Tutorial

The F distribution and the basic principle behind ANOVAs

Bodo Winter[1]

Last updated: September 21, 2011

Why care about this tutorial?

Before I start, let me explain why I think that this tutorial is useful for you. You might think that fancy techniques such as generalized linear models and mixed effects models render traditional approaches such as ANOVAs (="Analysis of Variance") obsolete. For example, in psycholinguistics, linear mixed effects models are used for many problems that have formerly been addressed with simple repeated measures ANOVAs. So why should you care about "going back" to ANOVAs?

Well, first of all, in order to understand some of the more "fancy" statistics, it is often useful to first understand the basic statistics. And, often, the fancy statistics subsume some of the simpler approaches. For example, a usual generalized linear model gives you an ANOVA table as output, and I would contend that you only really understand what is reported in such a table if you have – at least once – computed everything yourself. Therefore, this tutorial focuses on *understanding* rather than simply using ANOVAs. We will actually go through an ANOVA analysis step-by-step, using R merely as a calculator.

Situating ANOVAs in the world of statistical tests

ANOVA is a type of regression that is used for testing the effect of a categorical predictor variable (a so-called "fixed effect", independent variable or factor) on a continuous dependent variable (what was measured in your study). An example of a categorical predictor would be "male versus female" or condition A versus condition B. Continuous measures can be anything from pitch to reaction times, anything that has the property of being interval-scaled (e.g. a frequency of 200 Hertz is double a frequency of 100 Hertz).

[1] For updates and other tutorials, check my webpage www.bodowinter.com. If you have any suggestions, please write me an email: bodo@bodowinter.com

The following table shows how ANOVA differs from other types of regression, namely in what the nature of the dependent and the independent variable are.

Regression	continuous dependent measure, continuous predictors
Logistic Regression	categorical dependent measure, continuous predictors
ANOVA	continuous dependent measure, categorical predictors

There are different kinds of ANOVAs, some of which are listed below:

One-way independent	one factor, each observations are independent
Two-way independent	two factors, each observations are independent
One-way repeated measures	one factor, multiple observations from the same subjects
…	…

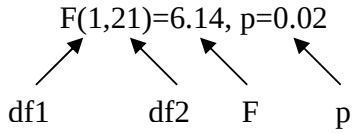
In a one-way independent ANOVA, there is only one factor with multiple levels (two, three, four etc.). Each observation must come from one individual that is not re-used in the same experiment, i.e. each observation needs to be independent. You might ask the question: Hey, but if there's only one factor with two levels, that's exactly like a t-test, isn't it? For example, in an independent samples t-test, you would have two conditions and test whether there's a difference. And yes, you would be spot on with this question as a one-way independent ANOVA and an independent t-test lead to exactly the same result. Therefore, you should use an ANOVA as soon as you have more than two levels…. or if you have more than two factors (e.g. two-way, three-way, four-way ANOVAs).

In this tutorial, we will focus on the one-way independent ANOVA and in our example our one predictor has three levels. In using this test, we are looking at three groups or conditions, and we ask the question: is there a difference between any one of these groups that is unlikely due to chance?

So, let's get started with the basics of ANOVAs!

The F-value

At the heart of every type of ANOVA lies the F-value. Whenever the result of an ANOVA appears in a published research article, usually something such as the following is reported:



Journals almost always require researchers to provide the degrees of freedom, the F-value and the p-value… but unfortunately, many people (including reviewers) only look at the p-value, overlooking the degrees of freedom ("1" and "21" in this case) and the F-value ("6.14"). This is dangerous, because if the degrees of freedom are not correct, the F-value and the p-value are practically meaningless (cf. Hurlbert, 1984). In other cases, the F-value can actually be more informative than the p-value. So, let's go through each of these values and see what they actually mean.

The F-value is actually the quotient of the following ratio:

$$F = \frac{\text{Effect Variance (or "Treatment Variance")}}{\text{Error Variance}}$$

Or, sometimes the variances in the ratio are labeled like this:

$$F = \frac{\text{Between-group Variance}}{\text{Within-group Variance}}$$

Imagine you have the dataset below of average voice pitch (measured in Hertz) for male and female participants.

