Chapter 7

One-way ANOVA

One-way ANOVA examines equality of population means for a quantitative outcome and a single categorical explanatory variable with any number of levels.

The t-test of Chapter 6 looks at quantitative outcomes with a categorical explanatory variable that has only two levels. The one-way **Analysis of Variance** (**ANOVA**) can be used for the case of a quantitative outcome with a categorical explanatory variable that has two or more levels of treatment. The term one-way, also called one-factor, indicates that there is a single explanatory variable ("treatment") with two or more levels, and only one level of treatment is applied at any time for a given subject. In this chapter we assume that each subject is exposed to only one treatment, in which case the treatment variable is being applied "between-subjects". For the alternative in which each subject is exposed to several or all levels of treatment (at different times) we use the term "within-subjects", but that is covered Chapter 14. We use the term two-way or two-factor ANOVA, when the levels of two different explanatory variables are being assigned, and each subject is assigned to one level of *each* factor.

It is worth noting that the situation for which we can choose between one-way ANOVA and an independent samples t-test is when the explanatory variable has exactly two levels. In that case we always come to the same conclusions regardless of which method we use.

The term "analysis of variance" is a bit of a misnomer. In ANOVA we use variance-like quantities to study the equality or non-equality of population means. So we are analyzing means, not variances. There are some unrelated methods,

such as "variance component analysis" which have variances as the primary focus for inference.

7.1 Moral Sentiment Example

As an example of application of one-way ANOVA consider the research reported in "Moral sentiments and cooperation: Differential influences of shame and guilt" by de Hooge, Zeelenberg, and M. Breugelmans (Cognition & Emotion,21(5): 1025-1042, 2007).

As background you need to know that there is a well-established theory of Social Value Orientations or SVO (see Wikipedia for a brief introduction and references). SVOs represent characteristics of people with regard to their basic motivations. In this study a questionnaire called the Triple Dominance Measure was used to categorize subjects into "proself" and "prosocial" orientations. In this chapter we will examine simulated data based on the results for the proself individuals.

The goal of the study was to investigate the effects of emotion on cooperation. The study was carried out using undergraduate economics and psychology students in the Netherlands.

The sole explanatory variable is "induced emotion". This is a nominal categorical variable with three levels: control, guilt and shame. Each subject was randomly assigned to one of the three levels of treatment. Guilt and shame were induced in the subjects by asking them to write about a personal experience where they experienced guilt or shame respectively. The control condition consisted of having the subject write about what they did on a recent weekday. (The validity of the emotion induction was tested by asking the subjects to rate how strongly they were feeling a variety of emotions towards the end of the experiment.)

After inducing one of the three emotions, the experimenters had the subjects participate in a one-round computer game that is designed to test cooperation. Each subject initially had ten coins, with each coin worth 0.50 Euros for the subject but 1 Euro for their "partner" who is presumably connected separately to the computer. The subjects were told that the partners also had ten coins, each worth 0.50 Euros for themselves but 1 Euro for the subject. The subjects decided how many coins to give to the interaction partner, without knowing how many coins the interaction partner would give. In this game, both participants would earn 10 Euros when both offered all coins to the interaction partner (the

cooperative option). If a cooperator gave all 10 coins but their partner gave none, the cooperator could end up with nothing, and the partner would end up with the maximum of 15 Euros. Participants could avoid the possibility of earning nothing by keeping all their coins to themselves which is worth 5 Euros plus 1 Euro for each coin their partner gives them (the selfish option). The number of coins offered was the measure of cooperation.

The number of coins offered (0 to 10) is the outcome variable, and is called "cooperation". Obviously this outcome is related to the concept of "cooperation" and is in some senses a good measure of cooperation, but just as obviously, it is not a complete measure of the concept.

Cooperation as defined here is a discrete quantitative variable with a limited range of possible values. As explained below, the Analysis of Variance statistical procedure, like the t-test, is based on the assumption of a Gaussian distribution of the outcome at each level of the (categorical) explanatory variable. In this case, it is judged to be a reasonable approximation to treat "cooperation" as a continuous variable. There is no hard-and-fast rule, but 11 different values might be considered borderline, while, e.g., 5 different values would be hard to justify as possibly consistent with a Gaussian distribution.

Note that this is a randomized experiment. The levels of "treatment" (emotion induced) are randomized and assigned by the experimenter. If we do see evidence that "cooperation" differs among the groups, we can validly claim that induced emotion *causes* different degrees of cooperation. If we had only measured the subjects' current emotion rather than manipulating it, we could only conclude that emotion is *associated* with cooperation. Such an association could have other explanations than a causal relationship. E.g., poor sleep the night before could cause more feelings of guilt and more cooperation, without the guilt having any direct effect on cooperation. (See section 8.1 for more on causality.)

