interest. Since you certainly want to know if your new drug Joyzepam is better than a placebo, it would be nice to know how well it stacks up against an existing commercial alternative (i.e., Anxifree). It would even be useful to check the performance of Anxifree against the placebo. Even if Anxifree has already been extensively tested against placebos by other researchers, it can still be very useful to check that your study is producing similar results to earlier work.

When we characterise the null hypothesis in terms of these three distinct propositions, it becomes clear that there are eight possible "states of the world" that we need to distinguish between (@tbl-tab13-9).

```{r}

#| label: tbl-tab13-9

#| tbl-cap: The null hypothesis and eight possible 'states of the world'

huxtabs[[13]][[9]]

```

By rejecting the null hypothesis, we've decided that we don't believe that #1 is the true state of the world. The next question to ask is, which of the other seven possibilities \*do\* we think is right? When faced with this situation, its usually helps to look at the data. For instance, if we look at the plots in @fig-fig13-1, it's tempting to conclude that Joyzepam is better than the placebo and better than Anxifree, but there's no real difference between Anxifree and the placebo. However, if we want to get a clearer answer about this, it might help to run some tests.

### Running "pairwise" t-tests

How might we go about solving our problem? Given that we've got three separate pairs of means (placebo versus Anxifree, placebo versus Joyzepam, and Anxifree versus Joyzepam) to compare, what we could do is run three separate t-tests and see what happens. This is easy to do in jamovi. Go to the ANOVA 'Post Hoc Tests' options, move the 'drug' variable across into the active box on the right, and then click on the 'No correction' checkbox. This will produce a neat table showing all the pairwise t-test comparisons amongst the three levels of the drug variable, as in @fig-fig13-4

```{r}

#| label: fig-fig13-4

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: Uncorrected pairwise t-tests as post hoc comparisons in jamovi

knitr::include\_graphics("images/fig13-4.png")

```

### Corrections for multiple testing

In the previous section I hinted that there's a problem with just running lots and lots of t-tests. The concern is that, when running these analyses, what we're doing is going on a "fishing expedition". We're running lots and lots of tests without much theoretical guidance in the hope that some of them come up significant. This kind of theory-free search for group differences is referred to as \*\*post hoc analysis\*\* ("post hoc" being Latin for "after this").[^13-comparing-several-means-one-way-anova-10]

[^13-comparing-several-means-one-way-anova-10]: If you \*do\* have some theoretical basis for wanting to investigate some comparisons but not others, it's a different story. In those circumstances you're not really running "post hoc" analyses at all, you're making "planned comparisons". I do talk about this situation later in the book - @sec-The-method-of-planned-comparisons, but for now I want to keep things simple.

It's okay to run post hoc analyses, but a lot of care is required. For instance, the analysis that I ran in the previous section should be avoided, as each individual t-test is designed to have a 5% Type I error rate (i.e., $\alpha = .05$) and I ran three of these tests. Imagine what would have happened if my ANOVA involved 10 different groups, and I had decided to run 45 "post hoc" t-tests to try to find out which ones were significantly different from each other, you'd expect 2 or 3 of them to come up significant by chance alone. As we saw in @sec-Hypothesis-testing, the central organising principle behind null hypothesis testing is that we seek to control our Type I error rate, but now that I'm running lots of t-tests at once in order to determine the source of my ANOVA results, my actual Type I error rate across this whole family of tests has gotten completely out of control.

The usual solution to this problem is to introduce an adjustment to the p-value, which aims to control the total error rate across the family of tests (see @Shaffer1995). An adjustment of this form, which is usually (but not always) applied because one is doing post hoc analysis, is often referred to as a \*\*correction for multiple comparisons\*\*, though it is sometimes referred to as "simultaneous inference". In any case, there are quite a few different ways of doing this adjustment. I'll discuss a few of them in this section and in @sec-Post-hoc-tests in the next chapter, but you should be aware that there are many other methods out there (see, e.g., @Hsu1996).

