Likelihood

Author: Cole Brookson **Date:** 29 August 2022

In our discussion of comparing sample means we discussed how to use R to compare two sample means. This is usually done through a t-test of some form. However, it's not uncommon that we start asking questions wherein we want to compare the sample means across a larger number of groups. For example, say we have four groups 1-4.

This brings us to a conundrum, as statistical theory has to re-consider what to do to be able to formulate our null and alternate hypotheses in such a way that is useful. If we want to compare the means of four groups of a single grouping variable, we will use a single-factor ANOVA to do so. Our task here is to figure out how much variance is likely present between the group means due to sampling error, and then what amount of variance on top of that would denote a significantly different mean.

To be clear, when we are discussing this one-factor ANOVA, we are stating that our null hypothesis is that there is no difference between the sample means, so $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. In our alternative hypothesis then, it does not have to be true that more than one μ are significantly different, but only that at least one is.

F-statistic

Just like when we did t-tests and had the t-statistic, our test-statistic for the one-factor ANOVA is the F-statistic. Since our interest here is variance, the F-statistic is best thought of as a ratio of two variances. We will think the variances in the terms of mean squares. We first have the group mean square which we calculate as the error sum of squares and dividing by the degrees of freedom:

$$MSG = \frac{\sum (y_i - \hat{y_i})^2}{n-2}$$

and then we have the mean square error which we calculate by summing the error sum of squares and dividing by the associated degrees of freedom:

$$MSE = \frac{\sum (y_i - \hat{y_i})^2}{n-2}$$

. These two quantities can be thought of as representing the variance among group means (MSG) and the variance between subjects in the same group (MSE). Thus, it may be somewhat logical then, that since the F-statistic is the simple ratio

$$F = \frac{MSG}{MSE}$$

, if the ratio is equal to 1, then the variances are in fact the same, and there is no additional variance between the sample means (represented by MSG) compared to just the error within groups (represented by MSE).

We then use our p-value as usual to determine the probability of getting our present F-statistic by chance. Then we can reject or fail to reject the null hypothesis.

Performing the ANOVA

When performing the ANOVA, similar to all statistical tests there are some processes that need to take place before we actually run the test. We first gather and organize our data as need be. Then we must check that our data meet the assumptions of the test we are planning on carrying out, then we can perform our test and interpret.

For our example here, let's consider the sizes of penguins bills.

In our example here, we may hypothesize that between the species in our dataset, there may be differences in the mean bill length. Bill length in a penguin looks vaguely like this:

So first we'll look at the species on hand

It's probably best to check that each species has ~roughly the same amount of samples per group before we continue. Let's use the table() function for this:

We can also plot this:

Okay so we can see that while they're not completely equal, they're close enough for our purposes, especially with an ANOVA, as it's one of the less sensitive tests to number of samples per group.

Assumptions

The assumptions of this one-factor ANOVA are:

1) Observations are independent and random
2) Data in each level of the groupings are normally distributed
3) The populations have common variances

So we'll start by assuming independence and random sampling. Next, we need to check if each of the groups have data that are roughly normally distributed.

Normality A note: while plotting our data can be helpful, it can often lead to p-hacking and HARKing. While it is technically acceptable to plot the density plots for each group, that will give us an insight, prior to performing our tests, about what the means of each group might be. Again, this in and of itself is not bad but we can get the same information without the possible pitfalls by using QQ plots and significance tests, which at the end of the day are often required for publications/reports anyways. Here we'll use QQ plots and a Shapiro-Wilks test to check for normality in each of our groups:

Visually, all of these look reasonably acceptable, but let's compare with a Shapiro-Wilks test on each:

Well, it's a good thing we checked with our tests!! Our Shapiro-Wilks test for our Gentoo species shows that actually the data are non-normal. We note that the p-value is not especially large, but still <0.05.

Two things to note here: First, ANOVAs are actually relatively robust to violations of normality assumptions. This means that given we can see the data are not that far off of a normal distribution (we can see this via our QQ plot), we could in theory proceed with our test here with relative confidence. Second, log-transforming our data is completely valid for an ANOVA. This is helpful for us, as while we could go ahead with the test (after checking our other assumptions), it's still true that there is an easy transformation available to us through the log transformation, that prevents us from having to technically violate the assumptions of our test. An additional note is that if our assumptions are not met completely, and perhaps a transformation does not help us, it is possible to use a non-parametric alternative such as the Kruskal-Wallis test which we do not cover in this section, but is similarly easy to run.

So let's go ahead and transform our data:

Now we'll go ahead and re-do our QQ plots and our tests:

So these all look pretty good. Time for the tests:

Ok great, so we can now see that each of the three tests has a p-value > 0.05, so we can assume normality for these values.

Homoskedasticity Next to check our variances.

Our second assumption was that the variances of all samples be equal. If they are not (heteroskedasticity), all is not lost and we can use a different version of the t-test. We will test our homoskedascitiy with the Bartlett's test. Again, similar to the normality test, the null hypothesis of the F-test is that the variances are equal, so if $p < 0.05$ then for our purposes we could say that the assumption of homoskedascitiy is violated. As a sidenote, if our assumptions of normality were not quite met, it would be preferable to use the Levene's test which is less sensitive to this problem.

To do the Barlett's test let's use the stats package function bartlett.test(). We will now actually re-format our data first so that it's in long form. This will make our lives easier. To do this, we want one column with the actual values, and another with the grouping variable (in this case, our species names). To get the species names, what we'll do is combine a few arguments such that we can do it all in one command. We'll use the rep() function which takes two arguments: a) the value to be repeated, and b) the number of times to repeat it. Because there are different numbers of observations for each species, the second argument in each call to rep() will simply be the length of the vector of the log_length for that species. Also, we want the grouping variable to be a factor. That looks like this:

Great, now for our bartlett's test. The first argument is the column that has the values of interest, and we use the tilde (~) to denote after the column of interest, the name of the grouping column. Last, tell the function what the dataframe itself is called.

Wonderful, and we see that the p-value is well above 0.05, so our variances are equal enough to continue.

The test

Now to perform the test itself. It is, similarly to the t-test, very easy to compute in R, and looks pretty much identical to our syntax for the Bartlett's test. The first argument is the column that has the values of interest, and we use the tilde (~) to denote after the column of interest, the name of the grouping column. The function itself we use comes from the stats package again. We can now proceed to our test and run it all like this:

So we see our p-value is very small, which indicates we can reject the null hypothesis - there is a difference between the means of the species.

Interpretation

There are still a few steps we might want to take. First, we still don't know what the actual means of the groups are, nor which are different than each other. Also, it's almost always advantageous to plot our output so let's do that now.

Tukey Comparison To actually see what the pairwise comparison of the groups are, we can see which groups are actually different than one another. This is useful as it allows us to identify if it's only one group that is different, (i.e. $\mu_1 = \mu_2$ but $\mu_3 > \mu_2$) or it they're all quite a bit different. We can do this comparison through the Tukey test (Tukey Honest Signfiicant Differences) which again can be done using the stats package:

So here we get four key outputs: diff which is the difference between the two groups at hand (this is read as the chinstraps are ~0.23 greater than adelies for the first row), the upper and lower bounds of the 95% confidence interval, and the p-value denoting if said difference is significant.

Plot the result Now let's plot our results. This can be done in any number of ways, but I am personally keen on plotting the density of the values, in the same way we would if we wanted to use it to check assumptions of normality as it shows all components of the data at hand and doesn't have the chance of hiding anything. I also think it's just a useful plot to have. We'll pair that with denotations of the means of each group as well as annotating the p-value onto the plot like this: