15. Analysis of Variance

A. Introduction

B. ANOVA Designs

C. One-Factor ANOVA (Between-Subjects)

D. Multi-Factor ANOVA (Between-Subjects)

E. Unequal Sample Sizes

F. Tests Supplementing ANOVA

G. Within-Subjects ANOVA

Introduction

by David M. Lane

Prerequisites

- Chapter 3: Variance
- Chapter 11: Significance Testing
- Chapter 12: All Pairwise Comparisons among Means

Learning Objectives

1. What null hypothesis is tested by ANOVA
2. Describe the uses of ANOVA

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called "Analysis of Variance" rather than "Analysis of Means." As you will see, the name is appropriate because inferences about means are made by analyzing variance.

ANOVA is used to test general rather than specific differences among means. This can be seen best by example. In the case study "Smiles and Leniency," the effect of different types of smiles on the leniency shown to a person was investigated. Four different types of smiles (neutral, false, felt, miserable) were investigated. The chapter "All Pairwise Comparisons among Means" showed how to test differences among means. The results from the Tukey HSD test are shown in Table 1.

Table 1. Six pairwise comparisons.

Comparison	Mi-Mj	Q	p
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.010
Felt - Miserable	0.00	0.00	1.000
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

Notice that the only significant difference is between the False and Neutral conditions.

ANOVA tests the non-specific null hypothesis that all four population means are equal. That is

$$\mu_{false} = \mu_{felt} = \mu_{miserable} = \mu_{neutral}.$$

This non-specific null hypothesis is sometimes called the omnibus null hypothesis. When the omnibus null hypothesis is rejected, the conclusion is that at least one population mean is different from at least one other mean. However, since the ANOVA does not reveal which means are different from which, it offers less specific information than the Tukey HSD test. The Tukey HSD is therefore preferable to ANOVA in this situation. Some textbooks introduce the Tukey test only as a follow-up to an ANOVA. However, there is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA.

You might be wondering why you should learn about ANOVA when the Tukey test is better. One reason is that there are complex types of analyses that can be done with ANOVA and not with the Tukey test. A second is that ANOVA is by far the most commonly-used technique for comparing means, and it is important to understand ANOVA in order to understand research reports.

Analysis of Variance Designs

by David M. Lane

Prerequisites

• Chapter 15: Introduction to ANOVA

Learning Objectives

1. Be able to identify the factors and levels of each factor from a description of an experiment
2. Determine whether a factor is a between-subjects or a within-subjects factor
3. Define factorial design

There are many types of experimental designs that can be analyzed by ANOVA. This section discusses many of these designs and defines several key terms used.

Factors and Levels

The section on variables defined an independent variable as a *variable* manipulated by the experimenter. In the case study "Smiles and Leniency," the effect of different types of smiles on the leniency showed to a person was investigated. Four different types of smiles (neutral, false, felt, miserable, on leniency) were shown. In this experiment, "Type of Smile" is the independent variable. In describing an ANOVA design, the term factor is a synonym of independent variable. Therefore, "Type of Smile" is the factor in this experiment. Since four types of smiles were compared, the factor "Type of Smile" has four *levels*.

An ANOVA conducted on a design in which there is only one factor is called a *one-way ANOVA*. If an experiment has two factors, then the ANOVA is called a *two-way ANOVA*. For example, suppose an experiment on the effects of age and gender on reading speed were conducted using three age groups (8 years, 10 years, and 12 years) and the two genders (male and female). The factors would be age and gender. Age would have three levels and gender would have two levels.

Between- and Within-Subjects Factors

In the "Smiles and Leniency" study, the four levels of the factor "Type of Smile" were represented by four separate groups of subjects. When different subjects are used for the levels of a factor, the factor is called a *between-subjects factor* or a *between-subjects variable*. The term "between subjects" reflects the fact that comparisons are between different groups of subjects.

In the "ADHD Treatment" study, every subject was tested with each of four dosage levels (0, 0.15, 0.30, 0.60 mg/kg) of a drug. Therefore there was only one group of subjects, and comparisons were not between different groups of subjects but between conditions within the same subjects. When the same subjects are used for the levels of a factor, the factor is called a *within-subjects factor* or a *within-subjects variable*. Within-subjects variables are sometimes referred to as repeated-measures variables since there are repeated measurements of the same subjects.

Multi-Factor Designs

It is common for designs to have more than one factor. For example, consider a hypothetical study of the effects of age and gender on reading speed in which males and females from the age levels of 8 years, 10 years, and 12 years are tested. There would be a total of six different groups as shown in Table 1.

Table 1. Gender x Age Design

Group	Gender	Age
1	Female	8
2	Female	10
3	Female	12
4	Male	8
5	Male	10
6	Male	12

This design has two factors: age and gender. Age has three levels and gender has two levels. When all combinations of the levels are included (as they are here), the design is called a *factorial design*. A concise way of describing this design is as a Gender (2) x Age (3) factorial design where the numbers in parentheses indicate

the number of levels. Complex designs frequently have more than two factors and may have combinations of between- and within-subjects factors.

One-Factor ANOVA (Between Subjects)

by David M. Lane

Prerequisites

- Chapter 3: Variance
- Chapter 7: Introduction to Normal Distributions
- Chapter 11: Significance Testing
- Chapter 11: One- and Two-Tailed Tests
- Chapter 12: t Test of Differences Between Groups
- Chapter 15: Introduction to ANOVA
- Chapter 15: ANOVA Designs

Learning Objectives

1. State what the Mean Square Error (MSE) estimates when the null hypothesis is true and when the null hypothesis is false
2. State what the Mean Square Between (MSB) estimates when the null hypothesis is true and when the null hypothesis is false
3. State the assumptions of a one-way ANOVA
4. Compute MSE
5. Compute MSB
6. Compute F and its two degrees of freedom parameters
7. Describe the shape of the F distribution
8. Explain why ANOVA is best thought of as a two-tailed test even though literally only one tail of the distribution is used
9. State the relationship between the t and F distributions
10. Partition the sums of squares into conditions and error
11. Format data to be used with a computer statistics program

This section shows how ANOVA can be used to analyze a one-factor between-subjects design. We will use as our main example the "Smiles and Leniency" case study. In this study there were four conditions with 34 subjects in each condition. There was one score per subject. The null hypothesis tested by ANOVA is that the population means for all conditions are the same. This can be expressed as follows:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$

where H_0 is the null hypothesis and k is the number of conditions. In the smiles and leniency study, k = 4 and the null hypothesis is

$$H_0: \mu_{false} = \mu_{felt} = \mu_{miserable} = \mu_{neutral}.$$

If the null hypothesis is rejected, then it can be concluded that at least one of the population means is different from at least one other population mean.

Analysis of variance is a method for testing differences among means by analyzing variance. The test is based on two estimates of the population variance (σ^2). One estimate is called the mean square error (MSE) and is based on differences among scores within the groups. MSE estimates σ^2 regardless of whether the null hypothesis is true (the population means are equal). The second estimate is called the mean square between (MSB) and is based on differences among the sample means. MSB only estimates σ^2 if the population means are equal. If the population means are not equal, then MSB estimates a quantity larger than σ^2. Therefore, if the MSB is much larger than the MSE, then the population means are unlikely to be equal. On the other hand, if the MSB is about the same as MSE, then the data are consistent with the hypothesis that the population means are equal.

Before proceeding with the calculation of MSE and MSB, it is important to consider the assumptions made by ANOVA:

1. The populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the values are not independent. The analysis of data with two scores per subject is shown in the section on within-subjects ANOVA later in this chapter.

These assumptions are the same as for a t test of differences between groups except that they apply to two or more groups, not just to two groups.

The means and variances of the four groups in the "Smiles and Leniency" case study are shown in Table 1. Note that there are 34 subjects in each of the four conditions (False, Felt, Miserable, and Neutral).

Table 1. Means and Variances from “Smiles and Leniency” Study

Condition	Mean	Variance
FALSE	5.3676	3.3380
Felt	4.9118	2.8253
Miserable	4.9118	2.1132
Neutral	4.1176	2.3191

Sample Sizes

The first calculations in this section all assume that there is an equal number of observations in each group. Unequal sample size calculations are shown in the section on sources of variation. We will refer to the number of observations in each group as n and the total number of observations as N. For these data there are four groups of 34 observations. Therefore n = 34 and N = 136.

Computing MSE

Recall that the assumption of homogeneity of variance states that the variance within each of the populations (σ^2) is the same. This variance, σ^2, is the quantity estimated by MSE and is computed as the mean of the sample variances. For these data, the MSE is equal to 2.6489.

Computing MSB

The formula for MSB is based on the fact that the variance of the sampling distribution of the mean is

$$\sigma_M^2 = \frac{\sigma^2}{n}$$

where n is the sample size of each group. Rearranging this formula, we have

$$\sigma^2 = n\sigma_M^2$$

Therefore, if we knew the variance of the sampling distribution of the mean, we could compute σ^2 by multiplying it by n. Although we do not know the variance of the sampling distribution of the mean, we can estimate it with the variance of the

sample means. For the leniency data, the variance of the four sample means is 0.270. To estimate σ^2, we multiply the variance of the sample means (0.270) by n (the number of observations in each group, which is 34). We find that MSB = 9.179.

To sum up these steps:

1. Compute the means.
2. Compute the variance of the means.
3. Multiply the variance of the means by n.

Recap

If the population means are equal, then both MSE and MSB are estimates of σ^2 and should therefore be about the same. Naturally, they will not be exactly the same since they are just estimates and are based on different aspects of the data: The MSB is computed from the sample means and the MSE is computed from the sample variances.

If the population means are not equal, then MSE will still estimate σ^2 because differences in population means do not affect variances. However, differences in population means affect MSB since differences among population means are associated with differences among sample means. It follows that the larger the differences among sample means, the larger the MSB. In short, MSE estimates σ^2 whether or not the population means are equal, whereas MSB estimates σ^2 only when the population means are equal and estimates a larger quantity when they are not equal.

Comparing MSE and MSB

The critical step in an ANOVA is comparing MSE and MSB. Since MSB estimates a larger quantity than MSE only when the population means are not equal, a finding of a larger MSB than an MSE is a sign that the population means are not equal. But since MSB could be larger than MSE by chance even if the population means are equal, MSB must be much larger than MSE in order to justify the conclusion that the population means differ. But how much larger must MSB be? For the "Smiles and Leniency" data, the MSB and MSE are 9.179 and 2.649, respectively. Is that difference big enough? To answer, we would need to know the probability of getting that big a difference or a bigger difference if the population means were all equal. The mathematics necessary to answer this question were

worked out by the statistician R. Fisher. Although Fisher's original formulation took a slightly different form, the standard method for determining the probability is based on the ratio of MSB to MSE. This ratio is named after Fisher and is called the F ratio.

For these data, the F ratio is

```
F = 9.179/2.649 = 3.465.
```

Therefore, the MSB is 3.465 times higher than MSE. Would this have been likely to happen if all the population means were equal? That depends on the sample size. With a small sample size, it would not be too surprising because results from small samples are unstable. However, with a very large sample, the MSB and MSE are almost always about the same, and an F ratio of 3.465 or larger would be very unusual. Figure 1 shows the *sampling distribution* of F for the sample size in the "Smiles and Leniency" study. As you can see, it has a positive skew.

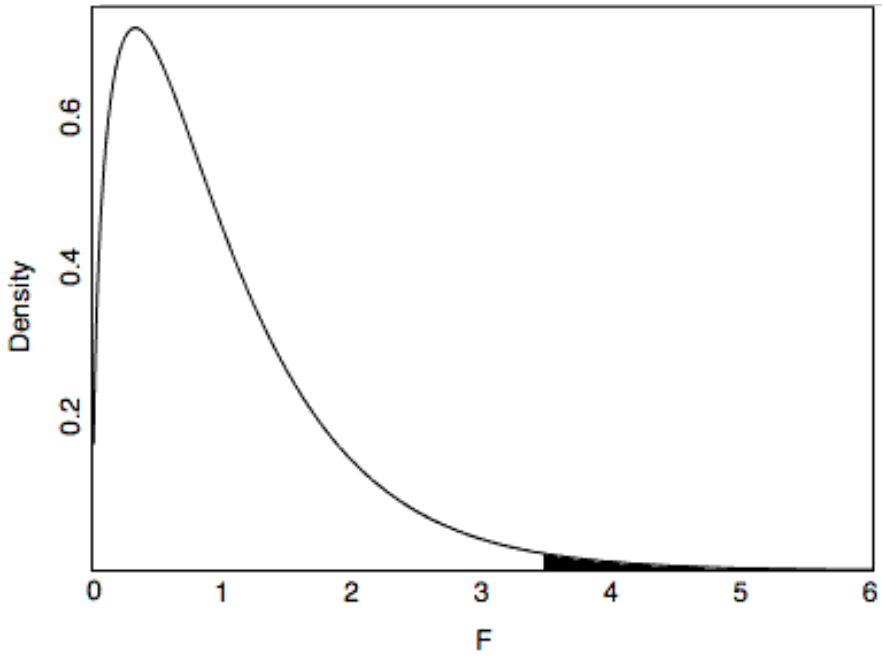


Figure 1. Distribution of F.

From Figure 1, you can see that F ratios of 3.465 or above are unusual occurrences. The area to the right of 3.465 represents the probability of an F that large or larger and is equal to 0.018 and therefore the null hypothesis can be rejected. The conclusion that at least one of the population means is different from at least one of the others is justified.

The shape of the F distribution depends on the sample size. More precisely, it depends on two degrees of freedom (df) parameters: one for the numerator (MSB) and one for the denominator (MSE). Recall that the degrees of freedom for an estimate of variance is equal to the number of observations minus one. Since the MSB is the variance of k means, it has k - 1 df. The MSE is an average of k variances, each with n-1 df. Therefore, the df for MSE is k(n - 1) = N - k. where N is the total number of observations, n is the number of observations in each group, and k is the number of groups. To summarize:

$df_{numerator} = k-1$
$df_{denominator} = N-k$

For the "Smiles and Leniency" data,

$df_{numerator} = k-1 = 4-1 = 3$
$df_{denominator} = N-k = 136-4 = 132$
$F = 3.465$

The F distribution calculator shows that $p = 0.018$.

One-Tailed or Two?

Is the probability value from an F ratio a one-tailed or a two-tailed probability? In the literal sense, it is a one-tailed probability since, as you can see in Figure 1, the probability is the area in the right-hand tail of the distribution. However, the F ratio is sensitive to any pattern of differences among means. It is, therefore, a test of a two-tailed hypothesis and is best considered a two-tailed test.

Relationship to the t test

Since an ANOVA and an independent-groups t test can both test the difference between two means, you might be wondering which one to use. Fortunately, it does not matter since the results will always be the same. When there are only two groups, the following relationship between F and t will always hold:

$$F(1,dfd) = t^2(df)$$

where dfd is the degrees of freedom for the denominator of the F test and df is the degrees of freedom for the t test. dfd will always equal df.