### Bonferroni corrections

The simplest of these adjustments is called the \*\*Bonferroni correction\*\* [@Dunn1961], and it's very very simple indeed. Suppose that my post hoc analysis consists of m separate tests, and I want to ensure that the total probability of making \*any\* Type I errors at all is at most $\alpha$.[^13-comparing-several-means-one-way-anova-11] If so, then the Bonferroni correction just says "multiply all your raw p-values by m". If we let $p$ denote the original p-value, and let $p\_j^{'}$ be the corrected value, then the Bonferroni correction tells that:

[^13-comparing-several-means-one-way-anova-11]: It's worth noting in passing that not all adjustment methods try to do this. What I've described here is an approach for controlling "family wise Type I error rate". However, there are other post hoc tests that seek to control the "false discovery rate", which is a somewhat different thing.

$$p\_j^{'}=m \times p$$

And therefore, if you're using the Bonferroni correction, you would reject the null hypothesis if $p\_j^{'} < \alpha$. The logic behind this correction is very straightforward. We're doing m different tests, so if we arrange it so that each test has a Type I error rate of at most $\frac{\alpha}{m}$, then the \*total\* Type I error rate across these tests cannot be larger than $\alpha$. That's pretty simple, so much so that in the original paper, the author writes:

>The method given here is so simple and so general that I am sure it must have been used before this. I do not find it, however, so can only conclude that perhaps its very simplicity has kept statisticians from realizing that it is a very good method in some situations (@Dunn1961, pp 52-53).

To use the Bonferroni correction in jamovi, just click on the 'Bonferroni' checkbox in the 'Correction' options, and you will see another column added to the ANOVA results table showing the adjusted p-values for the Bonferroni correction (@tbl-tab13-8). If we compare these three p-values to those for the uncorrected, pairwise t-tests, it is clear that the only thing that jamovi has done is multiply them by $3$.

### Holm corrections

Although the Bonferroni correction is the simplest adjustment out there, it's not usually the best one to use. One method that is often used instead is the \*\*Holm correction\*\* [@Holm1979]. The idea behind the Holm correction is to pretend that you're doing the tests sequentially, starting with the smallest (raw) p-value and moving onto the largest one. For the j-th largest of the p-values, the adjustment is \*either\*

$$p\_j^{'}=j \times p\_j$$

(i.e., the biggest p-value remains unchanged, the second biggest p-value is doubled, the third biggest p-value is tripled, and so on), or

$$p\_j^{'}=p\_{j+1}^{'}$$

whichever one is <u>larger</u>. This might sound a little confusing, so let's go through it a little more slowly. Here's what the Holm correction does. First, you sort all of your p-values in order, from smallest to largest. For the smallest p-value all you do is multiply it by $m$, and you're done. However, for all the other ones it's a two-stage process. For instance, when you move to the second smallest p value, you first multiply it by $m - 1$. If this produces a number that is bigger than the adjusted p-value that you got last time, then you keep it. But if it's smaller than the last one, then you copy the last p-value. To illustrate how this works, consider @tbl-tab13-10 which shows the calculations of a Holm correction for a collection of five p-values.

```{r}

#| label: tbl-tab13-10

#| tbl-cap: Holm corrected p values

huxtabs[[13]][[10]]

```

Hopefully that makes things clear.

Although it's a little harder to calculate, the Holm correction has some very nice properties. It's more powerful than Bonferroni (i.e., it has a lower Type II error rate) but, counter-intuitive as it might seem, it has the same Type I error rate. As a consequence, in practice there's never any reason to use the simpler Bonferroni correction since it is always outperformed by the slightly more elaborate Holm correction. Because of this, the Holm correction should be your \*go to\* multiple comparison correction. @fig-fig13-4 also shows the Holm corrected p-values and, as you can see, the biggest p-value (corresponding to the comparison between Anxifree and the placebo) is unaltered. At a value of .15, it is exactly the same as the value we got originally when we applied no correction at all. In contrast, the smallest p-value (Joyzepam versus placebo) has been multiplied by three.

### Writing up the post hoc test

Finally, having run the post hoc analysis to determine which groups are significantly different to one another, you might write up the result like this:

> \*Post hoc tests (using the Holm correction to adjust p) indicated that Joyzepam produced a significantly larger mood change than both Anxifree (p = .001) and the placebo ($(p = 9.0 \times{10^{-5}}$). We found no evidence that Anxifree performed better than the placebo ($p = .15$).\*

Or, if you don't like the idea of reporting exact p-values, then you'd change those numbers to $p < .01$, $p < .001$ and $p > .05$ respectively. Either way, the key thing is that you indicate that you used Holm's correction to adjust the p-values. And of course, I'm assuming that elsewhere in the write up you've included the relevant descriptive statistics (i.e., the group means and standard deviations), since these p-values on their own aren't terribly informative.

## The assumptions of one-way ANOVA

Like any statistical test, analysis of variance relies on some assumptions about the data, specifically the residuals. There are three key assumptions that you need to be aware of: normality, homogeneity of variance and independence.

