ANOVA

6

Contents

What You Will Learn in This Chapter

In Chap. 3, we examined how to compare the means of two groups. In this chapter, we will examine how to compare means of more than two groups.

6.1 One-Way Independent Measures ANOVA

Suppose we wish to examine the effects of geographic region on tree heights. We might sample trees from near the equator, from the 49th parallel, and from the 60th parallel. We want to know whether the mean tree heights from all three regions are the same (Fig. 6.1). Since we have three regions we cannot use the *t*-test because the *t*-test only works if we are comparing two groups.

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_6

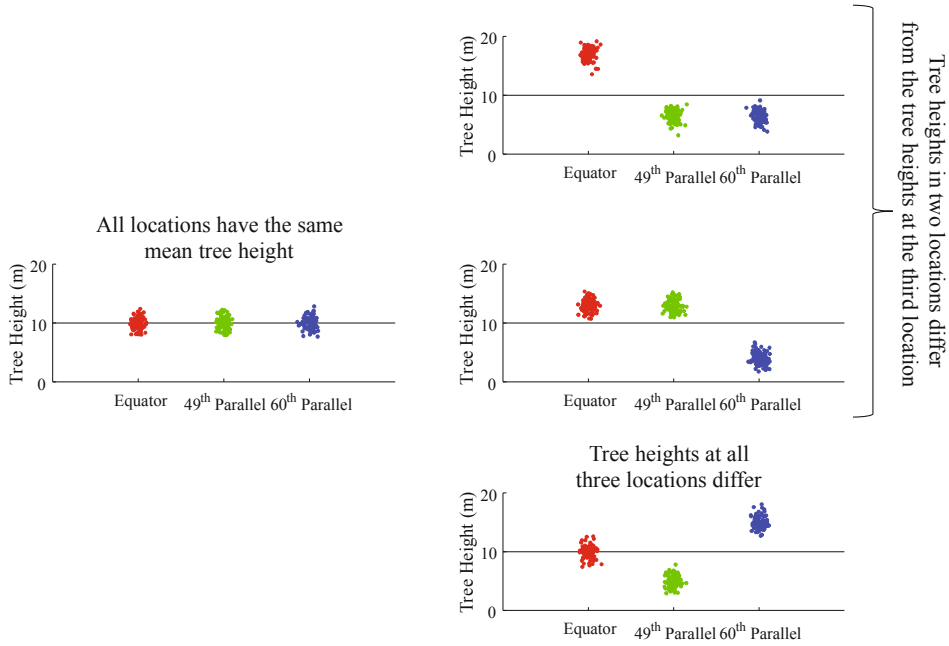


Fig. 6.1 We will examine the tree heights at three different latitudes. Left: the mean heights at all three latitudes is the same, shown by the horizontal line. Right: At least one latitude has a mean height that differs from the other two. The horizontal line shows the mean tree height for all three groups, called the " Grand Mean"

In principle we could compute three t-tests to compare all possible pairs of means (equator vs 49, equator vs 60, and 49 vs 60). However in this case, as shown in Chap. 5, we would face the multiple testing problem with the unpleasant side effect of increasing our Type I error rate as the number of comparisons increases. Situations like this are a case for an analysis of variance (ANOVA), which uses a clever trick to avoid the multiple testing problem.

6.2 Logic of the ANOVA

Terms

There are many terms for a 1-way ANOVA with m-groups:

- way = factor
- group = treatment = level

The logic of the ANOVA is simple. We simplify our alternative hypothesis by asking whether or not at least one of the tree populations is larger than the others. Hence, we are stating *one* hypothesis instead of three by lumping all alternative hypotheses together:

Null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Lumped alternative hypotheses

$$H_1 : \mu_1 = \mu_2 \neq \mu_3$$
$$H_1 : \mu_1 \neq \mu_2 = \mu_3$$
$$H_1 : \mu_1 \neq \mu_3 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$

The ANOVA assumes, similarly to the t-test, that all groups have the same population variances σ. If the null hypothesis is true, then the *population* means are equal for all three groups of trees. Any observed differences in the *sample* means come from the variance σ alone, which is due to random differences in tree heights (noise), but not to systematic differences in tree heights with geographic region (Fig. 6.1). It turns out that when the null hypothesis is true, the variability between means can be used to estimate σ (by multiplying by the sample sizes). An ANOVA compares this between means estimate to a direct estimate that is computed within each group.