The data can be found in MoralSent.dat. The data look like this:

emotion	cooperation
Control	3
Control	0
Control	0

Typical exploratory data analyses include a tabulation of the frequencies of the levels of a categorical explanatory variable like "emotion". Here we see 39 controls, 42 guilt subjects, and 45 shame subjects. Some sample statistics of cooperation broken down by each level of induced emotion are shown in table 7.1, and side-by-

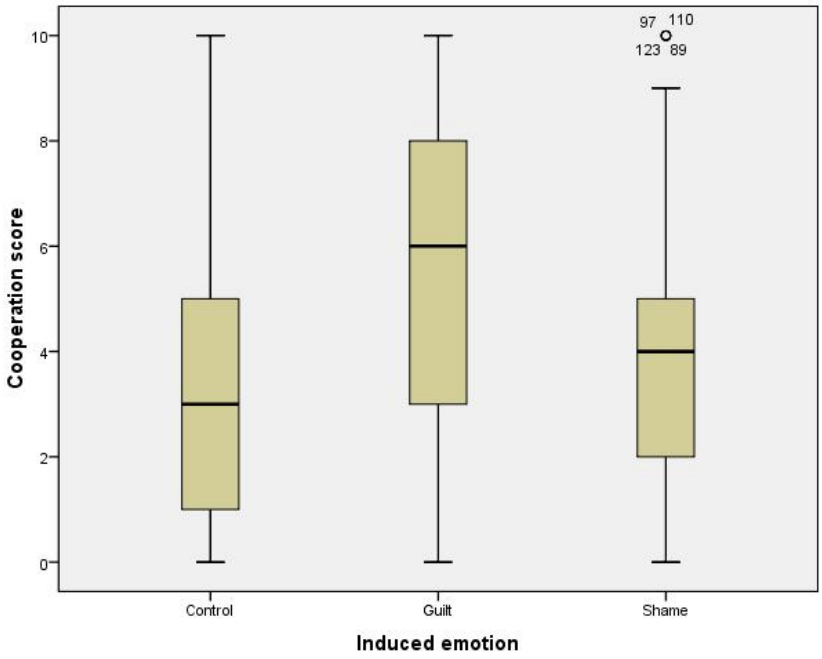

Figure 7.1: Boxplots of cooperation by induced emotion.

side boxplots shown in figure 7.1.

Our initial impression is that cooperation is higher for guilt than either shame or the control condition. The mean cooperation for shame is slightly lower than for the control. In terms of pre-checking model assumptions, the boxplots show fairly symmetric distributions with fairly equal spread (as demonstrated by the comparative IQRs). We see four high outliers for the shame group, but careful thought suggests that this may be unimportant because they are just one unit of measurement (coin) into the outlier region and that region may be "pulled in' a bit by the slightly narrower IQR of the shame group.

	Induced emotion			Statistic	Std.Error
Cooperation score	Control	Mean		3.49	0.50
		95% Confidence Interval for Mean	Lower Bound	2.48	
			Upper Bound	4.50	
		Median		3.00	
		Std. Deviation		3.11	
		Minimum		0	
		Maximum		10	
		Skewness		0.57	0.38
		Kurtosis		-0.81	0.74
	Guilt	Mean		5.38	0.50
		95% Confidence Interval for Mean	Lower Bound	4.37	
			Upper Bound	6.39	
		Median		6.00	
		Std. Deviation		3.25	
		Minimum		0	
		Maximum		10	
		Skewness		-0.19	0.36
		Kurtosis		-1.17	0.72
	Shame	Mean		3.78	0.44
		95% Confidence Interval for Mean	Lower Bound	2.89	
			Upper Bound	4.66	
		Median		4.00	
		Std. Deviation		2.95	
		Minimum		0	
		Maximum		10	
		Skewness		0.71	0.35
		Kurtosis		-0.20	0.70

Table 7.1: Group statistics for the moral sentiment experiment.

7.2 How one-way ANOVA works

7.2.1 The model and statistical hypotheses

One-way ANOVA is appropriate when the following model holds. We have a single "treatment" with, say, k levels. "Treatment" may be interpreted in the loosest possible sense as any categorical explanatory variable. There is a population of interest for which there is a true quantitative outcome for each of the k levels of treatment. The population outcomes for each group have mean parameters that we can label μ_1 through μ_k with no restrictions on the pattern of means. The population variances for the outcome for each of the k groups defined by the levels of the explanatory variable all have the same value, usually called σ^2, with no restriction other than that $\sigma^2 > 0$. For treatment i, the distribution of the outcome is assumed to follow a Normal distribution with mean μ_i and variance σ^2, often written $N(\mu_i, \sigma^2)$.

Our model assumes that the true deviations of observations from their corresponding group mean parameters, called the "errors", are independent. In this context, independence indicates that knowing one true deviation would not help us predict any other true deviation. Because it is common that subjects who have a high outcome when given one treatment tend to have a high outcome when given another treatment, using the same subject twice would violate the independence assumption.