[Additional technical detail [^13-comparing-several-means-one-way-anova-12]]

[^13-comparing-several-means-one-way-anova-12]: If you remember back to [A worked example], which I hope you at least skimmed even if you didn't read the whole thing, I described the statistical models underpinning ANOVA in this way: $$H\_0:Y\_{ik}=\mu + \epsilon\_{ik}$$ $$H\_1:Y\_{ik}=\mu\_k + \epsilon\_{ik}$$ In these equations $\mu$ refers to a single grand population mean which is the same for all groups, and µk is the population mean for the k-th group. Up to this point we've been mostly interested in whether our data are best described in terms of a single grand mean (the null hypothesis) or in terms of different group-specific means (the alternative hypothesis). This makes sense, of course, as that's actually the important research question! However, all of our testing procedures have, implicitly, relied on a specific assumption about the residuals, $\epsilon\\_{ik}$, namely that $$\epsilon\_{ik} \sim Normal(0,\sigma^2)$$ None of the maths works properly without this bit. Or, to be precise, you can still do all the calculations and you'll end up with an F-statistic, but you have no guarantee that this F-statistic actually measures what you think it's measuring, and so any conclusions that you might draw on the basis of the F test might be wrong.

So, how do we check whether the assumption about the residuals is accurate? Well, as I indicated above, there are three distinct claims buried in this one statement, and we'll consider them separately.

- \*\*Homogeneity of variance\*\*. Notice that we've only got the one value for the population standard deviation (i.e., $\sigma$), rather than allowing each group to have it's own value (i.e., $\sigma\_k$). This is referred to as the homogeneity of variance (sometimes called homoscedasticity) assumption. ANOVA assumes that the population standard deviation is the same for all groups. We'll talk about this extensively in the [Checking the homogeneity of variance assumption] section.

- \*\*Normality\*\*. The residuals are assumed to be normally distributed. As we saw in @sec-Checking-the-normality-of-a-sample, we can assess this by looking at QQ plots (or running a Shapiro-Wilk test. I'll talk about this more in an ANOVA context in the [Checking the normality assumption] section.

- \*\*Independence\*\*. The independence assumption is a little trickier. What it basically means is that, knowing one residual tells you nothing about any other residual. All of the $\epsilon\_{ik}$ values are assumed to have been generated without any "regard for" or "relationship to" any of the other ones. There's not an obvious or simple way to test for this, but there are some situations that are clear violations of this. For instance, if you have a repeated measures design, where each participant in your study appears in more than one condition, then independence doesn't hold. There's a special relationship between some observations, namely those that correspond to the same person! When that happens, you need to use something like a [Repeated measures one-way ANOVA].

### Checking the homogeneity of variance assumption {#sec-Checking-the-homogeneity-of-variance-assumption}

> \*To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!\* \

> -- George Box [@Box1953]

There's more than one way to skin a cat, as the saying goes, and more than one way to test the homogeneity of variance assumption, too (though for some reason no-one made a saying out of that). The most commonly used test for this that I've seen in the literature is the Levene test [@Levene1960], and the closely related Brown-Forsythe test [@BrownForsythe1974].

Regardless of whether you're doing the standard Levene test or the Brown-Forsythe test, the test statistic, which is sometimes denoted $F$ but also sometimes written as $W$, is calculated in exactly the same way that the F-statistic for the regular ANOVA is calculated, just using a $Z\_{ik}$ rather than $Y\_{ik}$. With that in mind, we can go on to look at how to run the test in jamovi.

[Additional technical detail [^13-comparing-several-means-one-way-anova-13]]

[^13-comparing-several-means-one-way-anova-13]: The Levene test is shockingly simple. Suppose we have our outcome variable $Y\_{ik}$. All we do is define a new variable, which I'll call $Z\_{ik}$, corresponding to the absolute deviation from the group mean $$Z\_{ik}=Y\_{ik}-\bar{Y}\_{k}$$ Okay, what good does this do us? Well, let's take a moment to think about what $Z\_{ik}$ actually is and what we're trying to test. The value of $Z\_{ik}$ is a measure of how the $i$-th observation in the $k$-th group deviates from its group mean. And our null hypothesis is that all groups have the same variance, i.e., the same overall deviations from the group means! So the null hypothesis in a Levene test is that the population means of $Z$ are identical for all groups. Hmm. So what we need now is a statistical test of the null hypothesis that all group means are identical. Where have we seen that before? Oh right, that's what ANOVA is, and so all that the Levene test does is run an ANOVA on the new variable $Z\_{ik}$. What about the Brown-Forsythe test? Does that do anything particularly different? Nope. The only change from the Levene test is that it constructs the transformed variable Z in a slightly different way, using deviations from the group medians rather than deviations from the group means. That is, for the Brown-Forsythe test: $$Z\_{ik}=Y\_{ik}-median\_k(Y)$$ where $median\_k(Y)$ is the median for group k.