Now assume that the mean tree heights in the three geographic regions are in fact different. In this case, the individual tree heights depend on both the variance within a group σ and the variability between the group means. In this case, the estimate of σ based on the variability between means tends to be larger than σ. In contrast, the estimate of σ based on the variability within each group tends to be similar to the true value. The ANOVA divides the two estimated variances and obtains an F-value:

$$F = \frac{\text{Variance estimate based on variability between group means}}{\text{Variance estimate based on variability within groups}}$$

Formally, this equation is expressed as:

$$F = \frac{\frac{\sum_{j=1}^{k} n_j (M_j - M_G)^2}{k-1}}{\sum_{j=1}^{k} \frac{\sum_{i=1}^{n_j} (x_{ij} - M_j)^2}{n_j - 1}}$$

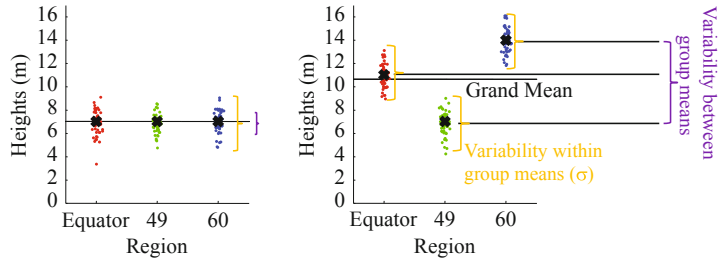


Fig. 6.2 Logic of the ANOVA. Left: the null hypothesis is true and, hence, all population means are identical. In this case, all the variability is within-group variability, which we consider to be the noise. The ANOVA assumes that variability in the data is the same for all three populations of trees. Right: the null hypothesis is false. Here, we show an extreme case where the variability of the means is larger than the variability of the data about the means. In this case, most of the variability in the data is explained by the effect of the three different regions on the tree heights. Usually the situation is somewhere in between these extremes. The null hypothesis for an ANOVA is that any observed differences are only due to the noise. The purpose of the ANOVA is to distinguish variability caused by the independent variable from variability of the data points around their individual treatment means

where k is the number of groups (three tree populations), n_j is the number of scores in group j (the number of trees within each sampled geographic region), M_j is the mean for group j (mean of geographic region sample j), M_G is the grand mean of all scores pooled together, and x_{ij} is the ith score for group j (the height of a single tree). To make it easier to distinguish the means from individual scores we use the symbols M_j and M_G rather than the traditional symbol for a sample mean $\bar{x}$. The multiplication by n_j in the numerator weights the deviations of the group means around the grand mean by the number of trees in each group so that the numbers of scores contributing to the variance estimates are equated between the numerator and denominator.

Consider two extreme examples. First, the null hypothesis is true (as in Fig. 6.2 left). In this case, the variance estimates for both the numerator and the denominator are similar and will produce an F-value that is close to 1. Next, let us consider an example of an alternative hypothesis where the differences between tree heights in the three geographic regions are large and σ is very small, i.e., the tree heights differ largely between the three populations but are almost the same within each population (as in Fig. 6.2 right). The variability in the measurements is mostly determined by the group differences and the F-value is large.

Just as in the t-test, a criterion is chosen for statistical significance to set the Type I error rate to a desired rate (e.g., $\alpha = 0.05$). When F exceeds the criterion, we conclude that there is a significant difference (i.e., we reject the null hypothesis of equality between the group means).

The tree example is a one-way ANOVA, where there is one factor (tree location) with three groups (regions) within the factor. The groups are also called levels and the factors are also called ways. There can be as many levels as you wish within a factor, e.g. many more regions, from which to sample trees. A special case is a one-way independent measures ANOVA with two levels, which compares two means as does the t-test. In fact, there is a close relationship between the two tests and in this case it holds that: $F = t^2$. The p-value here will be the same for the ANOVA and the two-tailed t-test. Hence, the ANOVA is a generalization of the t-test.

As with the t-test, the degrees of freedom play an important role in computing the p-value. For a one-way independent measures ANOVA with k levels, there are two types of degrees of freedom df_1 and df_2, respectively. In general, $df_1 = k - 1$ and $df_2 = n - k$ where n is the total number of sampled scores pooled over all groups, e.g., all trees in the three groups. The total of the degrees of freedom is $df_1 + df_2 = n - 1$.