Subjects are randomly selected from the population, and then randomly assigned to exactly one treatment each. The number of subjects assigned to treatment i (where $1 \leq i \leq k$) is called n_i if it differs between treatments or just n if all of the treatments have the same number of subjects. For convenience, define $N = \sum_{i=1}^{k} n_i$, which is the total sample size.

(In case you have forgotten, the Greek capital sigma (Σ) stands for summation, i.e., adding. In this case, the notation says that we should consider all values of n_i where i is set to 1, 2, ..., k, and then add them all up. For example, if we have $k = 3$ levels of treatment, and the group samples sizes are 12, 11, and 14 respectively, then $n_1 = 12$, $n_2 = 11$, $n_3 = 14$ and $N = \sum_{i=1}^{k} n_i = n_1 + n_2 + n_3 = 12 + 11 + 14 = 37$.)

Because of the random treatment assignment, the sample mean for any treatment group is representative of the population mean for assignment to that group for the entire population.

Technically, the sample group means are unbiased estimators of the population group means when treatment is randomly assigned. The meaning of unbiased here is that the true mean of the sampling distribution of any group sample mean equals the corresponding population mean. Further, under the Normality, independence and equal variance assumptions it is true that the sampling distribution of $\bar{Y}_i$ is $N(\mu_i, \sigma^2/n_i)$, exactly.

The statistical model for which one-way ANOVA is appropriate is that the (quantitative) outcomes for each group are normally distributed with a common variance (σ^2). The errors (deviations of individual outcomes from the population group means) are assumed to be independent. The model places no restrictions on the population group means.

The term **assumption** in statistics refers to any specific part of a statistical model. For one-way ANOVA, the assumptions are normality, equal variance, and independence of errors. Correct assignment of individuals to groups is sometimes considered to be an implicit assumption.

The null hypothesis is a point hypothesis stating that "nothing interesting is happening." For one-way ANOVA, we use $H_0 : \mu_1 = \cdots = \mu_k$, which states that all of the population means are equal, without restricting what the common value is. The alternative must include everything else, which can be expressed as "at least one of the k population means differs from all of the others". It is *definitely wrong* to use $H_A : \mu_1 \neq \cdots \neq \mu_k$ because some cases, such as $\mu_1 = 5,\ \ \mu_2 = 5,\ \ \mu_3 = 10$, are neither covered by H_0 nor this incorrect H_A. You can write the alternative hypothesis as "H_A : Not $\mu_1 = \cdots = \mu_k$ or "the population means are not all equal".

One way to correctly write H_A mathematically is $H_A : \exists\, i, j : \mu_i \neq \mu_j$.

This null hypothesis is called the "overall" null hypothesis and is the hypothesis tested by ANOVA, per se. If we have only two levels of our categorical explanatory

variable, then retaining or rejecting the overall null hypothesis, is all that needs to be done in terms of hypothesis testing. But if we have 3 or more levels ($k \geq 3$), then we usually need to followup on rejection of the overall null hypothesis with more specific hypotheses to determine for which population group means we have evidence of a difference. This is called contrast testing and discussion of it will be delayed until chapter 13.

The overall null hypothesis for one-way ANOVA with k groups is $H_0 : \mu_1 = \cdots = \mu_k$. The alternative hypothesis is that "the population means are not all equal".

7.2.2 The F statistic (ratio)

The next step in standard inference is to select a statistic for which we can compute the null sampling distribution and that tends to fall in a different region for the alternative than the null hypothesis. For ANOVA, we use the "F-statistic". The single formula for the F-statistic that is shown in most textbooks is quite complex and hard to understand. But we can build it up in small understandable steps.

Remember that a sample variance is calculated as SS/df where SS is "sum of squared deviations from the mean" and df is "degrees of freedom" (see page 69). In ANOVA we work with variances and also "variance-like quantities" which are not really the variance of anything, but are still calculated as SS/df. We will call all of these quantities **mean squares** or MS. i.e., $MS = SS/df$, which is a key formula that you should memorize. Note that these are not really means, because the denominator is the df, not n.

For one-way ANOVA we will work with two different MS values called "mean square within-groups", $\text{MS}_{\text{within}}$, and "mean square between-groups", $\text{MS}_{\text{between}}$. We know the general formula for any MS, so we really just need to find the formulas for $\text{SS}_{\text{within}}$ and $\text{SS}_{\text{between}}$, and their corresponding df.

The F statistic denominator: $\text{MS}_{\text{within}}$

$\text{MS}_{\text{within}}$ is a "pure" estimate of σ^2 that is unaffected by whether the null or alternative hypothesis is true. Consider figure 7.2 which represents the within-group

deviations used in the calculation of $\text{MS}_{\text{within}}$ for a simple two-group experiment with 4 subjects in each group. The extension to more groups and/or different numbers of subjects is straightforward.