### Running the Levene test in jamovi

Okay, so how do we run the Levene test? Simple really - under the ANOVA 'Assumption Checks' option, just click on the 'Homogeneity tests' checkbox. If we look at the output, shown in @fig-fig13-5, we see that the test is non-significant ($F\_{2,15} = 1.45, p = .266$), so it looks like the homogeneity of variance assumption is fine. However, looks can be deceptive! If your sample size is pretty big, then the Levene test could show up a significant effect (i.e. p < .05) even when the homogeneity of variance assumption is not violated to an extent which troubles the robustness of ANOVA. This was the point George Box was making in the quote above. Similarly, if your sample size is quite small, then the homogeneity of variance assumption might not be satisfied and yet a Levene test could be non-significant (i.e. p > .05). What this means is that, alongside any statistical test of the assumption being met, you should always plot the standard deviation around the means for each group / category in the analysis...just to see if they look fairly similar (i.e. homogeneity of variance) or not.

```{r}

#| label: fig-fig13-5

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: Levene test output for one-way ANOVA in jamovi

knitr::include\_graphics("images/fig13-5.png")

```

### Removing the homogeneity of variance assumption

In our example, the homogeneity of variance assumption turned out to be a pretty safe one: the Levene test came back non-significant (notwithstanding that we should also look at the plot of standard deviations), so we probably don't need to worry. However, in real life we aren't always that lucky. How do we save our ANOVA when the homogeneity of variance assumption is violated? If you recall from our discussion of t-tests, we've seen this problem before. The Student t-test assumes equal variances, so the solution was to use the Welch t-test, which does not. In fact, @Welch1951 also showed how we can solve this problem for ANOVA too (the \*\*Welch one-way test\*\*). It's implemented in jamovi using the One-Way ANOVA analysis. This is a specific analysis approach just for one-way ANOVA, and to run the Welch one-way ANOVA for our example, we would re-run the analysis as previously, but this time use the jamovi ANOVA - One Way ANOVA analysis command, and check the option for Welch's test (see @fig-fig13-6). To understand what's happening here, let's compare these numbers to what we got earlier when [Running an ANOVA in jamovi] originally. To save you the trouble of flicking back, this is what we got last time: $F(2, 15) = 18.611, p = .00009$, also shown as the Fisher's test in the One-Way ANOVA shown in @fig-fig13-6.

```{r}

#| label: fig-fig13-6

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: Welch's test as part of the One Way ANOVA analysis in jamovi

knitr::include\_graphics("images/fig13-6.png")

```

Okay, so originally our ANOVA gave us the result $F(2, 15) = 18.6$, whereas the Welch one way test gave us $F(2, 9.49) = 26.32$. In other words, the Welch test has reduced the within-groups degrees of freedom from 15 to 9.49, and the F-value has increased from 18.6 to 26.32.

### Checking the normality assumption {#sec-Checking-the-normality-assumption}

Testing the normality assumption is relatively straightforward. We covered most of what you need to know in @sec-Checking-the-normality-of-a-sample. The only thing we really need to do is draw a QQ plot and, in addition if it is available, run the Shapiro-Wilk test. The QQ plot is shown in @fig-fig13-7 and it looks pretty normal to me. If the Shapiro-Wilk test is not significant (i.e. $p > .05$) then this indicates that the assumption of normality is not violated. However, as with Levene's test, if the sample size is large then a significant Shapiro-Wilk test may in fact be a false positive, where the assumption of normality is not violated in any substantive problematic sense for the analysis. And, similarly, a very small sample can produce false negatives. That's why a visual inspection of the QQ plot is important.

```{r}

#| label: fig-fig13-7

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: QQ plot in the One Way ANOVA analysis in jamovi

knitr::include\_graphics("images/fig13-7.png")

```

Alongside inspecting the QQ plot for any deviations from normality, the Shapiro-Wilk test for our data does show a non-significant effect, with p = 0.6053 (see @fig-fig13-6. This therefore supports the QQ plot assessment; both checks find no indication that normality is violated.

### Removing the normality assumption

Now that we've seen how to check for normality, we are led naturally to ask what we can do to address violations of normality. In the context of a one-way ANOVA, the easiest solution is probably to switch to a non-parametric test (i.e., one that doesn't rely on any particular assumption about the kind of distribution involved). We've seen non-parametric tests before, in the chapter [Comparing two means]. When you only have two groups, the Mann-Whitney or the Wilcoxon test provides the non-parametric alternative that you need. When you've got three or more groups, you can use the \*\*Kruskal-Wallis rank sum test\*\* [@KruskalWallis1952]. So that's the test we'll talk about next.

