23. Analysis of variance

An ANalysis Of VAriance (ANOVA) is, as the name implies, a method for analysing variances in a dataset. This is confusing, at first, because the most common application of an ANOVA is to test for differences among group *means*. That is, an ANOVA can be used to test the same null hypothesis as the independent samples student's t-test introduced Chapter 21.2; are 2 groups sampled from a population that has the same mean? The t-test works fine when we have only 2 groups, but it does not work when there are 3 or more groups and we want two know if the groups all have the same mean. An ANOVA can be used to test the null hypothesis that *all* groups in a dataset are sampled from a population with the same mean. For example, we might want to know if mean wing length is the same for 5 species of fig wasps sampled from the same area (Duthie et al., 2015). What follows is an explanation of why this can be done by looking at the variance within and between groups (note, 'groups' are also sometimes called 'factors' or 'levels'). Groups are categorical data (see Chapter 5). In the case of the fig wasp example, the groups are the different species (Table 23.1).

Table 23.1.: Wing lengths (mm) measured for 5 unnamed species of non-pollinating fig wasps collected from fig trees in 2010 near La Paz in Baja, Mexico. Note, for readability, this table is not presented in a tidy format.

Het1	Het2	LO1	SO1	SO2
2.122	1.810	1.869	1.557	1.635
1.938	1.821	1.957	1.493	1.700
1.765	1.653	1.589	1.470	1.407
1.700	1.547	1.430	1.541	1.378

Why is any of this necessary? If we want to know if the 5 species of fig wasps in Table 23.1 have the same mean wing length, can we not just use t-tests to compare the means between each species? There are a couple problems with this approach. First, there are a lot of group combinations to compare (Het1 vs Het2, Het1 vs LO1, Het1 vs SO1, etc.). For the 5 fig wasp species in Table 21.2, there are 10 pair-wise combinations that would need to be tested. And the number of combinations grows exponentially[1] with each new group added to the dataset (Table 23.2)

[1]Technically polynomially, but the distinction really is not important for understanding the concept.

Table 23.2.: The number of individual t-tests that would need to be run to compare the means given different numbers of groups (e.g., if a dataset had measurements from 2-10 species)

Groups	2	3	4	5	6	7	8	9	10
Tests	1	3	6	10	15	21	28	36	45

Aside from the tedium of testing every possible combination of group means, there is a more serious problem having to do with the Type I error. Recall from Chapter 20.3 that a Type I error occurs when we reject the null hypothesis (H_0) and erroneously conclude that H_0 is false when it is actually true (i.e., a false positive). If we reject H_0 at a threshold level of $\alpha = 0.05$ (i.e., reject H_0 when $P < 0.05$, as usual), then we will erroneously reject the null hypothesis about 5% of the time that we run a statistical test and H_0 is true. But if we run 10 separate t-tests to see if the fig wasp species in Table 23.1 have different mean wing lengths, then the probability of making an error increases considerably. The probability of erroneously rejecting **at least 1** of the 10 null hypotheses increases from 0.05 to about 0.40. In other words, about 40% of the time, we would conclude that at least 2 species differ in their mean wing lengths[2], even when all species *really do* have the same wing length. This is not a mistake that we want to make, which is why we should first test if all of the means are equal:

- H_0 : All mean species wing lengths are the same
- H_A : Mean species wing lengths are not all the same

We can use an ANOVA to test the null hypothesis above against the alternative hypothesis. If we reject H_0, then we can start comparing pairs of group means (more on this in Chapter 25).

How do we test the above H_0 by looking at *variances* instead of *means*? Before getting into the details of how an ANOVA works, we will first look at the F-distribution. This is relevant because the test statistic calculated in an ANOVA is called an F-statistic, which is compared to an F-distribution in the same way that a t-statistic is compared to a t-distribution for a t-test (see Chapter 21).

In general, the number of possible comparisons between groups is described by a binomial coefficient,

$$\binom{g}{2} = \frac{g!}{2\,(g-2)!}.$$

The number of combinations therefore increases with increasing group number (g).