Male Participants	Female Participants
100 Hz	190 Hz
85 Hz	230 Hz
130 Hz	200 Hz

As you can see, male participants have lower voice pitch than female participants, but then, there's also a lot of variation within groups. Each participant will have slightly different voice pitch due to physiological or psychological differences. These differences within the male group and within the female group are called "within-group variance" or "error variance". It's the variation that you can't control for, the variation that is due to individual differences.

Now, what you're usually interested in is the variation that is caused by your experimental manipulation or your independent variable. In the above case, you could be interested in the difference between male and female voice pitch – for this, you would need the "between-group variance" or the effect variance. This is

what you're interested in, this is the systematic effect your study sets out to investigate. So, looking back at the ratio…

$$F = \frac{\text{Between-group Variance}}{\text{Within-group Variance}}$$

…we can see that a large amount of between-group variance (= "effect variance") will lead to a higher F ratio (because the between-group variance is in the numerator), and a large amount of variance that is due to chance will lead to a smaller F ratio (because the within-group variance is in the denominator). Now, in any kind of statistical testing, it is usually the case that the more random variation there is in a sample, the more difficult it will be to detect any consistent patterns. Or, if we do find consistent patterns, we will be less confident in these patterns because with more random variation, the patterns could actually be due to chance. It is also the case that the larger the difference between conditions, the easier it will be to find a pattern (namely that difference) despite random variation. This makes sense intuitively: a needle in the haystack is more difficult to find than a baseball bat. We will later see that the effect variance or the between-groups variance reflects this difference between conditions.

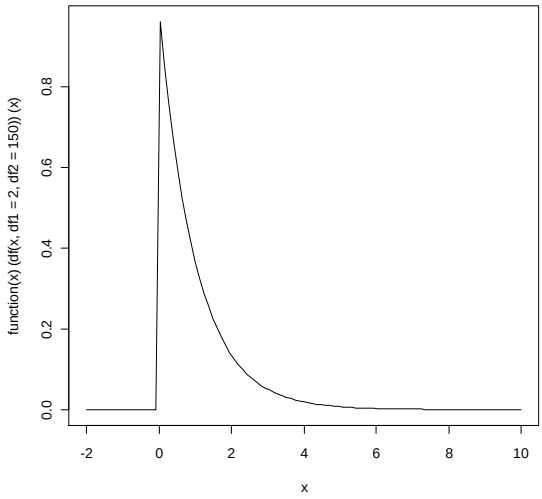
All of this means that the larger an F-value, the better for you to find a "significant" effect, a consistent pattern that is unlikely due to chance. So, the simple rule in most research situations is: the higher the F value, the better…

This was an explanation of the F-value. Now, basically, what we're doing with an ANOVA is the following: we look at *how unexpected* an F-value that we obtained in our study is. A very large F-value means that the between-group variance (the effect variance) exceeds the within-group variance (the error variance) by a substantial amount. The p-value then just gives a number to how likely a particular F-value is going to occur, with lower p-values indicating that the probability of obtaining that particular F-value is pretty low.

The core of any ANOVA, the F-test

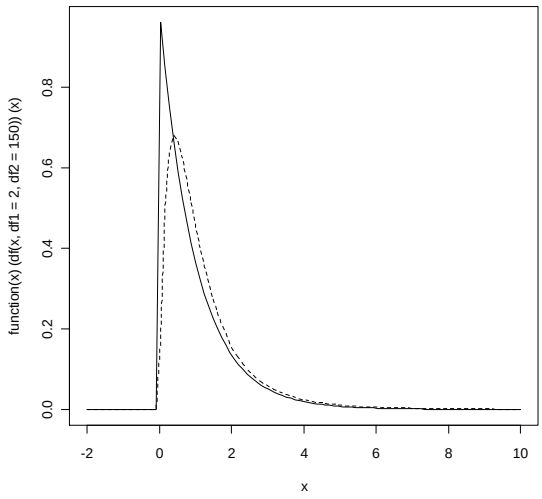
Now, we can use the F-value that we obtained based on our data and compare it to a probability distribution of F-values. This probability distribution is the distribution of F-values that we expect to occur by chance alone, it is the distribution of F-values that we expect to occur under the null hypothesis (that there is no difference between groups or conditions). It's important that we get a good feel of this distribution, so let's plot it! **[this is where you need to open R]**

```
plot(function(x)(df(x,df1=2,df2=150)),xlim=c(-
    2,10))
```