### The logic behind the Kruskal-Wallis test

The Kruskal-Wallis test is surprisingly similar to ANOVA, in some ways. In ANOVA we started with $Y\_{ik}$, the value of the outcome variable for the ith person in the kth group. For the Kruskal Wallis test what we'll do is rank order all of these $Y\_{ik}$ values and conduct our analysis on the ranked data. [^13-comparing-several-means-one-way-anova-14]

[^13-comparing-several-means-one-way-anova-14]: So let's let R\\_{ik} refer to the ranking given to the ith member of the kth group. Now, let's calculate $\bar{R}\_k$, the average rank given to observations in the kth group $$\bar{R}\_k=\frac{1}{N\_k}\sum\_i R\_{ik}$$ and let's also calculate $\bar{R}$, the grand mean rank $$\bar{R}=\frac{1}{N}\sum\_i\sum\_k R\_{ik}$$ Now that we've done this, we can calculate the squared deviations from the grand mean rank $\bar{R}$. When we do this for the individual scores, i.e., if we calculate $(R\_{ik} - \bar{R})^2$ , what we have is a "nonparametric" measure of how far the ik-th observation deviates from the grand mean rank. When we calculate the squared deviation of the group means from the grand means, i.e., if we calculate $(R\_{ik} - \bar{R})^2$, then what we have is a nonparametric measure of how much the group deviates from the grand mean rank. With this in mind, we'll follow the same logic that we did with ANOVA and define our ranked sums of squares measures, much like we did earlier. First, we have our "total ranked sums of squares" $$RSS\_{tot}=\sum\_k\sum\_i (R\_{ik}-\bar{R})^2$$ and we can define the "between groups ranked sums of squares" like this $$\begin{aligned} RSS\_{b}& =\sum{k}\sum\_{i}(\bar{R}\_{k}-\bar{R})^2 \\ &= \sum\_{k} N\_k (\bar{R}\_{k}-\bar{R})^2 \end{aligned}$$ So, if the null hypothesis is true and there are no true group differences at all, you'd expect the between group rank sums $RSS\_b$ to be very small, much smaller than the total rank sums $RSS\_{tot}$. Qualitatively this is very much the same as what we found when we went about constructing the ANOVA F-statistic, but for technical reasons the Kruskal-Wallis test statistic, usually denoted K, is constructed in a slightly different way, $$K=(N-1) \times \frac{RSS\_b}{RSS\_{tot}}$$ and if the null hypothesis is true, then the sampling distribution of K is approximately chi square with $G-1$ degrees of freedom (where $G$ is the number of groups). The larger the value of K, the less consistent the data are with the null hypothesis, so this is a one-sided test. We reject $H\_0$ when K is sufficiently large.

### Additional details

The description in the previous section illustrates the logic behind the Kruskal-Wallis test. At a conceptual level, this is the right way to think about how the test works.[^13-comparing-several-means-one-way-anova-15]

[^13-comparing-several-means-one-way-anova-15]: However, from a purely mathematical perspective it's needlessly complicated. I won't show you the derivation, but you can use a bit of algebraic jiggery-pokery$^b$ to show that the equation for K can be $$K=\frac{12}{N(N-1)}\sum\_k N\_k \bar{R}\_k^2 -3(N+1)$$ It's this last equation that you sometimes see given for K. This is way easier to calculate than the version I described in the previous section, but it's just that it's totally meaningless to actual humans. It's probably best to think of K the way I described it earlier, as an analogue of ANOVA based on ranks. But keep in mind that the test statistic that gets calculated ends up with a rather different look to it than the one we used for our original ANOVA. <br>---<br> $b$ A technical term

But wait, there's more! Dear lord, why is there always more? The story I've told so far is only actually true when there are no ties in the raw data. That is, if there are no two observations that have exactly the same value. If there are ties, then we have to introduce a correction factor to these calculations. At this point I'm assuming that even the most diligent reader has stopped caring (or at least formed the opinion that the tie-correction factor is something that doesn't require their immediate attention). So I'll very quickly tell you how it's calculated, and omit the tedious details about why it's done this way. Suppose we construct a frequency table for the raw data, and let fj be the number of observations that have the j-th unique value. This might sound a bit abstract, so here's a concrete example from the frequency table of mood.gain from the clinicaltrials.csv data set (@tbl-tab13-11)

```{r}

#| label: tbl-tab13-11

#| tbl-cap: Frequency table of mood gain from the clinicaltrials.csv data

huxtabs[[13]][[11]]