6.3 What the ANOVA Does and Does Not Tell You: Post-Hoc Tests

Assume our ANOVA found a significant result. What does it tell us? We reject the null hypothesis that all means are equal:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

thereby accepting the alternative hypothesis, which can mean that any of the following are true:

$$H_1 : \mu_1 = \mu_2 \neq \mu_3$$
$$H_1 : \mu_1 \neq \mu_2 = \mu_3$$
$$H_1 : \mu_1 \neq \mu_3 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$$

By rejecting the H_0, we accept one of the alternative hypotheses—but we do not know which one. This is the price of avoiding multiple testing by lumping the four alternative hypotheses into one.

Here, the ANOVA offers a second trick. If we rejected the null hypothesis, it is appropriate to compare pairs of means with what are called "post-hoc tests," which, roughly speaking, corresponds to computing pairwise comparisons. Contrary to the multiple testing situations discussed in Chap. 5, these multiple comparisons do not inflate the Type I error rate because they are only conducted if the ANOVA finds a main effect.

There are many post-hoc tests in the statistical literature. Commonly used post-hoc tests include: Scheffé, Tukey, and REGW-Q. The process is best described with an example, which is provided at the end of this chapter.

6.4 Assumptions

The assumptions of the ANOVA are similar to the assumptions for the t-test described in Chap. 4.

1. Independent samples.
2. Gaussian distributed populations.
3. The independent variable is discrete, while the dependent variable is continuous.
4. Homogeneity of variance: All groups have the same variance.
5. The sample size needs to be determined before the experiment.

6.5 Example Calculations for a One-Way Independent Measures ANOVA

6.5.1 Computation of the ANOVA

Suppose there is a sword fighting tournament with three different types of swords: light sabers, Hattori Hanzo katanas, and elvish daggers (see Fig. 6.3). We are asking whether there are differences in the number of wins across swords. Hence, our null hypothesis is that there is no difference. The data and the computation of the F-value are shown in Fig. 6.3.

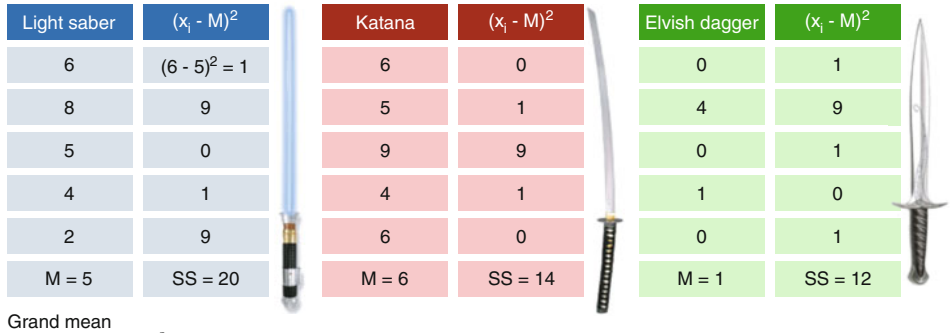
Our final[1] F-value is 9.14. This means that the variability of the group means around the grand mean is 9.14 times the variability of the data points around their individual group means. Hence, much of the variability comes from differences in the means, much less comes from variability within each population. An F-value of 9.14 leads to a p-value of $0.0039 < 0.05$ and we conclude that our results are significant, i.e., we reject the null hypothesis that all three sword types yield equal mean numbers of wins ($F(2, 12) = 9.14$, $p = 0.0039$). Furthermore, we can conclude that at least one sword type yields a different number of wins than the other sword types. We can now use one of the various post-hoc tests to find out which sword(s) is/are superior.

[1] If the data are analyzed by a statistics program, you will get $F = 9.13$. The difference is due to rounding of MS_{Within} in Fig. 6.3.

Light saber	$(x_i - M)^2$
6	$(6 - 5)^2 = 1$
8	9
5	0
4	1
2	9
M = 5	SS = 20

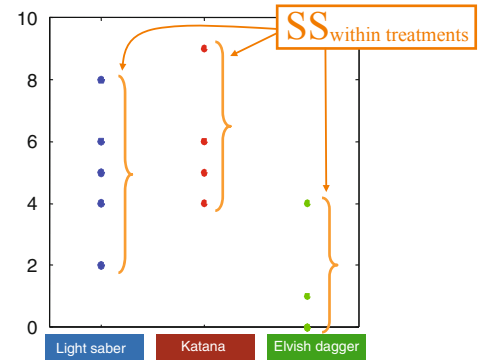
Katana	$(x_i - M)^2$
6	0
5	1
9	9
4	1
6	0
M = 6	SS = 14

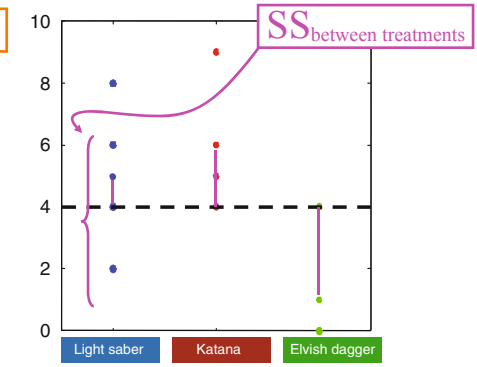
Elvish dagger	$(x_i - M)^2$
0	1
4	9
0	1
1	0
0	1
M = 1	SS = 12



Grand mean

$$M_G = \frac{\sum_{k=1}^{3}\sum_{i=1}^{n_k} x_{ik}}{N} = \frac{6+8+5+4+2+6+5+9+4+6+0+4+0+1+0}{15} = 4$$





$$SS_{within} = \sum_{k=1}^{3} SS_k = 20 + 14 + 12 = 46$$

$$SS_{between} = \sum_{k=1}^{3} n_k (M_k - M_G)^2 = 5(5-4)^2 + 5(6-4)^2\,5(1-4)^2 = 70$$

$$MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{46}{12} = 3.83$$

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{70}{2} = 35$$

$$F = \frac{MS_{between}}{MS_{within}} = \frac{35}{3.83} = 9.14$$

Source	SS	df	MS	F	p
Between treatments	70	2	35	9.14	0.0039
Within treatments	46	12	3.83		

Fig. 6.3 Example calculations for a one-way independent measures ANOVA. Each of the three swords is used by five independent fighters, making up all together 15 fighters. Hence, we have a 1×3 ANOVA. The opponents in the fights are not from the 15 test fighters. The upper left panel shows how many fights were won with the swords. The table below shows the data by numbers. First, we compute the mean for each sword type. For example, with the light sabers five wins occurred on average. Next, we compute the variability for each sword, also called the sum of squares inside a

6.5.2 Post-Hoc Tests

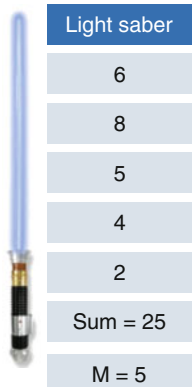
Various procedures exist for performing post-hoc tests, but we will focus here on the Scheffé test in order to illustrate some general principles.

The idea behind the Scheffé test is to perform multiple comparisons by computing pairwise ANOVA' s (e.g., light sabers vs. katanas, light sabers vs. elvish daggers, and katanas vs. elvish daggers). One assumption of the ANOVA is that all populations have equal variances. If this is true, then the best estimate of the variability within each population is the pooled estimate from the overall ANOVA calculated by MS_{within} (i.e., 3.83 in this case). The Scheffé test also uses $df_{between}$ from the overall ANOVA, and the calculations for performing this test are illustrated in Fig. 6.4.

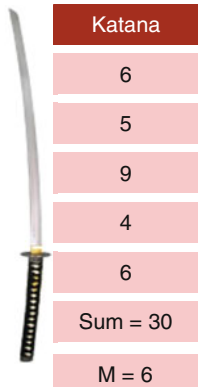
The p-value for each of the comparisons is computed using the degrees of freedom from the original ANOVA (i.e., $df_{between} = 2$ and $df_{within} = 12$). This yields the results in Table 6.1 for our post-hoc tests. Only the second and third comparisons are below our critical threshold of $\alpha = 0.05$, thus, we can conclude that the light sabers differ from the elvish daggers ($F(2, 12) = 5.22$, $p = 0.023$), and that katanas also differ from elvish daggers ($F(2, 12) = 8.15$, $p = 0.006$), but that we failed to find a significant difference between light sabers and katanas ($F(2, 12) = 0.33$, $p = 0.728$).

A common way to illustrate these differences is to plot a graph showing the mean number of wins for the three sword types with error bars denoting standard errors around each mean, and lines connecting the significantly different sword types with asterisks above them (Fig. 6.5).

Fig. 6.3 (continued) treatment. For the computation, we subtract each data point from the mean and square the value. To compute the variance within groups, we add the three sums of squares. In this case, we arrive at a value of 46. The next main step is to compute the variability between means. For this, we compute first the Grand Mean M_G, which is simply the mean number of wins over all 15 sword fights. In this example, the Grand Mean is 4. Next, for each sword, we subtract the means from the Grand Mean, square the values and multiply with the number of fights for each sword type (five in this example). We arrive at a value of 70. Next we divide our two sum of squares values by the degrees of freedom df_1 and df_2 in order to arrive at variances. We had three types of swords, hence, $df_1 = 3 - 1$, so we divide 70 by 2. We had 15 fights, hence, $df_2 = 12$, so we divide 46 by 12 (MS means mean square). For the test statistic, we divide the two values of 35 and 3.83 and arrive at $F = 9.14$. As with the t-value, the F-value can easily be computed by hand. For the p-value, we use statistics software, which delivers $p = 0.0039$. The output of software packages summarize the computation in panels similar to the one shown here. In publications, a result like this is presented as ($F(2, 12) = 9.14$, $p = 0.0039$)



Light saber
6
8
5
4
2
Sum = 25
M = 5



Katana
6
5
9
4
6
Sum = 30
M = 6

Elvish dagger
0
4
0
1
0
Sum = 5
M = 1

$$df_{between} = 2$$
$$MS_{within} = 3.83$$

From the original ANOVA table

$$SS_{between} = \sum_{k \in comp} n_k \left(M_i - G_{comp}\right)^2$$

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$F = \frac{MS_{between}}{MS_{within}}$$

Light saber versus Katana

$$G_{comp} = \frac{25 + 30}{10} = 5.5$$

$$SS_{between} = 5\left(5 - 5.5\right)^2 + 5\left(6 - 5.5\right)^2 = 2.5$$

$$MS_{between} = \frac{2.5}{2} = 1.25$$

$$F = \frac{1.25}{3.83} = 0.3264$$

Light saber versus Elvish dagger

$$G_{comp} = \frac{25 + 5}{10} = 3$$

$$SS_{between} = 5\left(5 - 3\right)^2 + 5\left(1 - 3\right)^2 = 40.0$$

$$MS_{between} = \frac{40}{2} = 20$$

$$F = \frac{20}{3.83} = 5.2219$$

Katana versus Elvish dagger

$$G_{comp} = \frac{30 + 5}{10} = 3.5$$

$$SS_{between} = 5\left(6 - 3.5\right)^2 + 5\left(1 - 3.5\right)^2 = 62.5$$

$$MS_{between} = \frac{62.5}{2} = 31.25$$

$$F = \frac{31.25}{3.83} = 8.1593$$

Fig. 6.4 Calculations for Scheffé post-hoc comparisons on our sword example. First we calculate the grand mean for each comparison (G_{Comp}), which consists of the mean of the pooled scores for the pairs of groups being considered. Next, using this mean we compute the sum of squared deviations between group means and the grand mean for the two groups being ($SS_{Between}$). The $MS_{between}$ scores are then obtained by dividing by $df_{between}$ from the overall ANOVA (i.e., two in this case). Finally, the F-ratio for each comparison is computed by dividing by the MS_{within} term from the overall ANOVA (i.e., 3.83 in this example)

Table 6.1 Post-hoc Scheffé test results for our three comparisons

Comparison	Result
1 vs. 2[a]	$F(2,12) = 0.33$, $p = 0.728$
1 vs. 3[b]	$F(2,12) = 5.22$, $p = 0.023$
2 vs. 3[c]	$F(2,12) = 8.16$, $p = 0.006$

[a]Light sabers versus katanas
[b]Light sabers versus elvish daggers
[c]Katanas versus elvish daggers

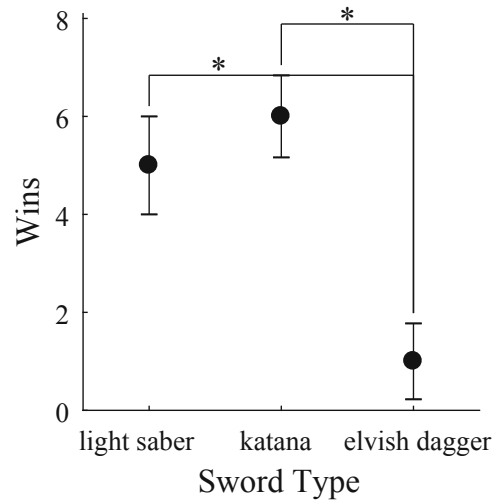


Fig. 6.5 Mean numbers of wins for the three sword types plotted with standard errors and lines with asterisks above them connecting sword types that are significantly different

6.6 Effect Size

As with the t-test, the p-value from an ANOVA confounds the effect size and the sample size. It is always important to look at the effect size, which for an ANOVA is denoted by η^2. The effect size η^2 tells you the proportion of the total variability in the dependent variable that is explained by the variability of the independent variable. The calculation is:

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

with

$$SS_{between} = \sum_{j=1}^{k} n_j(\bar{x}_j - M_G)^2 \tag{6.1}$$

$$SS_{total} = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - M_G)^2 \tag{6.2}$$

Table 6.2 Effect size guidelines according to Cohen

	Small	Medium	Large
Effect size	0.01	0.09	0.25

where M_G is the grand mean (i.e., average over all data points). This ratio tells you the proportion of the total variability in the data explained by variability due to the treatment means. For the above sword example the effect size is $\eta^2 = 0.60$, which is a large effect according to Cohen, who provided guidelines for effect sizes (Table 6.2).

6.7 Two-Way Independent Measures ANOVA

The one-way independent measures ANOVA generalizes nicely to cases with more than one factor. Here, we will discuss the simplest of such cases, the two-factor design.

Suppose that you and your friends acquire super powers in a science experiment and you are preparing to start your life of fighting crime. You and your super hero friends do not want your enemies to hurt your loved ones so you need costumes to conceal your identity. Furthermore, sometimes you might fight crime during the day, and other times you might fight crime at night. You want to know which costume material (spandex, cotton, or leather) will be the best for crime fighting as measured by the number of evil villains a hero can catch while wearing costumes made from each material, and you want to know if there is an effect of time of day on which material is best. You assign each friend to a costume material and time of day condition and count the number of evil villains each hero catches. You have different friends in each group. In this case, there are three separate hypotheses that we can make about the data:

1. H_0: There is no effect of time of day on the number of villains caught.
   H_1: The number of villains caught during the day are different from the number of villains caught at night.
2. H_0: There is no effect of costume material on the number of villains caught.
   H_1: At least one costume material yields different numbers of villains caught than the other costume materials.
3. H_0: The effect of time of day on the number of villains caught does not depend on costume material.
   H_1: The effect of time of day on the number of villains caught does depend on costume material.

The first two null hypotheses relate to what are called *main effects*. The two main hypotheses are exactly the same as computing two one-way ANOVAs. The third hypothesis is a new type of hypothesis and pertains to the *interaction* between the two factors, costume and day time. To measure the main effect of costume material, we take the average number of villains caught in the spandex group, averaging over both day and

night conditions, and compare this with the same averages for the cotton and leather costume conditions. To measure the main effect of time of day, we look at the average number of villains caught for the day condition, averaging over the spandex, cotton, and leather costume material conditions, and compare this with the same average for the night condition.

For the interaction, we consider all groups separately, looking at the number of villains caught for spandex, cotton and leather costume groups separately as a function of day- and night-time crime-fighting conditions. If there is a significant interaction, then the effects of time of day on the number of villains caught will depend on which costume material we are looking at. Conversely, the effect of costume material on the number of villains caught will depend on which time of day our friends are fighting crime at.

Testing these three null hypotheses requires three separate F-statistics. Each F-statistic will use the same denominator as in the one-way ANOVA (i.e., the pooled variance of the data about the treatment means, or MS_{within} as shown in Fig. 6.3), but the numerators ($MS_{between}$) will be specific for the particular hypotheses tested.