[2]To get the 0.4, we can first calculate the probability that we (correctly) do not reject H_0 for all 10 pair-wise species combinations $(1-0.05)^{10} \approx 0.60$, then subtract from 1, $P(Do\ not\ reject\ H_0) = 1-(1-0.05)^{10} \approx 0.4$. That is, we find the probability of there not being a Type I error in the first test (1 - 0.05), **and** the second test (1 - 0.05), and so forth, thereby multiplying (1 - 0.05) by itself 10 times. This gives the probability of not committing any Type I error across all 10 tests, so the probability that we commit at least 1 Type I error is 1 minus this probability.

23.1. The F-distribution

If we want to test whether or not 2 variances are the same, then we need to know what the null distribution should be if 2 different samples came from a population with the same variance. The general idea is the same as it was for the distributions introduced in Chapter 14.4. For example, if we wanted to test whether or not a coin is fair, then we could flip it 10 times and compare the number of times it comes up heads to probabilities predicted by the binomial distribution when $P(Heads) = 0.5$ and $N = 10$ (see Chapter 14.4.1 Figure 14.5). To test variances, we will calculate the ratio of variances (F), then compare it to the F probability density function[3]. For example, the ratio of the variances for samples 1 and 2 is (Sokal and Rohlf, 1995),

$$F = \frac{Variance\,1}{Variance\,2}.$$

Note that if the variances of samples 1 and 2 are the exact same, then F = 1. If the variances are very different, then F is either very low (if Variance 1 < Variance 2) or very high (if Variance 1 > Variance 2). To test the null hypothesis that samples 1 and 2 have the same variance, we therefore need to map the calculated F to the probability density of the F distribution. Again, the general idea is the same as comparing a t-score to the t-distribution in Chapter 21.1. Recall that the shape of the t-distribution is slightly different for different degrees of freedom (df). As df increases, the t-distribution starts to resemble the normal distribution. For the F-distribution, there are actually 2 degrees of freedom to consider. One degree of freedom is needed for Variance 1, and a second degree of freedom is needed for Variance 2. Together, these 2 degrees of freedom will determine the shape of the F-distribution (Figure 23.1).

Figure 23.1 shows an F distribution for 3 different combinations of degrees of freedom. The F distribution changes its shape considerably given different df values. Visualising this is much, much easier using an interactive application.

> Click here for an interactive application demonstrating how the F distribution changes with different degrees of freedom.

It is not necessary to memorise how the F distribution changes with different degrees of freedom. The point is that the probability distribution changes given different degrees of freedom, and that the relationship between probability and the value on the x-axis (F) works like other distributions such as the normal or t-distribution. The entire area under the curve must sum to 1, and we can calculate the area above and below any F value (rather, we can get statistical programs such as Jamovi and R to do this for us). Consequently, we can use the F-distribution as the null distribution for the ratio of two

[3]The F-distribution was originally discovered in the context of the ratio of random variables with chi-square distributions, with each variable being divided by its own degree of freedom (Miller and Miller, 2004). We will look at the Chi-square distribution in Week 9.