Let's try to understand the command that we just used. *df( )* gives you the probability density function of the F-value. The command takes three arguments – an F-value and two degrees of freedom – and from this, it calculates the likelihood of observing an F-value equal to or greater than the one that we specify. In this case, the F-value is simply "x", handed down from the command *function(x)*. We will later talk about degrees of freedom in more detail, but for now, let's recognize that different values for these degrees of freedom change the shape of the F distribution. We can see this by adding another F distribution with slightly different degrees of freedom [you need to have the preceding plot being displayed in your R interface].

```
plot(function(x)(df(x,df1=4,df2=10)),xlim=c(-
     2,10),lty="dashed",add=T)
```



So, the F distribution is not just one particular distribution, it is a *family of distributions*, and the degrees of freedom are the parameters that determine the exact shape of a particular distribution (e.g. how far spread out it is).

Let's look at the F distribution graph above in somewhat more detail. The x-axis gives you F-values and the y-axis gives you the probability of observing such an F-value ("the density"). The density for each distribution (the area under the line) adds up to 1. There are several important properties of the F distribution: First, the distribution starts at the point x=0, y=0. Because of this, an F-value of "0" will never occur, which makes sense because the F-value is a ratio, and ratios are always above 0… And, as you can see, the F-value has no negative values. However, you can get an F-value of between 0 and 1 – that happens, of course, when the denominator (the within-group variance) is larger than the numerator (the between-group variance). In published research, when the authors want to convince you that they did not obtain a significant difference in one particular condition, you can often read something like "all F < 1". This basically means that the error variance exceeds the effect variance.

Towards the right side of the F distribution, you can see that this distribution has a long tail that asymptotes out to the x-axis. By "asymptoting" towards the x-axis, I mean that it will never reach a y-value of exactly 0 with these very high values. Therefore, very very high F-values such as "5657" still have some probability, although a very very very small one. But in any case, extremely high F-values are possible. Such values might, for example, occur if there is an extremely strong difference between conditions (increasing the between-group variance, the numerator of the F-value), or if a dataset has an extremely low amount of individual differences (decreasing the within-group variance, the denominator).

The command *pf( )* gives us the probability of observing a value *lower* than specific F-value given two degrees of freedom. Therefore, if you take the reverse, *1-pf( )*, you get the probability of observing a value as high or higher as your F-value. Again, let's not worry about the degrees of freedom for now, just look at the output that R gives you if you type the following two commands. The first argument specifies the F-value for which you want to obtain a probability value, df1 and df2 specify the degrees of freedom.

```
1-pf(4,df=1,df2=150)
1-pf(2,df=1,df2=150)
```

```
> 1-pf(2,df=1,df2=150)
[1] 0.1593715
> 1-pf(4,df=1,df2=150)
[1] 0.04730553
```

The output says that an F-value of 4 (the first argument) or higher has the probability of p=0.04730553, whereas an F-value of 2 or higher has the

probability of p=0.1593715. So, an F-value of 4 is much less likely going to occur than an F-value of 2. From a conceptual viewpoint, this makes sense: The world is full of randomness, and there's almost always bound to be small differences between whatever two treatments, conditions or groups we are investigating. For example, even if the experimental manipulation of a study actually doesn't do anything to the dependent measure, we will find "apparent" differences between conditions that are due to individual differences or intra-individual variation (two means are almost never be exactly the same). Based on this reasoning, we would expect a lot of small F-values and the higher the F-value, the more are we going to be surprised about it because it is likely to occur due to chance under the null hypothesis. This is why the F distribution has high densities between 1 and 2 and much lower densities towards the right end, its tail.