Figure 6.6 shows example raw data and the means being compared for the three hypotheses being tested (see margins). When pooling over time of day it looks like costume material has very little effect on crime fighting abilities. When pooling over costume material, it looks like time of day has also little effect on crime fighting abilities. It is only when we consider each mean individually that we can see the true effects of time

	Spandex	Cotton	Leather	Time of Day Means
Day	18	10	3	9.7
	10	8	5	
	16	12	1	
	12	6	7	
	14	14	9	
Day Means	14	10	5	
Night	5	6	15	10
	7	14	13	
	3	10	17	
	9	8	11	
	1	12	19	
Night Means	5	10	15	
Costume Means	9.5	10	10	Grand Mean = 9.8

Fig. 6.6 Number of villains caught for each super hero. As well as means for main effects (time of day and costume type), and individual cell means (spandex during the day, spandex at night, cotton during the day, etc.). The grand mean is the mean overall data points

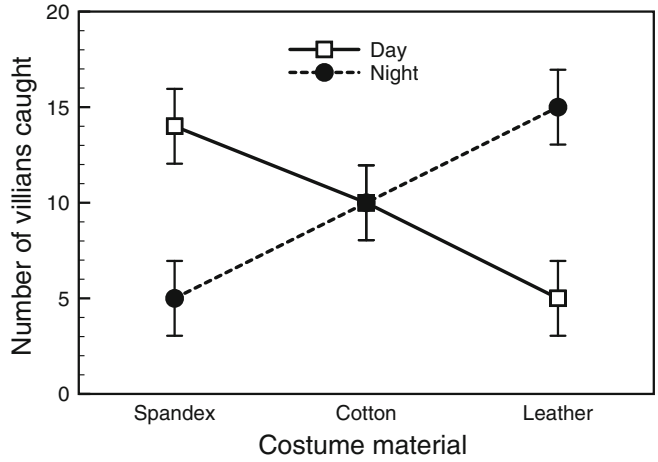


Fig. 6.7 Interaction plot for the number of villains caught as a function of costume material and time of day. Here we can see that the effect of costume material on the number of villains caught depends critically on the time of day

of day and costume material on how many villains our friends are catching: there is an interaction between costume material and time of day in relating the number of villains caught (Fig. 6.7). This interaction is such that spandex is better during the day and leather better at night, with cotton always being somewhere in the middle.

This example illustrates the value of a two-factor design. Had we done only separate one-way ANOVAs examining the relationships between costume material and number of villains caught, or time of day and number of villains caught, we would have found little or no effects. Including both variables reveals the true nature of both effects, showing the effect of one to depend on the level of the other. Figure 6.8 demonstrates three possible outcome patterns that isolate just one significant effect (Main effect of A, Main effect of B, Interaction) without any of the other effects. It is also possible to have combinations of main and interaction effects.

Another virtue of a two-factor design relative to a one-factor design is that variability that would otherwise be included in the error term (i.e., MS_{within}) is now partly explained by variability due to another factor, thereby reducing MS_{within} and increasing the power for detecting effects when present.

Thus, it may seem that the more factors we add the better we will understand the data and obtain significant results. However, this is not true because we lose power for each factor we are adding due to the fact that we have fewer scores contributing to each mean. Typically, larger samples are needed when the number of factors increases.

Importantly, if we find a significant interaction, the main effect varies depending on the other factor. Thus, we should usually refrain from making conclusions about the main effect if there is an interaction.