Sources of Variation

Why do scores in an experiment differ from one another? Consider the scores of two subjects in the "Smiles and Leniency" study: one from the "False Smile" condition and one from the "Felt Smile" condition. An obvious possible reason that the scores could differ is that the subjects were treated differently (they were in different conditions and saw different stimuli). A second reason is that the two subjects may have differed with regard to their tendency to judge people leniently. A third is that, perhaps, one of the subjects was in a bad mood after receiving a low grade on a test. You can imagine that there are innumerable other reasons why the scores of the two subjects could differ. All of these reasons except the first (subjects were treated differently) are possibilities that were not under experimental investigation and, therefore, all of the differences (variation) due to these possibilities are unexplained. It is traditional to call unexplained variance error even though there is no implication that an error was made. Therefore, the variation in this experiment can be thought of as being either variation due to the condition the subject was in or due to error (the sum total of all reasons the subjects' scores could differ that were not measured).

One of the important characteristics of ANOVA is that it partitions the variation into its various sources. In ANOVA, the term *sum of squares* (SSQ) is used to indicate variation. The total variation is defined as the sum of squared differences between each score and the mean of all subjects. The mean of all subjects is called the grand mean and is designated as GM. (When there is an equal number of subjects in each condition, the grand mean is the mean of the condition means.) The total sum of squares is defined as

$$SSQ_{total} = \sum (X - GM)^2$$

which means to take each score, subtract the grand mean from it, square the difference, and then sum up these squared values. For the "Smiles and Leniency" study, $SSQ_{total} = 377.19$.

The sum of squares condition is calculated as shown below.

$$SSQ_{condition} = n \sum (M_1 - GM)^2 + (M_2 - GM)^2 + \cdots + (M_k - GM)^2$$

where n is the number of scores in each group, k is the number of groups, M_1 is the mean for Condition 1, M_2 is the mean for Condition 2, and M_k is the mean for Condition k. For the "Smiles and Leniency" study, the values are:

$$SSQ_{condition} = 34[(5.37-4.83)^2 + (4.91-4.83)^2 + (4.91-4.83)^2 + (4.12-4.83)^2]$$

$$= 27.5$$

If there are unequal sample sizes, the only change is that the following formula is used for the sum of squares condition:

$$SSQ_{condition} = \sum n_1(M_1 - GM)^2 + n_2(M_2 - GM)^2 + \cdots + n_k(M_k - GM)^2$$

where n_i is the sample size of the i^{th} condition. SSQ_{total} is computed the same way as shown above.

The sum of squares error is the sum of the squared deviations of each score from its group mean. This can be written as

$$SSQ_{error} = \sum (X_{i1} - M_1)^2 + \sum (X_{i2} - M_2)^2 + \cdots + \sum (X_{ik} - M_k)^2$$

where X_{i1} is the i^{th} score in group 1 and M_1 is the mean for group 1, X_{i2} is the i^{th} score in group 2 and M_2 is the mean for group 2, etc. For the "Smiles and Leniency" study, the means are: 5.368, 4.912, 4.912, and 4.118. The SSQ_{error} is therefore:

$$(2.5-5.368)^2 + (5.5-5.368)^2 + \ldots + (6.5-4.118)^2 = 349.65$$

The sum of squares error can also be computed by subtraction:

$$SSQ_{error} = SSQ_{total} - SSQ_{condition}$$

$$SSQ_{error} = 377.189 - 27.535 = 349.65.$$

Therefore, the total sum of squares of 377.19 can be partitioned into $SSQ_{condition}$ (27.53) and SSQ_{error} (349.66).

Once the sums of squares have been computed, the mean squares (MSB and MSE) can be computed easily. The formulas are:

$$MSB = SSQ_{condition}/dfn$$

where dfn is the degrees of freedom numerator and is equal to k - 1 = 3.

$$MSB = 27.535/3 = 9.18$$

which is the same value of MSB obtained previously (except for rounding error). Similarly,

$$MSE = SSQ_{error}/dfn$$

where dfd is the degrees of freedom for the denominator and is equal to N - k.

$$dfd = 136 - 4 = 132$$

$$MSE = 349.66/132 = 2.65$$

which is the same as obtained previously (except for rounding error). Note that the dfd is often called the dfe for *degrees of freedom error*.

The Analysis of Variance Summary Table shown below is a convenient way to summarize the partitioning of the variance. The rounding errors have been corrected.