Degrees of freedom

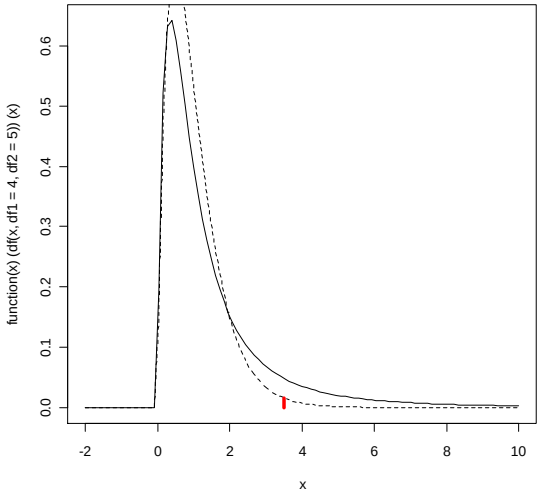
So far, we've evaded the concept of "degrees of freedom". The easiest way to get a grasp of this concept is by means of an example: Let's say you have 5 positive numbers that add up to 10 and assume that you already *know* that the numbers add up to ten. The first number could be any number, let's say 2. Now, you can't predict what the second number will be … because there are different combinations that will lead to the same total 10. For example, five numbers that add up to ten could be 2-1-3-1-3 or 2-1-1-1-5. But let's say that we picked another 2. Again, you can't predict the next number of the sequence 2-2- because there's still different combinations that add up to ten. This Spiel can go on and on until you reach the fourth number. Let's say that the combination is now 2-2-1-3. The last number can only be 2 in order for the whole sequence of numbers to add up to 10. So while the first four numbers are "allowed" to vary freely, the last number is set (given that they have to add up to a specific number). In this particular example, we therefore had 4 degrees of freedom (= 5 minus 1).

Now, you might ask: but in this made-up example, you already now the total 10 "in advance." How could that be the case in actual research, if you don't know a summary number (a mean or a sum) in advance? Why are the degrees of freedom $n - 1$ rather than just n. In the case of ANOVAs, the reason for this is simple: ANOVAs works by comparing variances. Now, as we will see later in more detail, you need the mean in order to compute the variance, therefore, the mean is already known before you compute the variance, and that's why the variance has $n - 1$ rather than just n degrees of freedom.

The degrees of freedom reflect the independent pieces of information in your sequence of numbers. The last number of the above-mentioned sequence wasn't independent in that its value completely depended on the values of the other numbers. But the first four numbers were independent in that they could have had other values.

The degrees of freedom thus reflect the true sample size: with more data points (e.g. more people that participate in your study), you will have more independent pieces of information, more stuff that is allowed to "vary freely". This is actually why the degrees of freedom influence the F distribution. Because with "more stuff" that is allowed to vary, observing the same difference between conditions will have more meaning, simply because there are more reasons why it could have been otherwise. Let's do some more plotting to understand this aspect of the degrees of freedom:

```
plot(function(x)(df(x,df1=4,df2=5)),xlim=c(-2,10))
plot(function(x)(df(x,df1=4,df2=100)),xlim=c(-
     2,10),add=T,lty="dashed")
lines(x=c(3.5,3.5),y=c(0,df(3.5,df1=4,df2=100)),col
     ="red",lwd=4)
```



I highlighted the F-value 3.5 with a red line under the curve. You see that the dashed F distribution, the one with higher degrees of freedom has less density than the solid distribution. This means that an F-value of "3.5" is much more surprising (= much more unexpected under chance circumstances) if it is based on many independent observations rather than on just a few. Therefore, smaller F-values are more likely to be "significant" with more *independent* data points. Again, this makes sense conceptually: if you have a small sample, you might obtain a big F-value simply because you happened to have chosen some subjects or items that did not vary a lot (low error variance) or that happened to exhibit a large difference (big effect variance). Therefore, with small datasets, it's relatively easier to find big F-values (they have a higher probability). This means that for smaller samples, you need relatively larger F-values in order for it to count more towards statistical significance in a testing context. The smaller the sample, the higher your F-value has to be in order to reach significance.