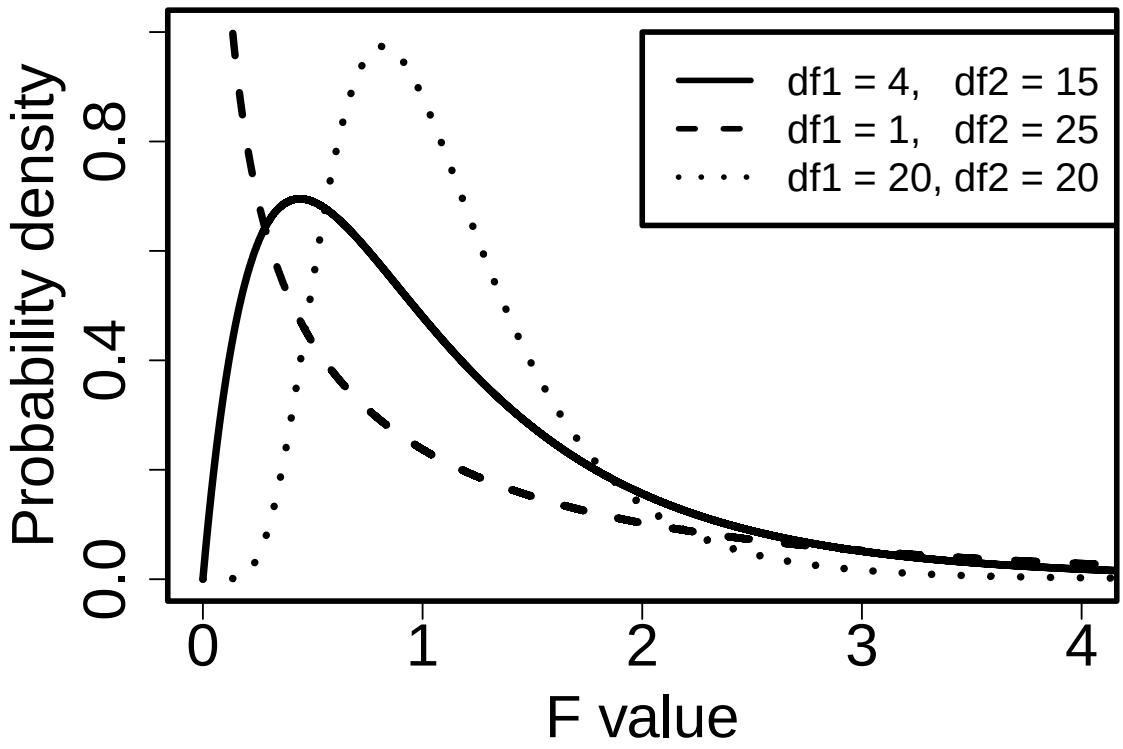

Figure 23.1.: Probability density functions for 3 different F distributions, each of which have different degrees of freedom for the variances in the numerator (df1) and denominator (df2).

variances. If the null hypothesis that the 2 variances are the same is true (i.e., F = 1), then the F-distribution gives us the probability of the ratio of 2 variances being as or more extreme (i.e., further from 1) than a specific value.

23.2. One-way ANOVA

We can use the F-distribution to test the null hypothesis mentioned at the beginning of the chapter (that fig wasp species have the same mean wing length). The general idea is to compare the mean variance among groups to the mean variance within groups, so our F value (i.e., "F statistic") is calculated,

$$F = \frac{Mean\ variance\ among\ \ groups}{Mean\ variance\ within\ \ groups}.$$

The rest of this section works through the details of how to calculate this F statistic. It is easy to get lost in these details, but the calculations that follow do not need to be done by hand. As usual, Jamovi or R will do all of this work for us (The Jamovi Project, 2022; R Core Team, 2022). The reason for going through the ANOVA step-by-step is to show how the total variation in the dataset is being partitioned into the variance among versus within groups, and to provide some conceptual understanding of what the numbers in ANOVA output actually mean.

23.2.1. ANOVA mean variance among groups

To get the mean variance among groups (i.e., mean squares; MS_{among}), we need to use the sum of squares (SS). The SS was introduced to show how the variance is calculated in Chapter 12.3,

$$SS = \sum_{i=1}^{N} (x_i - \bar{x})^2.$$

Instead of dividing SS by N - 1 (i.e., the total df), as we would do to get a sample variance, we will need to divide it by the df *among groups* (df_{among}) and df within groups (df_{within}). We can then use these SS_{among}/df_{among} and SS_{within}/df_{within} values to calculate our F[4].

This all sounds a bit abstract at first, so an example will be helpful. We can again consider the wing length measurements from the 5 species of fig wasps shown in Table

[4]Note that the SS divided by the degrees of freedom (N - 1) is a variance. For technical reasons (Sokal and Rohlf, 1995), we cannot simply calculate the mean variance of groups (i.e., the mean of s^2_{Het1}, s^2_{Het2}, etc.). We need to sum up all the squared deviations from group means *before* dividing by the relevant degrees of freedom (i.e., dfs for the among and within group variation).

23.1. First, note that the **grand mean** (i.e., the mean across all species) is $\bar{\bar{x}} = 1.6691$. We can also get the sample mean values of each group, individually. For example, for Het1,

$$\bar{x}_{Het1} = \frac{2.122 + 1.938 + 1.765 + 1.7}{4} = 1.88125$$

We can calculate the means for all 5 fig wasps (Table 23.3).

Table 23.3.: Mean wing lengths (mm) from 5 unnamed species of non-pollinating fig wasps collected from fig trees in 2010 near La Paz in Baja, Mexico. Each species mean was calculated from 4 wasps (N = 4).

Het1	Het2	LO1	SO1	SO2
1.88125	1.70775	1.71125	1.51525	1.53

To get the mean variance among groups, we need to calculate the sum of the squared deviations of each species wing length ($\bar{x}_{Het1} = 1.88125$, $\bar{x}_{Het2} = 1.70775$, etc.) from the grand mean ($\bar{\bar{x}} = 1.6691$). We also need to weigh the squared deviation of each species by the number of samples for each species[5]. For example, for Het1, the squared deviation would be $4(1.88125 - 1.6691)^2$ because there are 4 fig wasps, so we multiply the squared deviation from the mean by 4. We can then calculate the sum of squared deviations of the species means from the grand mean,

$$SS_{among} = 4(1.88125 - 1.6691)^2 + 4(1.70775 - 1.6691)^2 \; + \; ... \; + \; 4(1.53 - 1.6691)^2.$$

Calculating the above across the 5 species of wasps gives a value of $SS_{among} = 0.3651868$. To get our mean variance among groups, we now just need to divide by the appropriate degrees of freedom (df_{among}). Because there are 5 total species ($N_{species} = 5$), $df_{among} = 5 - 1 = 4$. The mean variance among groups is therefore $MS_{among} = 0.3651868/4 = 0.0912967$.

23.2.2. ANOVA mean variance within groups

To get the mean variance within groups (MS_{within}), we need to calculate the sum of squared deviations of wing lengths from *species means*. That is, we need to take the wing length of each wasp, subtract the mean species wing length, then square it. For example, for Het1, we calculate,

[5] In this case, weighing by sample size is not so important because each species has the same number of samples. But when different groups have different numbers of samples, we need to multiply by sample number so that each group contributes proportionally to the SS.

$SS_{Het1} = (2.122 - 1.88125)^2 + (1.938 - 1.88125)^2 + (1.765 - 1.88125)^2 + (1.7 - 1.88125)^2$.

If we subtract the mean and square each term of the above,

$$SS_{Het1} = 0.0579606 + 0.0032206 + 0.0135141 + 0.0328516 = 0.1075467.$$

Table 23.4 shows what happens after taking the wing lengths from Table 22.1, subtracting the means, then squaring.

Table 23.4.: The squared deviations from species means for each wing length presented in Table 23.1

Het1	Het2	LO1	SO1	SO2
0.0579606	0.0104551	0.0248851	0.0017431	0.011025
0.0032206	0.0128256	0.0603931	0.0004951	0.028900
0.0135141	0.0029976	0.0149451	0.0020476	0.015129
0.0328516	0.0258406	0.0791016	0.0006631	0.023104

If we sum each column (i.e., do what we did for SS_{Het1} for each species), then we get the SS for each species (Table 23.5).

Table 23.5.: The sum of squared deviations from species means for each wing length presented in Table 23.1

Het1	Het2	LO1	SO1	SO2
0.1075467	0.0521188	0.1793248	0.0049487	0.078158

If we sum the squared deviations in Table 23.5, we get a $SS_{within} = 0.422097$. Note that each species included 4 wing lengths. We lose a degree of freedom for each of the 5 species (because we had to calculate the species mean), so our total df is 3 for each species, and $5 \times 3 = 15$ degrees of freedom within groups (df_{within}). To get the mean variance within groups (denominator of F), we calculate $MS_{within} = SS_{within}/df_{within} = 0.0281398$.

23.2.3. ANOVA F statistic calculation

From Chapter 23.2.1, we have the mean variance among groups,

$$MS_{among} = 0.0912967.$$

From Chapter 23.2.2, we have the mean variance within groups,

$$MS_{within} = 0.0281398$$

To calculate F, we just need to divide MS_{among} by MS_{within},

$$F = \frac{0.0912967}{0.0281398} = 3.2443976.$$

Remember that if the mean variance among groups is the same as the mean variance within groups (i.e., $MS_{among} = MS_{within}$), then F = 1. We can test the null hypothesis that $MS_{among} = MS_{within}$ against the alternative hypothesis that there is more variation among groups than within groups ($H_A : MS_{among} > MS_{within}$) using the F distribution (note that this is a one-tailed test). In the example of 5 fig wasp species, $df_{among} = 4$ and $df_{within} = 15$, so we need an F distribution with 4 degrees of freedom in the numerator and 15 degrees of freedom in the denominator[6]. We can use the interactive app to get the F-distribution and p-value (Figure 23.2).

The area shaded in grey in Figure 23.2, where F > 3.2443976, is approximately $P = 0.041762$. This is our p-value. Since $P < 0.05$, we can reject the null hypothesis that all mean species wing lengths are the same because the variance among species wing lengths is significantly higher than the variance within species wing lengths. Note that the critical value of F (i.e., for which $P = 0.05$) is 3.0555683, so for any F value above this (for df1 = 5 and df2 = 19), we would reject H_0.

When running an ANOVA in a statistical program, output includes (at least) the calculated F statistic, degrees of freedom, and the p-value. Figure 23.3 shows the one-way ANOVA output of the test of fig wasp wing lengths.