Table 2. ANOVA Summary Table

Source	df	SSQ	MS	F	p
Condition	3	27.5349	9.1783	3.465	0.0182
Error	132	349.6544	2.6489		
Total	135	377.1893			

The first column shows the sources of variation, the second column shows the degrees of freedom, the third shows the sums of squares, the fourth shows the

mean squares, the fifth shows the F ratio, and the last shows the probability value. Note that the mean squares are always the sums of squares divided by degrees of freedom. The F and p are relevant only to Condition. Although the mean square total could be computed by dividing the sum of squares by the degrees of freedom, it is generally not of much interest and is omitted here.

Formatting data for Computer Analysis

Most computer programs that compute ANOVAs require your data to be in a specific form. Consider the data in Table 3.

Table 3. Example Data

Group 1	Group 2	Group 3
3	2	8
4	4	5
5	6	5

Here there are three groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. The reformatted version of the data in Table 3 is shown in Table 4.

Table 4. Reformatted Data

G	Y
1	3
1	4
1	5
2	2
2	4
2	6
3	8
3	5
3	5

Multi-Factor Between-Subjects Designs

by David M. Lane

Prerequisites
- Chapter 15: Introduction to ANOVA
- Chapter 15: ANOVA Designs

Learning Objectives
1. Define main effect, simple effect, interaction, and marginal mean
2. State the relationship between simple effects and interaction
3. Compute the source of variation and df for each effect in a factorial design
4. Plot the means for an interaction
5. Define three-way interaction

Basic Concepts and Terms

In the "Bias Against Associates of the Obese" case study, the researchers were interested in whether the weight of a companion of a job applicant would affect judgments of a male applicant's qualifications for a job. Two *independent variables* were investigated: (1) whether the companion was obese or of typical weight and (2) whether the companion was a girlfriend or just an acquaintance. One approach could have been to conduct two separate studies, one with each independent variable. However, it is more efficient to conduct one study that includes both independent variables. Moreover, there is a much bigger advantage than efficiency for including two variables in the same study: it allows a test of the *interaction* between the variables. There is an interaction when the effect of one variable differs depending on the *level* of a second variable. For example, it is possible that the effect of having an obese companion would differ depending on the relationship to the companion. Perhaps there is more prejudice against a person with an obese companion if the companion is a girlfriend than if she is just an acquaintance. If so, there would be an interaction between the obesity factor and the relationship factor.

There are three effects of interest in this experiment:

1. Weight: Are applicants judged differently depending on the weight of their companion?
2. Relationship: Are applicants judged differently depending on their relationship with their companion?

3. Weight x Relationship Interaction: Does the effect of weight differ depending on the relationship with the companion?

The first two effects (Weight and Relationship) are both *main effects*. A main effect of an independent variable is the effect of the variable averaging over the levels of the other variable(s). It is convenient to talk about main effects in terms of *marginal means*. A marginal mean for a level of a variable is the mean of the means of all levels of the other variable. For example, the marginal mean for the level "Obese" is the mean of "Girlfriend Obese" and "Acquaintance Obese." Table 1 shows that this marginal mean is equal to the mean of 5.65 and 6.15, which is 5.90. Similarly, the marginal mean for the level "Typical" is the mean of 6.19 and 6.59, which is 6.39. The main effect of Weight is based on a comparison of these two marginal means. Similarly, the marginal means for "Girlfriend" and "Acquaintance" are 5.92 and 6.37..

Table 1. Means for All Four Conditions

		Companion Weight		
		Obese	Typical	Marginal Mean
Relationship	Girlfriend	5.65	6.19	5.92
	Acquaintance	6.15	6.59	6.37
	Marginal Mean	5.90	6.39	

In contrast to a main effect, which is the effect of a variable averaged across levels of another variable, the simple effect of a variable is the effect of the variable at a single level of another variable. The simple effect of Weight at the level of "Girlfriend" is the difference between the "Girlfriend Typical" and the "Girlfriend Obese" conditions. The difference is 6.19-5.65 = 0.54. Similarly, the simple effect of Weight at the level of "Acquaintance" is the difference between the "Acquaintance Typical" and the "Acquaintance Obese" conditions. The difference is 6.59-6.15 = 0.44.