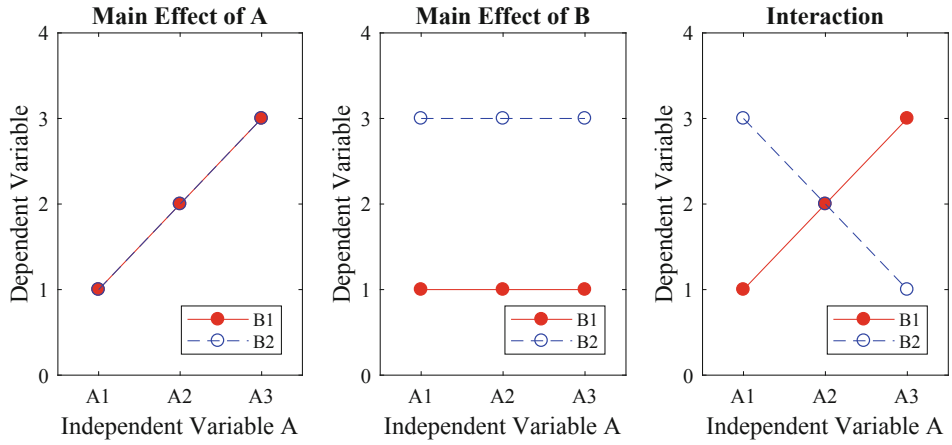


Fig. 6.8 The results of a two-factor ANOVA may reveal three general types of significant results: a main effect of variable A, a main effect of variable B, and an interaction between A and B. In the following example, we have a 2×3 ANOVA, where A1, A2, A3 may indicate the superhero's costume materials (Spandex, Cotton, Leather) and B1 and B2 times of day (night and day, respectively). The dependent variable would be villains caught. Left. Main effect of A. The costume material matters. More villains are caught while wearing leather than cotton. Time of day plays no role. As many villains are caught during the day as during the night. For this reason B1 and B2 are on top of each other. Center. Main effect of B. The costume material does not matter but the time of day does. More villains are caught during the day. Right. Interaction as in Fig. 6.7. In the case of a significant interaction, significant main effects are typically not examined because the more relevant comparison involves looking at the effect of one variable within the levels of the other variable

The one-way ANOVA avoids the multiple testing problem. However, a multi-way ANOVA reintroduces a kind of multiple testing problem. For example, consider a 2×2 ANOVA, with a significance criterion of 0.05. A truly null data set (where all four population means are equal to each other) has a 14% chance of producing at least one $p < 0.05$ among the two main effects and the interaction. If you use ANOVA to *explore* your data set by identifying significant results, you should understand that such an approach has a higher Type I error rate than you might have intended.

A typical statistical software package outputs the results of a two-way ANOVA as in Table 6.3.

6.8 Repeated Measures ANOVA

The ANOVA we have discussed up to now is a straightforward extension of the independent samples t-test. There also exists a generalization of the dependent samples t-test called the *repeated measures* ANOVA. You can use this kind of ANOVA when, for

Table 6.3 Typical statistical software outputs for a two-way ANOVA

Source	SS	df	MS	F	p	η^2
Costume material	1.67	2	0.83	0.083	0.920	0.0069
Time of day	0.83	1	0.83	0.083	0.775	0.0035
Costume × time	451.67	2	225.83	22.58	0.000003	0.6530
Error	240.00	24	10.00			

The columns show the source of variability (Source), the sums of squares (SS), the degrees of freedom (df), the mean squares (MS), the F-ratios (F), the p-values (p—sometimes also labeled "Sig."), and the effect sizes (η^2). The row labeled "Error" holds the computations for the within subjects variability, while the remaining rows show between subjects variability for the main effects and interactions

Table 6.4 Typical statistical software outputs for a repeated measures ANOVA

Source	SS	df	MS	F	p	η^2
Between times	70	2	35	70	0.00000009	0.94
Within times	40	12				
Between subjects	36	4				
Error	110	14				

Here the example is for patient symptoms measured before, during, and after treatment (i.e., at different times). The first row (*Between times*) shows the effect of time of measurement on symptoms. The within times row shows the variability due to subjects within each time condition. It is broken down into consistent trends for each individual subject (*Between subjects*) and random error caused by things like drug diffusion rates (*Error*). The columns show the source of variability (Source), the sums of squares (SS), the degrees of freedom (df), the mean squares (MS), the F-ratios (F), the p-values (p—sometimes also labeled "Sig."), and the effect sizes (η^2). The row labeled " Error" holds the computations for the error term which is used in the denominator of the F-ratio. In the independent measures ANOVA this term is the within treatments term. Here, however, we remove from the within treatments term the variability due to subjects, so we now simply call the remaining variability " error variability." The remaining rows show between subjects variability for the main effects and interactions. To summarize these results we would say that there is a significant effect of time of measurement on symptoms $F(2, 14) = 70$, $p = 0.00000009$. Here, we have taken the degrees of freedom from the between times and error rows, and have taken the F- and p-values from the between times row

example, measuring some aspect of patient health before, during, and after some treatment program. In this case, the very same patients undergo three measurements. A repeated measures ANOVA has higher power than the independent measures ANOVA because it compares the differences within patients first before comparing across the patients, thus, reducing variability in the data. Example output from a repeated measures ANOVA is provided in Table 6.4.

Take Home Messages

1. With an ANOVA you can avoid the multiple testing problem—to some extent.
2. More factors may improve or deteriorate power.