```

Looking at this table, notice that the third entry in the frequency table has a value of 2. Since this corresponds to a mood.gain of 0.3, this table is telling us that two people's mood increased by 0.3. [^13-comparing-several-means-one-way-anova-16]

[^13-comparing-several-means-one-way-anova-16]: More to the point, in the mathematical notation I introduced above, this is telling us that $f\_3 = 2$. Yay. So, now that we know this, the tie correction factor (TCF) is: $$TCF=1-\frac{\sum\_j f\_j^3 - f\_j}{N^3 - N}$$ The tie-corrected value of the Kruskal-Wallis statistic is obtained by dividing the value of K by this quantity. It is this tie-corrected version that jamovi calculates.

And so jamovi uses a tie-correction factor to calculate the tie-corrected Kruskall-Wallis statistic. And at long last, we're actually finished with the theory of the Kruskal-Wallis test. I'm sure you're all terribly relieved that I've cured you of the existential anxiety that naturally arises when you realise that you don't know how to calculate the tie-correction factor for the Kruskal-Wallis test. Right?

### How to run the Kruskal-Wallis test in jamovi

Despite the horror that we've gone through in trying to understand what the Kruskal Wallis test actually does, it turns out that running the test is pretty painless, since jamovi has an analysis as part of the ANOVA analysis set called 'Non-Parametric' - 'One Way ANOVA (Kruskall-Wallis)' Most of the time you'll have data like the clinicaltrial.csv data set, in which you have your outcome variable mood.gain and a grouping variable drug. If so, you can just go ahead and run the analysis in jamovi. What this gives us is a Kruskal-Wallis $\chi^2 =12.076, df = 2, p = 0.00239$, as in @fig-fig13-8

```{r}

#| label: fig-fig13-8

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: Kruskall-Wallis one-way non-parametric ANOVA in jamovi

knitr::include\_graphics("images/fig13-8.png")

```

## Repeated measures one-way ANOVA

The one-way repeated measures ANOVA test is a statistical method of testing for significant differences between three or more groups where the same participants are used in each group (or each participant is closely matched with participants in other experimental groups). For this reason, there should always be an equal number of scores (data points) in each experimental group. This type of design and analysis can also be called a 'related ANOVA' or a 'within subjects ANOVA'.

The logic behind a repeated measures ANOVA is very similar to that of an independent ANOVA (sometimes called a 'between-subjects' ANOVA). You'll remember that earlier we showed that in a between-subjects ANOVA total variability is partitioned into between-groups variability ($SS\_b$) and within-groups variability ($SS\_w$), and after each is divided by the respective degrees of freedom to give MSb and MSw (see Table 13.1) the F-ratio is calculated as:

$$F=\frac{MS\_b}{MS\_w}$$

In a repeated measures ANOVA, the F-ratio is calculated in a similar way, but whereas in an independent ANOVA the within-group variability ($SS\_w$) is used as the basis for the $MS\_w$ denominator, in a repeated measures ANOVA the $SS\_w$ is partioned into two parts. As we are using the same subjects in each group, we can remove the variability due to the individual differences between subjects (referred to as SSsubjects) from the within-groups variability. We won't go into too much technical detail about how this is done, but essentially each subject becomes a level of a factor called subjects. The variability in this within-subjects factor is then calculated in the same way as any between-subjects factor. And then we can subtract SSsubjects from $SS\_w$ to provide a smaller SSerror term:

$$\text{Independent ANOVA: } SS\_{error} = SS\_w$$ $$\text{Repeated Measures ANOVA: } SS\_{error} = SS\_w - SS\_{subjects}$$

This change in $SS\_{error}$ term often leads to a more powerful statistical test, but this does depend on whether the reduction in the $SS\_{error}$ more than compensates for the reduction in degrees of freedom for the error term (as degrees of freedom go from $(n - k)$ [^13-comparing-several-means-one-way-anova-17] to $(n - 1)(k - 1)$ (remembering that there are more subjects in the independent ANOVA design).

[^13-comparing-several-means-one-way-anova-17]: (n - k) : (number of subjects - number of groups)

### Repeated measures ANOVA in jamovi

First, we need some data. @Geschwind1972 has suggested that the exact nature of a patient's language deficit following a stroke can be used to diagnose the specific region of the brain that has been damaged. A researcher is concerned with identifying the specific communication difficulties experienced by six patients suffering from Broca's Aphasia (a language deficit commonly experienced following a stroke) (@tbl-tab13-12).

```{r}

#| label: tbl-tab13-12

#| tbl-cap: Word recognition task scores in stroke patients

huxtabs[[13]][[12]]

```

The patients were required to complete three word recognition tasks. On the first (speech production) task, patients were required to repeat single words read out aloud by the researcher. On the second (conceptual) task, designed to test word comprehension, patients were required to match a series of pictures with their correct name. On the third (syntax) task, designed to test knowledge of correct word order, patients were asked to reorder syntactically incorrect sentences. Each patient completed all three tasks. The order in which patients attempted the tasks was counterbalanced between participants. Each task consisted of a series of 10 attempts. The number of attempts successfully completed by each patient are shown in @tbl-tab13-11. Enter these data into jamovi ready for analysis (or take a short-cut and load up the broca.csv file).

To perform a one-way related ANOVA in jamovi, open the one-way repeated measures ANOVA dialogue box, as in @fig-fig13-9, via ANOVA - Repeated Measures ANOVA.

```{r}

#| label: fig-fig13-9

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: Repeated measures ANOVA dialogue box in jamovi

knitr::include\_graphics("images/fig13-9.png")

```

Then:

- Enter a Repeated Measures Factor Name. This should be a label that you choose to describe the conditions repeated by all participants. For example, to describe the speech, conceptual and syntax tasks completed by all participants a suitable label would be 'Task'. Note that this new factor name represents the independent variable in the analysis.

- Add a third level in the Repeated Measures Factors text box, as there are three levels representing the three tasks: speech, conceptual and syntax. Change the labels of the levels accordingly.

- Then move each of the levels variables across to the Repeated Measures Cells text box.

- Finally, under the Assumption Checks option, tick the "Sphericity checks" text box.

jamovi output for a one-way repeated measures ANOVA is produced as shown in @fig-fig13-10 to @fig-fig13-13. The first output we should look at is Mauchly's Test of Sphericity, which tests the hypothesis that the variances of the differences between the conditions are equal (meaning that the spread of difference scores between the study conditions is approximately the same). In @fig-fig13-10 Mauchly's test significance level is $p = .720$. If Mauchly's test is non-significant (i.e. p < .05, as is the case in this analysis) then it is reasonable to conclude that the variances of the differences are not significantly different (i.e. they are roughly equal and sphericity can be assumed.).

```{r}

#| label: fig-fig13-10

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: One-way repeated measures ANOVA output - Mauchly's Test of Sphericity

knitr::include\_graphics("images/fig13-10.png")

```

If, on the other hand, Mauchly's test had been significant (p < .05) then we would conclude that there are significant differences between the variance of the differences, and the requirement of sphericity has not been met. In this case, we should apply a correction to the F-value obtained in the one-way related ANOVA analysis:

- If the Greenhouse-Geisser value in the "Tests of Sphericity" table is > .75 then you should use the Huynh-Feldt correction

- But if the Greenhouse-Geisser value is < .75, then you should use the Greenhouse-Geisser correction.

Both these corrected F-values can be specified in the Sphericity Corrections check boxes under the Assumption Checks options, and the corrected F-values are then shown in the results table, as in Figure \*\*13.11\*\*.

```{r}

#| label: fig-fig13-11

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: One-way repeated measures ANOVA output - Tests of Within-Subjects Effects

knitr::include\_graphics("images/fig13-11.png")

```

In our analysis, we saw that the significance of Mauchly's Test of Sphericity was p = .720 (i.e. p > 0.05). So, this means we can assume that the requirement of sphericity has been met so no correction to the F-value is needed. Therefore, we can use the 'None' Sphericity Correction output values for the repeated measure 'Task': $F = 6.93$, $df = 2$, $p = .013$, and we can conclude that the number of tests successfully completed on each language task did vary significantly depending on whether the task was speech, comprehension or syntax based ($F(2, 10) = 6.93$, $p = .013$).