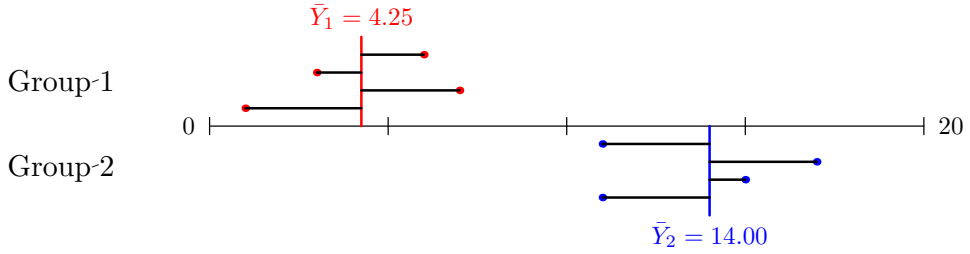

Figure 7.2: Deviations for within-group sum of squares

The deviation for subject j of group i in figure 7.2 is mathematically equal to $Y_{ij} - \bar{Y}_i$ where Y_{ij} is the observed value for subject j of group i and $\bar{Y}_i$ is the sample mean for group i.

I hope you can see that the deviations shown (black horizontal lines extending from the colored points to the colored group mean lines) are due to the underlying variation of subjects within a group. The variation has standard deviation σ, so that, e.g., about 2/3 of the times the deviation lines are shorter than σ. Regardless of the truth of the null hypothesis, for each individual group, $\text{MS}_i = \text{SS}_i/\text{df}_i$ is a good estimate of σ^2. The value of $\text{MS}_{\text{within}}$ comes from a statistically appropriate formula for combining all of the k separate group estimates of σ^2. It is important to know that $\text{MS}_{\text{within}}$ has $N - k$ df.

For an individual group, i, $\text{SS}_i = \sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$ and $\text{df}_i = n_i - 1$. We can use some statistical theory beyond the scope of this course to show that in general, $\text{MS}_{\text{within}}$ is a good (unbiased) estimate of σ^2 if it is defined as

$$\text{MS}_{\text{within}} = \text{SS}_{\text{within}}/\text{df}_{\text{within}}$$

where $\text{SS}_{\text{within}} = \sum_{i=1}^{k} SS_i$, and $\text{df}_{\text{within}} = \sum_{i=1}^{k} \text{df}_i = \sum_{i=1}^{k}(n_i - 1) = N - k$.

$\text{MS}_{\text{within}}$ is a good estimate of σ^2 (from our model) regardless of the truth of H_0. This is due to the way $\text{SS}_{\text{within}}$ is defined. $\text{SS}_{\text{within}}$ (and therefore $\text{MS}_{\text{within}}$) has N-k degrees of freedom with $n_i - 1$ coming from each of the k groups.

The F statistic numerator: $\text{MS}_{\text{between}}$

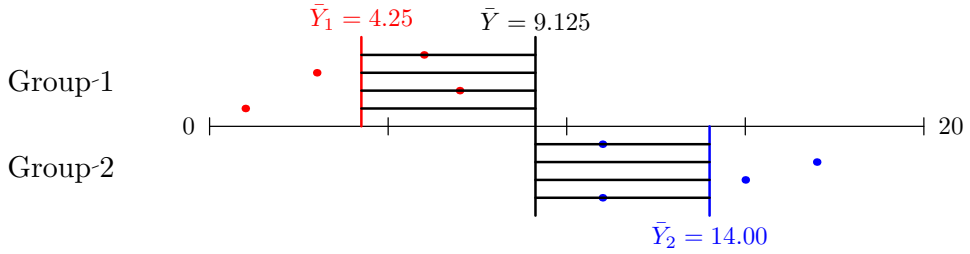

Figure 7.3: Deviations for between-group sum of squares

Now consider figure 7.3 which represents the between-group deviations used in the calculation of $\text{MS}_{\text{between}}$ for the same little 2-group 8-subject experiment as shown in figure 7.2. The single vertical black line is the average of all of the outcomes values in all of the treatment groups, usually called either the overall mean or the **grand mean**. The colored vertical lines are still the group means. The horizontal black lines are the deviations used for the between-group calculations. For each subject we get a deviation equal to the distance (difference) from that subject's group mean to the overall (grand) mean. These deviations are squared and summed to get $\text{SS}_{\text{between}}$, which is then divided by the between-group df, which is $k - 1$, to get $\text{MS}_{\text{between}}$.

$\text{MS}_{\text{between}}$ is a good estimate of σ^2 only when the null hypothesis is true. In this case we expect the group means to be fairly close together and close to the

grand mean. When the alternate hypothesis is true, as in our current example, the group means are farther apart and the value of $\text{MS}_{\text{between}}$ tends to be larger than σ^2. (We sometimes write this as "$\text{MS}_{\text{between}}$ is an inflated estimate of σ^2".)

$\text{SS}_{\text{between}}$ is the sum of the N squared between-group deviations, where the deviation is the same for all subjects in the same group. The formula is

$$\text{SS}_{\text{between}} = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{\bar{Y}})^2$$

where $\bar{\bar{Y}}$ is the grand mean. Because the k unique deviations add up to zero, we are free to choose only $k-1$ of them, and then the last one is fully determined by the others, which is why $\text{df}_{\text{between}} = k-1$ for one-way ANOVA.

Because of the way $\text{SS}_{\text{between}}$ is defined, $\text{MS}_{\text{between}}$ is a good estimate of σ^2 only if H_0 is true. Otherwise it tends to be larger. $\text{SS}_{\text{between}}$ (and therefore $\text{MS}_{\text{between}}$) has $k-1$ degrees of freedom.

The F statistic ratio

It might seem that we only need $\text{MS}_{\text{between}}$ to distinguish the null from the alternative hypothesis, but that ignores the fact that we don't usually know the value of σ^2. So instead we look at the ratio

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}$$

to evaluate the null hypothesis. Because the denominator is always (under null and alternative hypotheses) an estimate of σ^2 (i.e., tends to have a value near σ^2), and the numerator is either another estimate of σ^2 (under the null hypothesis) or is inflated (under the alternative hypothesis), it is clear that the (random) values of the F-statistic (from experiment to experiment) tend to fall around 1.0 when

the null hypothesis is true and are *bigger* when the alternative is true. So if we can compute the sampling distribution of the F statistic under the null hypothesis, then we will have a useful statistic for distinguishing the null from the alternative hypotheses, where large values of F argue for rejection of H_0.

The F-statistic, defined by $F = \frac{\mathbf{MS_{between}}}{\mathbf{MS_{within}}}$, tends to be larger if the alternative hypothesis is true than if the null hypothesis is true.

7.2.3 Null sampling distribution of the F statistic

Using the technical condition that the quantities $\text{MS}_{\text{between}}$ and $\text{MS}_{\text{within}}$ are independent, we can apply probability and statistics techniques (beyond the scope of this course) to show that the null sampling distribution of the F statistic is that of the "F-distribution" (see section 3.9.7). The F-distribution is indexed by two numbers called the numerator and denominator degrees of freedom. This indicates that there are (infinitely) many F-distribution pdf curves, and we must specify these two numbers to select the appropriate one for any given situation.

Not surprisingly the null sampling distribution of the F-statistic for any given one-way ANOVA is the F-distribution with numerator degrees of freedom equal to $\text{df}_{\text{between}} = k - 1$ and denominator degrees of freedom equal to $\text{df}_{\text{within}} = N - k$. Note that this indicates that the kinds of F-statistic values we will see if the null hypothesis is true depends only on the number of groups and the numbers of subjects, and not on the values of the population variance or the population group means. It is worth mentioning that the degrees of freedom are measures of the "size" of the experiment, where bigger experiments (more groups or more subjects) have bigger df.

We can quantify "large" for the F-statistic, by comparing it to its null sampling distribution which is the specific F-distribution which has degrees of freedom matching the numerator and denominator of the F-statistic.

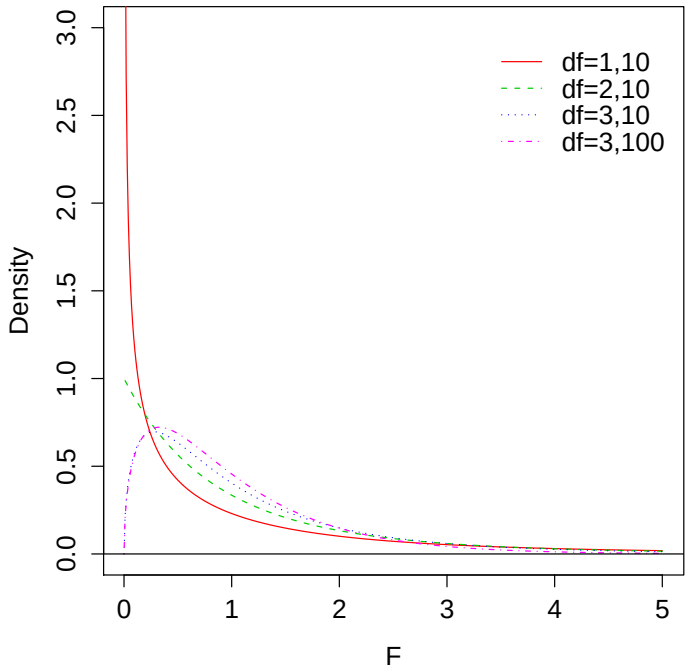

Figure 7.4: A variety of F-distribution pdfs.

The F-distribution is a non-negative distribution in the sense that F values, which are squares, can never be negative numbers. The distribution is skewed to the right and continues to have some tiny probability no matter how large F gets. The mean of the distribution is $s/(s-2)$, where s is the denominator degrees of freedom. So if s is reasonably large then the mean is near 1.00, but if s is small, then the mean is larger (e.g., k=2, n=4 per group gives s=3+3=6, and a mean of 6/4=1.5).

Examples of F-distributions with different numerator and denominator degrees of freedom are shown in figure 7.4. These curves are probability density functions, so the regions on the x-axis where the curve is high are the values most likely to occur. And the area under the curve between any two F values is equal to the probability that a random variable following the given distribution will fall between those values. Although very low F values are more likely for, say, the

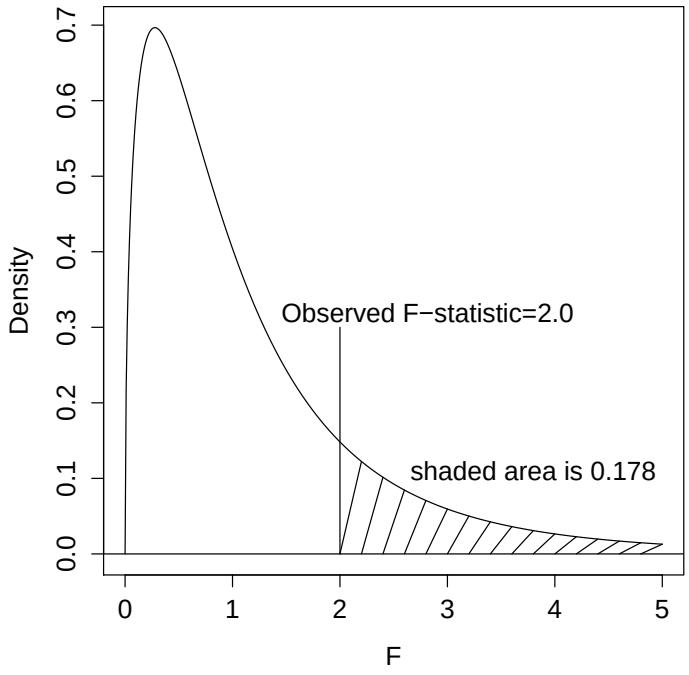

Figure 7.5: The F(3,10) pdf and the p-value for F=2.0.

F(1,10) distribution than the F(3,10) distribution, very high values are also more common for the F(1,10) than the F(3,10) values, though this may be hard to see in the figure. The bigger the numerator and/or denominator df, the more concentrated the F values will be around 1.0.

7.2.4 Inference: hypothesis testing

There are two ways to use the null sampling distribution of F in one-way ANOVA: to calculate a p-value or to find the "critical value" (see below).

A close up of the F-distribution with 3 and 10 degrees of freedom is shown in figure 7.5. This is the appropriate null sampling distribution of an F-statistic for an experiment with a quantitative outcome and one categorical explanatory variable (factor) with k=4 levels (each subject gets one of four different possible treatments) and with 14 subjects divided among the 4 groups. A vertical line marks an F-statistic of 2.0 (the observed value from some experiment). The p-value for this result is the chance of getting an F-statistic greater than or equal to

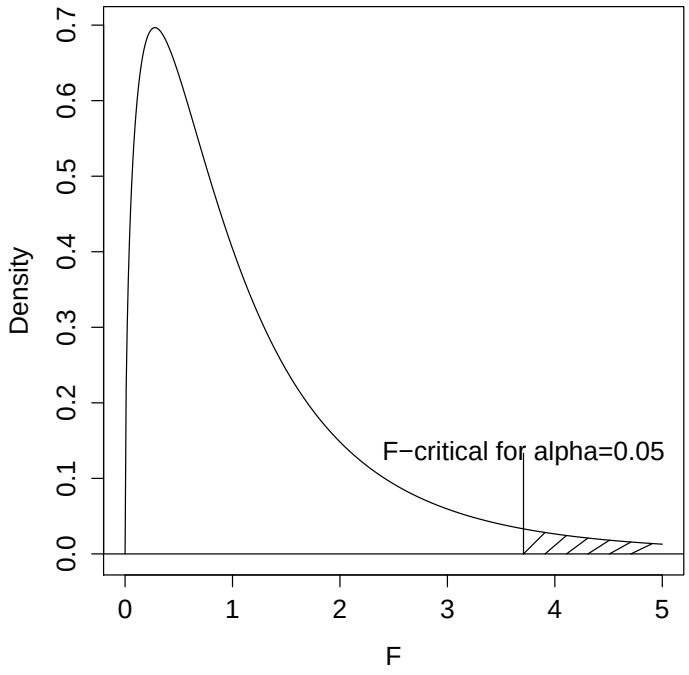

Figure 7.6: The F(3,10) pdf and its alpha=0.05 critical value.

2.0 when the null hypothesis is true, which is the shaded area. The total area is always 1.0, and the shaded area is 0.178 in this example, so the p-value is 0.178 (not significant at the usual 0.05 alpha level).

Figure 7.6 shows another close up of the F-distribution with 3 and 10 degrees of freedom. We will use this figure to define and calculate the **F-critical** value. For a given alpha (significance level), usually 0.05, the F-critical value is the F value above which $100\alpha\%$ of the null sampling distribution occurs. For experiments with 3 and 10 df, and using $\alpha = 0.05$, the figure shows that the F-critical value is 3.71. Note that this value can be obtained from a computer *before* the experiment is run, as long as we know how many subjects will be studied and how many levels the explanatory variable has. Then when the experiment is run, we can calculate the observed F-statistic and compare it to F-critical. If the statistic is smaller than the critical value, we retain the null hypothesis because the p-value must be bigger than α, and if the statistic is equal to or bigger than the critical value, we reject the null hypothesis because the p-value must be equal to or smaller than α.

7.2.5 Inference: confidence intervals

It is often worthwhile to express what we have learned from an experiment in terms of confidence intervals. In one-way ANOVA it is possible to make confidence intervals for population group means or for differences in pairs of population group means (or other more complex comparisons). We defer discussion of the latter to chapter 13.

Construction of a confidence interval for a population group means is usually done as an appropriate "plus or minus" amount around a sample group mean. We use $\text{MS}_{\text{within}}$ as an estimate of σ^2, and then for group i, the standard error of the mean is $\sqrt{\text{MS}_{\text{within}}/n_i}$. As discussed in section 6.2.7, the multiplier for the standard error of the mean is the so called "quantile of the t-distribution" which defines a central area equal to the desired confidence level. This comes from a computer or table of t-quantiles. For a 95% CI this is often symbolized as $t_{0.025,df}$ where df is the degrees of freedom of $\text{MS}_{\text{within}}$, $(N-k)$. Construct the CI as the sample mean plus or minus (SEM times the multiplier).

In a nutshell: In one-way ANOVA we calculate the F-statistic as the ratio $\text{MS}_{\text{between}}/\text{MS}_{\text{within}}$. Then the p-value is calculated as the area under the appropriate null sampling distribution of F that is bigger than the observed F-statistic. We reject the null hypothesis if $p \leq \alpha$.

7.3 Do it in SPSS

To run a one-way ANOVA in SPSS, use the Analyze menu, select Compare Means, then One-Way ANOVA. Add the quantitative outcome variable to the "Dependent List", and the categorical explanatory variable to the "Factor" box. Click OK to get the output. The dialog box for One-Way ANOVA is shown in figure 7.7.

You can also use the Options button to perform descriptive statistics by group, perform a variance homogeneity test, or make a means plot.

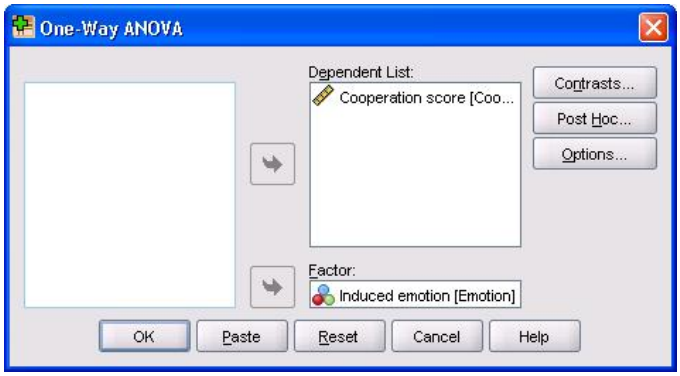

Figure 7.7: One-Way ANOVA dialog box.

You can use the Contrasts button to specify particular planned contrasts among the levels or you can use the Post-Hoc button to make unplanned contrasts (corrected for multiple comparisons), usually using the Tukey procedure for all pairs or the Dunnett procedure when comparing each level to a control level. See chapter 13 for more information.

7.4 Reading the ANOVA table

The **ANOVA table** is the main output of an ANOVA analysis. It always has the "source of variation" labels in the first column, plus additional columns for "sum of squares", "degrees of freedom", "means square", F, and the p-value (labeled "Sig." in SPSS).

For one-way ANOVA, there are always rows for "Between Groups" variation and "Within Groups" variation, and often a row for "Total" variation. In one-way ANOVA there is only a single F statistic ($\text{MS}_{\text{between}}/\text{MS}_{\text{within}}$), and this is shown on the "Between Groups" row. There is also only one p-value, because there is only one (overall) null hypothesis, namely $H_0 : \mu_1 = \cdots = \mu_k$, and because the p-value comes from comparing the (single) F value to its null sampling distribution. The calculation of MS for the total row is optional.

Table 7.2 shows the results for the moral sentiment experiment. There are several important aspects to this table that you should understand. First, as discussed above, the "Between Groups" lines refer to the variation of the group means around the grand mean, and the "Within Groups" line refers to the variation

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	86.35	2	43.18	4.50	0.013
Within Groups	1181.43	123	9.60		
Total	1267.78	125			

Table 7.2: ANOVA for the moral sentiment experiment.

of the subjects around their group means. The "Total" line refers to variation of the individual subjects around the grand mean. The Mean Square for the Total line is exactly the same as the variance of all of the data, ignoring the group assignments.

In any ANOVA table, the df column refers to the number of degrees of freedom in the particular SS defined on the same line. The MS on any line is always equal to the SS/df for that line. F-statistics are given on the line that has the MS that is the numerator of the F-statistic (ratio). The denominator comes from the MS of the "Within Groups" line for one-way ANOVA, but this is not always true for other types of ANOVA. It is always true that there is a p-value for each F-statistic, and that the p-value is the area under the null sampling distribution of that F-statistic that is above the (observed) F value shown in the table. Also, we can always tell which F-distribution is the appropriate null sampling distribution for any F-statistic, by finding the numerator and denominator df in the table.

An ANOVA is a breakdown of the total variation of the data, in the form of SS and df, into smaller independent components. For the one-way ANOVA, we break down the deviations of individual values from the overall mean of the data into deviations of the group means from the overall mean, and then deviations of the individuals from their group means. The independence of these sources of deviation results in additivity of the SS and df columns (but *not* the MS column). So we note that $\text{SS}_{\text{Total}} = \text{SS}_{\text{Between}} + \text{SS}_{\text{Within}}$ and $\text{df}_{\text{Total}} = \text{df}_{\text{Between}} + \text{df}_{\text{Within}}$. This fact can be used to reduce the amount of calculation, or just to check that the calculation were done and recorded correctly.

Note that we can calculate $\text{MS}_{\text{Total}} = 1267.78/125 = 10.14$ which is the variance of all of the data (thrown together and ignoring the treatment groups). You can see that MS_{Total} is certainly not equal to $\text{MS}_{\text{Between}} + \text{MS}_{\text{Within}}$.

Another use of the ANOVA table is to learn about an experiment when it is not full described (or to check that the ANOVA was performed and recorded

correctly). Just from this one-way ANOVA table, we can see that there were 3 treatment groups (because $\text{df}_{\text{Between}}$ is one less than the number of groups). Also, we can calculate that there were 125+1=126 subjects in the experiment.

Finally, it is worth knowing that $\text{MS}_{\text{within}}$ is an estimate of σ^2, the variance of outcomes around their group mean. So we can take the square root of $\text{MS}_{\text{within}}$ to get an estimate of σ, the standard deviation. Then we know that the majority (about $\frac{2}{3}$) of the measurements for each group are within σ of the group mean and most (about 95%) are within 2σ, assuming a Normal distribution. In this example the estimate of the s.d. is $\sqrt{9.60} = 3.10$, so individual subject cooperation values more than 2(3.10)=6.2 coins from their group means would be uncommon.

You should understand the structure of the one-way ANOVA table including that MS=SS/df for each line, SS and df are additive, F is the ratio of between to within group MS, the p-value comes from the F-statistic and its presumed (under model assumptions) null sampling distribution, and the number of treatments and number of subjects can be calculated from degrees of freedom.

7.5 Assumption checking

Except for the skewness of the shame group, the skewness and kurtosis statistics for all three groups are within 2SE of zero (see Table 7.1), and that one skewness is only slightly beyond 2SE from zero. This suggests that there is no evidence against the Normality assumption. The close similarity of the three group standard deviations suggests that the equal variance assumption is OK. And hopefully the subjects are totally unrelated, so the independent errors assumption is OK. Therefore we can accept that the F-distribution used to calculate the p-value from the F-statistic is the correct one, and we "believe" the p-value.

7.6 Conclusion about moral sentiments

With $p = 0.013 < 0.05$, we reject the null hypothesis that all three of the group population means of cooperation are equal. We therefore conclude that differences

in mean cooperation are caused by the induced emotions, and that among control, guilt, and shame, at least two of the population means differ. Again, we defer looking at *which* groups differ to chapter 13.

(A complete analysis would also include examination of residuals for additional evaluation of possible non-normality or unequal spread.)

The F-statistic of one-way ANOVA is easily calculated by a computer. The p-value is calculated from the F null sampling distribution with matching degrees of freedom. But only if we believe that the assumptions of the model are (approximately) correct should we believe that the p-value was calculated from the correct sampling distribution, and it is then valid.