Jamovi is quite minimalist for a one-way ANOVA (The Jamovi Project, 2022), but these 4 statistics (F, df1, df2, and p) are all that is really needed. Most statistical programs will show ANOVA output that includes the SS and mean squares among (MS_{among}) and within (MS_{within}) groups.

```
## Analysis of Variance Table
##
## Response: wing_length
##           Df  Sum Sq  Mean Sq F value  Pr(>F)
## Species    4 0.36519 0.091297  3.2444 0.04176 *
## Residuals 15 0.42210 0.028140
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[6]Note that $df_{among} = 4$ and $df_{within} = 15$ sum to 19, which is the total df of the entire dataset ($N - 1 = 20 - 1 = 19$). This is always the case for the ANOVA; the overall df constrains the degrees of freedom among and within groups.

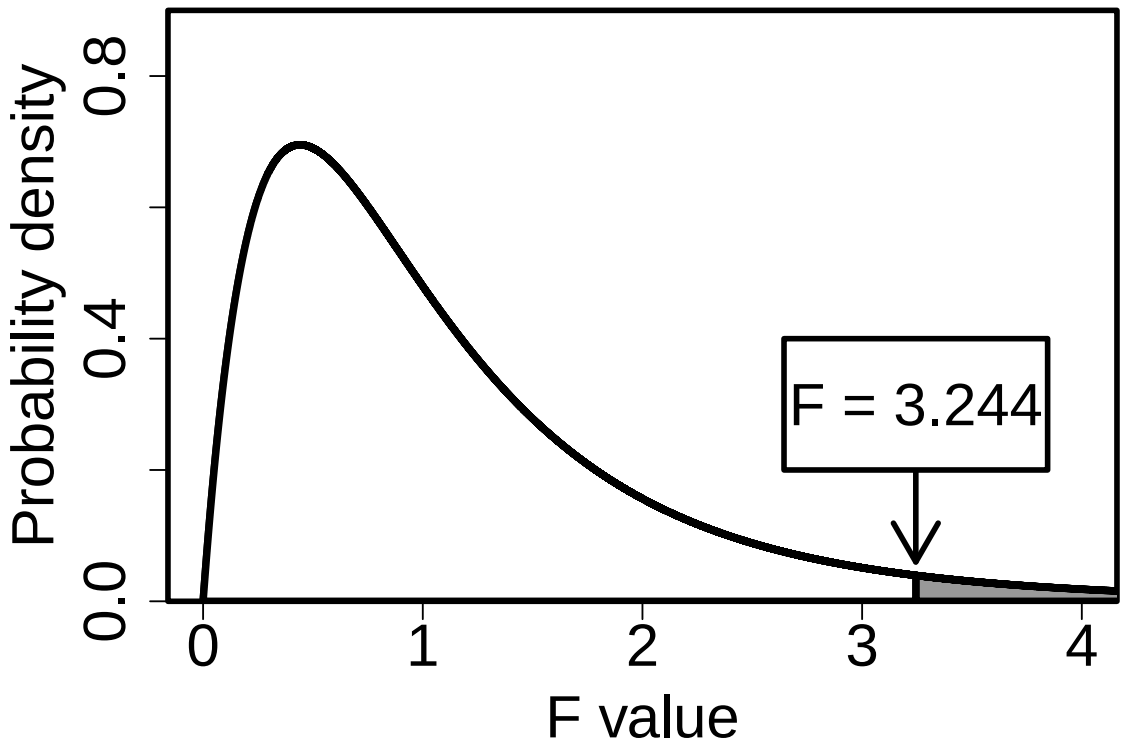

Figure 23.2.: F distribution with df = 4 for the numerator and df = 15 for the denominator. The arrow indicates an F value calculated from fig wasp species wing length measurements for 5 different species and 4 measurements per species. Fig wasp wing lengths were collected from a site near La Paz in Baja, Mexico 2010.

One-Way ANOVA (Fisher's)

	F	df1	df2	p
wing_length	3.24440	4	15	0.042

Figure 23.3.: Jamovi output for a one-way ANOVA of wing length measurements in 5 species of fig wasps collected in 2010 near La Paz in Baja, Mexico.

The above output, taken from R, includes the same information as Jamovi (F, df1, df2, and p), but also includes SS and mean variances. Note that we can also get this information from Jamovi if we want it (see Chapter 26).

23.3. Assumptions of ANOVA

As with the t-test (see Chapter 21.4), there are some important assumptions that we make when using an ANOVA. Violating these assumptions will mean that our Type I error rate (α) is, again, potentially misleading. Assumptions of ANOVA include the following (Box et al., 1978; Sokal and Rohlf, 1995):

1. Observations are sampled randomly
2. Observations are independent of one another
3. Groups have the same variance
4. Errors are normally distributed

Assumption 1 just means that observations are not biased in any particular way. For example, if the fig wasps introduced at the start of this chapter were used because they were the largest wasps that were collected for each species, then this would violate the assumption that the wing lengths were sampled randomly from the population.

Assumption 2 means that observations are not related to one another in some confounding way. For example, if all of the Het1 wasps came from one fig tree, and all of the Het2 wasps came from a different fig tree, then wing length measurements are not really independent within species. In this case, we could not attribute differences in mean wing length to species. The differences could instead be attributable to wasps being sampled from different trees (more on this in Chapter 26).

Assumption 3 is fairly self-explanatory. The ANOVA assumes that all of the groups in the dataset (e.g., species in the case of the fig wasp wing measurements) have the same variance. That is, we assume homogeneity of variances (as opposed to heterogeneity of variances). In general, ANOVA is reasonably robust to deviations from homogeneity, especially if groups have similar sample sizes (Blanca et al., 2018). This means that the Type I error rate is about what we want it to be (usually $\alpha = 0.05$), even when the assumption of homogeneity of variances is violated. In other words, we are not rejecting the null hypothesis when it is true more frequently than we intend! We can test the assumption that group variances are the same using a Levene's test in the same way that we did for the independent samples t-test in Chapter 22. If we reject the null hypothesis that groups have the same variance, then we should potentially consider a non-parametric alternative test such as the Kruskall-Wallis H test (see Chapter 25).

Assumption 4 is the equivalent of the t-test assumption from Chapter 21.4 that sample means are normally distributed around the true mean. What the assumption means is that if we were to repeatedly resample data from a population, the sample means that

we calculate would be normally distributed. For the fig wasp wing measurements, this means that if we were to go back out and repeatedly collect 4 fig wasps from each of the 5 species, then sample means of species wing length and overall wing length would be normally distributed around the true means. Due to the central limit theorem (see Chapter 15), this becomes less problematic with increasing sample size. We can test if the sample data are normally distributed using a Q-Q plot or Shapiro-Wilk test (the same procedure used for the t-test). Fortunately, the ANOVA is quite robust to deviations from normality (Schmider et al., 2010), but if data are not normally distributed, we should again consider a non-parametric alternative test such as the Kruskall-Wallis H test (see Chapter 25).