Post-hoc tests can also be specified in jamovi for repeated measures ANOVA in the same way as for independent ANOVA. The results are shown in @fig-fig13-12. These indicate that there is a significant difference between Speech and Syntax, but not between other levels.

```{r}

#| label: fig-fig13-12

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: Post-hoc tests in repeated measures ANOVA in jamovi

knitr::include\_graphics("images/fig13-12.png")

```

Descriptive statistics (marginal means) can be reviewed to help interpret the results, produced in the jamovi output as in @fig-fig13-13. Comparison of the mean number of trials successfully completed by participants shows that Broca's Aphasics perform reasonably well on speech production (mean = 7.17) and language comprehension (mean = 6.17) tasks. However, their performance was considerably worse on the syntax task (mean = 4.33), with a significant difference in post-hoc tests between Speech and Syntax task performance.

```{r}

#| label: fig-fig13-13

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: One-way repeated measures ANOVA output - Descriptive Statistics

knitr::include\_graphics("images/fig13-13.png")

```

## The Friedman non-parametric repeated measures ANOVA test

The Friedman test is a non-parametric version of a repeated measures ANOVA and can be used instead of the Kruskall-Wallis test when testing for differences between three or more groups where the same participants are in each group, or each participant is closely matched with participants in other conditions. If the dependent variable is ordinal, or if the assumption of normality is not met, then the Friedman test can be used.

As with the Kruskall-Wallis test, the underlying mathematics is complicated, and won't be presented here. For the purpose of this book, it is sufficient to note that jamovi calculates the tie-corrected version of the Friedman test, and in @fig-fig13-14 there is an example using the Broca's Aphasia data we have already looked at.

```{r}

#| label: fig-fig13-14

#| fig-align: left

#| fig-width: 0.9

#| fig-cap: The 'Repeated Measures ANOVA (Non-parametric)' dialogue box and results in jamovi

knitr::include\_graphics("images/fig13-14.png")

```

It's pretty straightforward to run a Friedman test in jamovi. Just select Analyses - ANOVA - Repeated Measures ANOVA (Non-parametric), as in @fig-fig13-14. Then highlight and transfer the names of the repeated measures variables you wish to compare (Speech, Conceptual, Syntax) into the 'Measures:' text box. To produce descriptive statistics (means and medians) for the three repeated measures variables, click on the Descriptives button

The jamovi results show descriptive statistics, chi-square value, degrees of freedom, and the p-value (@fig-fig13-14). Since the p-value is less than the level conventionally used to determine significance (p < .05), we can conclude that Broca's Aphasics perform reasonably well on speech production (median = 7.5) and language comprehension (median = 6.5) tasks. However, their performance was considerably worse on the syntax task (median = 4.5), with a significant difference in post-hoc tests between Speech and Syntax task performance.

## On the relationship between ANOVA and the Student t-test {#sec-On-the-relationship-between-ANOVA-and-the-Student-t-test}

There's one last thing I want to point out before finishing. It's something that a lot of people find kind of surprising, but it's worth knowing about. An ANOVA with two groups is identical to the Student t-test. No, really. It's not just that they are similar, but they are actually equivalent in every meaningful way. I won't try to prove that this is always true, but I will show you a single concrete demonstration. Suppose that, instead of running an ANOVA on our mood.gain \~ drug model, let's instead do it using therapy as the predictor. If we run this ANOVA we get an F-statistic of $F(1,16) = 1.71$, and a p-value = $0.21$. Since we only have two groups, I didn't actually need to resort to an ANOVA, I could have just decided to run a Student t-test. So let's see what happens when I do that: I get a t-statistic of $t(16) = -1.3068$ and a $p-value = 0.21$. Curiously, the p-values are identical. Once again we obtain a value of $p = .21$. But what about the test statistic? Having run a t-test instead of an ANOVA, we get a somewhat different answer, namely $t(16) = -1.3068$. However, there is a fairly straightforward relationship here. If we square the t-statistic then we get the F-statistic from before: $-1.3068^{2} = 1.7077$

## Summary

There's a fair bit covered in this chapter, but there's still a lot missing [^13-summary-1]. Most obviously, I haven't discussed how to run an ANOVA when you are interested in more than one grouping variable, but that will be discussed in a lot of detail in @sec-Factorial-ANOVA. In terms of what we have discussed, the key topics were:

- The basic logic behind [How ANOVA works] and [Running an ANOVA in jamovi]

- How to compute an [Effect size] for an ANOVA.

- [Multiple comparisons and post hoc tests] for multiple testing.

- [The assumptions of one-way ANOVA]

- [Checking the homogeneity of variance assumption] and what to do if it is violated: [Removing the homogeneity of variance assumption]

- [Checking the normality assumption] and what to do if it is violated: [Removing the normality assumption]

- [Repeated measures one-way ANOVA] and the non-parametric equivalent, [The Friedman non-parametric repeated measures ANOVA test]

[^13-summary-1]: As with all of the chapters in this book, there are quite a few different sources that I've relied upon, but the one stand-out text that I've been most heavily influenced by @Sahai2000. It's not a good book for beginners, but it's an excellent book for more advanced readers who are interested in understanding the mathematics behind ANOVA.