This aspect of the F distribution is one of the reasons why an F-value can sometimes tell you more than a p-value. Let's say you got a pretty high F-value

(e.g., one that is above 4), but the p-value indicates "non-significance", it is above 0.05. Then, a lot of people would say "oh, sad, I didn't find a significant result" …. but without looking at the degrees of freedom, you cannot conclude anything too quickly: a high F-value might – with more data and more degrees of freedom – become significant. A high F-value in a small dataset indicates that within that dataset the effect variance well exceeds the error variance … and therefore, it is useful to collect more data in order to see whether this high F-value is really just due to chance or actually something meaningful.

The upshot of this: more degrees of freedom mean that you will have smaller p-values, more "significant" results. Let's check this using the command *pf( )*. Remember, *pf( )* gives you the probability of observing any value *lower* than the F-value you specified. So, the probability of observing an F-value as high or higher as the one you found would be *1-pf()*.

```
1-pf(3.5,df1=4,df2=5)
1-pf(3.5,df1=4,df2=100)
```

The first command is not significant at all, the second one is … even though you're dealing with the same F-value in both cases.

What I've just talked about with respect to degrees of freedom is simply a variant of the general rule that with larger sample sizes, you are more likely to get significant results. However, the data points need to be *independent*. The degrees of freedom reflect the true sample size (the sample of individual subjects, the sample of unique experimental stimuli) and not the total dataset. For example, if you have 10 responses per subject, and you collected data from 8 subjects, your degrees of freedom have to stay below 7 (= 8-1).

Now, we've covered the basics (degrees of freedom, F value), and we can proceed to finally performing an actual ANOVA.

Independent one-factor ANOVA: A work-through example

Let's start by constructing an adequate dataset… let's simulate some data! Let's say you were interested in whether females, males and kids have different voice pitches. We'll sample four subjects in each group and measure their voice pitch in Hertz (= the dependent variable). We will use the *rnorm( )* function to generate four random numbers drawn from a normal distribution with the mean 200 Hertz and a standard deviation of 20. This will be our simulated female data points. Then, we will draw four values for male subjects with the mean 120 Hertz and four values for the child subjects with the mean 380 (Hertz).

We will save these in three vectors called "females", "males" and "kidz" … and for the sake of displaying the data, let's also put them together in a table using the command *data.frame( )* on the three vectors that we just constructed.

```
females = rnorm(4, mean=200, sd=20)
males = rnorm(4, mean=120, sd=20)
kidz = rnorm(4, mean=380, sd=20)
pitchstudy = data.frame(females, males, kidz)
```

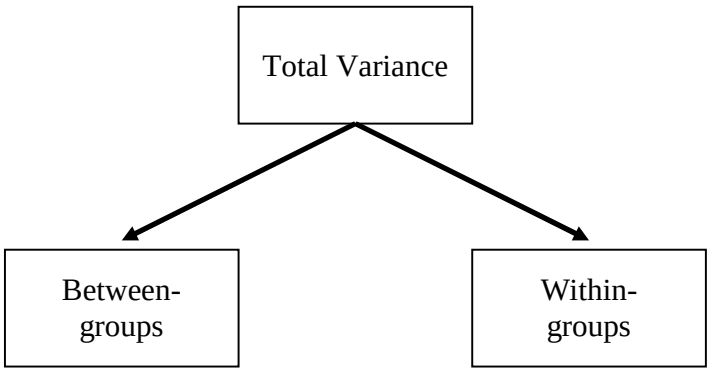
If you print the object "pitchstudy", you should see something like this. Of course, the actual values will be somewhat different for you because the numbers are generated randomly.

```
   females    males      kidz
1  213.6684   107.2392   381.5930
2  204.4267   148.6768   438.9887
3  184.4746   107.7622   359.4772
4  194.7927   95.1473    378.7686
```

As we simulated the data, females have higher pitch than males, and kids have higher pitch than both females and males. Now, the ANOVA tests whether this difference between the three groups is significant, whether it is unlikely due to chance.

Variance structure

As was mentioned above, the heart of the ANOVA is the F-test, and the F-value is a ratio of two variances – the "between group" variance and the "within-group" variance. To get an overview of the "variance structure" of our dataset above, have a look at the following variance breakdown.



This graph depicts the variance structure of the data in this fictive experiment. The total variance "consists of" or can be subdivided into the "between-groups" and the "within-groups" variance. What we actually want to compare is the between-groups and the within-groups variance, but often, it is easier to compute the total variance. Because the variance is structured as in this graph, we can calculate the within-groups variance by subtracting the between-groups variance from the total variance. Here, it becomes clear why some people call the within-group variance "residual variance": if you subtract the between-group variance or your "effect variance", then what's left (the "residual") is the within-group variance.

Two ways to calculate the variance

Dieser Abschnitt rechnet die ANOVA 'on Hand' aus - nicht verwirren lassen oder Überspringen

The variance (the statistical characterization of the everyday concept of "variation") is defined as follows:

$$\text{Variance} = \frac{\text{Sum of squares}}{\text{Degrees of freedom}}$$

The "sum of squares" is shorthand for saying *the sum of squared deviations from the mean*. There are two ways to calculate these sum of squares. Have a look at the following equation:

$$\frac{\sum(x-\bar{x})^2}{n-1} = \frac{\sum(x^2)-(\sum x)^2}{n-1}$$

The first method for calculating the sum of squares works by subtracting the mean of a set of numbers from each of the numbers in the set, squaring this difference, and summing all the differences together. This way of calculating actually quite well captures what "sum of squares" actually means: *the sum of squared deviations from the mean*. The formula on the right side of the equation seems to be somewhat more detached from the term "sum of squares"; it is calculated via adding all squared values together and subtracting the squared total. Both sides of the equation are divided by n-1, the degrees of freedom. Check whether both formulas actually lead to the same result by applying it to the vector of female data that we constructed.

```
sum((females-mean(females))^2)
sum(females^2)-(sum(females)^2)/4
```

Both commands give you the same number.

The reason why I bother you with both of these formulas is because the first one actually is conceptually easier to grasp (it makes sense that variation is measured by how much individual data points deviate from the mean), but the second one is actually the one that is used in a number of textbooks and, as you will see later, the second one has some computational advantages: with the second formula, we can simply calculate the sum of squares with *sum(vector^2)* of whatever part of the data we're looking at, and then we can subtract the so-called "correction factor" which is going to remain constant for the whole ANOVA computation. The correction factor is simply the squared total in the right side of the above equation, the part that is subtracted from the sum of squared values.

Let's start by calculating the total variance. To do this easily in R, we'll first put all the three groups together in one vector.

```
bigvector = c(females, males, kidz)
```

Then, we'll calculate the correction factor that we'll use again and again.

```
CF = (sum(bigvector))^2/length(bigvector)
```

Then we calculate the so-called "total sum of squares" (the numerator of the variance formula).

```
total.ss = sum(bigvector^2)-CF
```

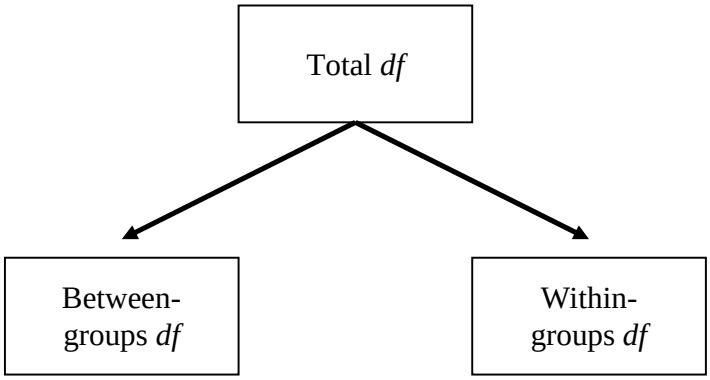
Now that we've calculated the total sum of squares, let's calculate the between group sum of squares. This is done by summing each column, squaring it and dividing it by the number of cells within each column. Add these squared group totals together and subtract by the correction factor to get "between.ss". This step might seem a little bit awkward, but you can think of it this way: each column represents a level of the factor "group" (female, male or kidz), and each level contributes a certain fixed amount of variance – when we add these together and subtract the correction factor, you get the "between-group sum of squares".

```
between.ss = (sum(females)^2)/4 + (sum(males)^2)/4
    + (sum(kidz)^2)/4 - CF
```

The variance breakdown shows how we can calculate "within.ss", namely by subtracting "between.ss" from "total.ss".

```
within.ss = total.ss - between.ss
```

Finally, to finish our calculation of the variance, we need to calculate the denominators for all the variances; the denominators will be the respective degrees of freedom. These can be broken down the same way the variance can be broken down:



The total degrees of freedom is simply the number of *all* data points (=*length(bigvector)*) minus 1.

```
df.total = length(bigvector)-1
```

This *df* value should be 11, because we have data from 12 subjects in total. The degrees of freedom for the between-groups variance is the number of columns (= the number of groups) minus one.

```
df.between = ncol(pitchstudy)-1
```

Finally, we can arrive at the within-group *df* by subtraction (have another look at the *df* breakdown): the within-groups degrees of freedom is the total degrees of freedom minus the between-groups degrees of freedom.

```
df.within = df.total-df.between
```

This is 9, the result of subtracting 2 (between *df*) from 11 (total *df*). However, there's another way to think about the within-groups degrees of freedom. In the female group, you have 4 participants, so the degrees of freedom for the female group is 3. Likewise, it is 3 for males and 3 for kidz. If you add this up, you get 9, the within-groups degrees of freedom.

Now, we can actually finish our formula to arrive at the variance estimate. This is often called "mean squares" (= the sum of squares divided by the respective degrees of freedom).

```
between.ms = between.ss/df.between
within.ms = within.ss/df.within
```

Finally, the F-value is the ratio of the between-group mean squares and the within-group mean squares.

```
F.value = between.ms/within.ms
```

Now we're almost finished. The last thing to do is to see how unlikely the F-value that we obtained is given the degrees of freedom that we have. This is actually the only part that is actually "inferential statistics" and not just descriptive statistics. What we're doing here is looking at the theoretical F distribution for the degrees of freedom 2 (the between-group *df*) and 9 (the within-group *df*)… and then we're trying to locate our F value on this distribution.

```
1-pf(F.value,2,9)
```

The p-value that you get *should* be well below 0.05 … I emphasize *should* because there is a probability (although a very low one) that you will get a non-significant result: remember, we generated a set of random numbers and even with very different means, the variation that we simulated allow for a dataset that has no statistically detectable difference between the conditions. But, given the big differences in means and the relatively small standard deviations that we specified, I suspect that this won't happen to you.

Now, the same thing, the "R way"

Auch hier wird etwas 'von Hand' getan um die Daten im richtigen Format zu generieren

The way we calculated the one-way independent samples ANOVA above, we would never do it in actual research practice. It's good to do it once for educational purposes because it shows you the inner workings of the ANOVA, but for practical reasons you would probably take one of the prefab ANOVA functions that are already provided with R.

However, to be able to use one of these functions, we need to have a table with a different structure. In the new table, each row will represent data from one subject, so each row actually represents an independent data point. This is how the end result will look like:

	subjects	groups	bigvector
1	1	female	213.6684
2	2	female	204.4267
3	3	female	184.4746
4	4	female	194.7927
5	5	male	107.2392
6	6	male	148.6768
7	7	male	107.7622
8	8	male	95.1473
9	9	kidz	381.5930
10	10	kidz	438.9887
11	11	kidz	359.4772
12	12	kidz	378.7686

The first step is to make a vector "groups" where we have labels for each group.

```
groups =
c(rep("female",4),rep("male",4),rep("kidz",4))
```

This command repeats the character string “female” four times, then “male” four times and “kidz” four times – this is concatenated by the *c( )* operator. Now, let’s put this new vector together with the actual data into a data frame. For this, we will overwrite the old data frame that we constructed, “pitchstudy”.

```
pitchstudy = data.frame(c(1:12),groups,bigvector)
```

With the command *c(1:12)* we added another column that contains the numbers 1 to 12 – let this represent IDs for our twelve subjects. Let’s rename the columns, so that the data frame looks more beautiful:

```
colnames(pitchstudy) =
c("subjects","groups","bigvector")
```

Now, if you display the data frame “pitchstudy”, you should pretty much see the same thing as in the table above. To calculate an ANOVA, simply use the *aov( )* command in conjunction with the *summary( )* command:

```
summary(aov(bigvector ~ groups + Error(subjects),
    data=pitchstudy))
```

The first argument of the *aov( )* is a formula that reads as the following: “try to predict *bigvector* with the fixed effect groups and the random factor *subjects*”. The “Error(subjects)” part is the part that tells the *aov( )* function that it should expect random variation from each participant.

However, if you compare the results of the *aov( )* output to the F value and the p-value above, you see that the results are not the same!! What happened?

Have a look at the data frame:

```
summary(pitchstudy)
```

R tries to calculate the “mean and median” of the subject column – this means that it treats this column as numeric. You can check how R treats this vector with the *class( )* command.

```
class(pitchstudy$subjects)
```

Das ist notwendig, um R zu sagen, dass die 'subjects'-Nummern Namen sind und keine Zahlen.

It should say “integer”. So you need to recode it as “factor”:

```
pitchstudy$subjects= as.factor(pitchstudy$subjects)
```

Now, rerun the *aov( )* command above, and this time, you should get the same results as in your hand computation.

This is how you report the results:

"We performed a one-way independent samples ANOVA with the fixed factor Group (three levels: females, males, children) and the random factor Subjects. There was a significant effect of Group ($F(2,9)= 172.98$, $p<0.0001$)."

Hopefully by now, you will have a much better understanding of what "($F(2,9)= 172.98$, $p<0.0001$)" actually means. The first degrees of freedom are the degrees of freedom that are associated with the numerator of the F ratio (between *df*, or effect *df*), the second degrees of freedom are associated with the denominator (within *df*, "residual" or error *df*). Finally, the p-value gives you the probability of observing the F value that you obtained – given the degrees of freedom that determine the shape of the theoretical F distribution on which you try to locate your F value.

Interpreting the result

What we've done so far only tells you that the factor "Group" plays a role. You can't make a claim yet about which groups differ with respect to each other, you can only say that there is an overall difference between groups. To assess where exactly the difference lies (e.g. between females and males or between males and children), you would need to do pair wise comparisons using, e.g. t-tests.

So, you might wonder: Hey, if I can use t-tests to look at the difference between pairs (females-males, females-children, males-children) – why didn't we do this to begin with? The reason is that with each statistical test that you perform, there is a probability that you find a statistically significant effect that is due to chance – this is expected to occur in 5% of all cases. By performing many tests, you greatly increase the chances of finding *at least one* statistically significant result in your data that, in fact, is a chance result. Therefore, the ANOVA is the preferred approach to the data above, because the so-called "study-wide error rate" is controlled for: by adopting a 5% significance level, there is an overall 5% chance of finding a significant effect of the factor Group that is actually spurious. And this 5% of uncertainty is (by convention) an accepted degree of uncertainty.

If you are interested in differences between specific groups (= specific levels of the factor "Group"), then, you would have to perform post-hoc comparisons and delve into the issue of corrections for multiple comparisons. This is quite a big topic in and of itself and will therefore not be covered in this text.

In any way, I hope that this tutorial was successful in making you familiar with the F ratio, the F distribution, degrees of freedom and ANOVA.

Acknowledgements

I thank Matthias Urban for proof-reading the manuscript. I thank Benjamin Bergen, Amy Schafer and Roger Mundry for teaching me almost everything I know about statistics.

References

Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54:2, 187-211.