Recall that there is an interaction when the effect of one variable differs depending on the level of another variable. This is equivalent to saying that there is an interaction when the simple effects differ. In this example, the simple effects of

weight are 0.54 and 0.44. As shown below, these simple effects are not significantly different.

Tests of Significance

The important questions are not whether there are main effects and interactions in the sample data. Instead, what is important is what the sample data allow you to conclude about the population. This is where Analysis of Variance comes in. ANOVA tests main effects and interactions for *significance*. An ANOVA Summary Table for these data is shown in Table 2.

Table 2. ANOVA Summary Table

Source	df	SSQ	MS	F	p
Weight	1	10.4673	10.4673	6.214	0.0136
Relation	1	8.8144	8.8144	5.233	0.0234
W x R	1	0.1038	0.1038	0.062	0.8043
Error	172	289.7132	1.6844		
Total	175	310.1818			

Consider first the effect of "Weight." The *degrees of freedom* (df) for "Weight" is 1. The degrees of freedom for a main effect is always equal to the number of levels of the variable minus one. Since there are two levels of the "Weight" variable (typical and obese), the df is 2 - 1 = 1. We skip the calculation of the sum of squares (SSQ) not because it is difficult, but because it is so much easier to rely on computer programs to compute it. The mean square (MS) is the sum of squares divided by the df. The F ratio is computed by dividing the MS for the effect by the MS for error (MSE). For the effect of "Weight," F = 10.4673/1.6844 = 6.214. The last column, p, is the probability of getting an F of 6.214 or larger given that there is no effect of weight in the population. The p value is 0.0136 and therefore the null *hypothesis* of no main effect of "Weight" is rejected. The conclusion is that being accompanied by an obese companion lowers judgments of qualifications.

The effect "Relation" is interpreted the same way. The conclusion is that being accompanied by a girlfriend leads to lower ratings than being accompanied by an acquaintance.

The df for an interaction is the product of the df's of variables in the interaction. For the "Weight x Relation" interaction (W x R), the df = 1 since both

Weight and Relation have one df: 1 x 1 = 1. The p value for the interaction is 0.8043, which is the probability of getting an interaction as big or bigger than the one obtained in the experiment if there were no interaction in the population. Therefore, these data provide no evidence for an interaction. Always keep in mind that the lack of evidence for an effect does not justify the conclusion that there is no effect. In other words, you do not accept the null hypothesis just because you do not reject it.

For "Error," the degrees of freedom is equal to the total number of observations minus the total number of groups. The sample sizes of the four conditions in this experiment are shown in Table 3. The total number of observations is 40 + 42 + 40 + 54 = 176. Since there are four groups, dfe = 176 - 4 = 172.

Table 3. Sample Sizes for All Four Conditions

		Companion Weight	
		Obese	Typical
Relationship	Girlfriend	40	42
	Acquaintance	40	54

The final row in the ANOVA Summary Table is "Total." The degrees of freedom total is equal to the sum of all degrees of freedom. It is also equal to the number of observations minus 1, or 176 - 1 = 175. When there are equal sample sizes, the sum of squares total will equal the sum of all other sums of squares. However, when there are unequal sample sizes, as there are here, this will not generally be true. The reasons for this are complex and are discussed in the section Unequal Sample Sizes.

Plotting Means

Although the plot shown in Figure 1 illustrates the main effects as well as the interaction (or lack of an interaction), it is called an *interaction plot*. It is important to consider the components of this plot carefully. First, the dependent variable is on the Y-axis. Second, one of the independent variables is on the X-axis. In this case, it is the variable "Weight." Finally, a separate line is drawn for each level of the other independent variable. It is better to label the lines right on the graph, as shown here, than with a legend.

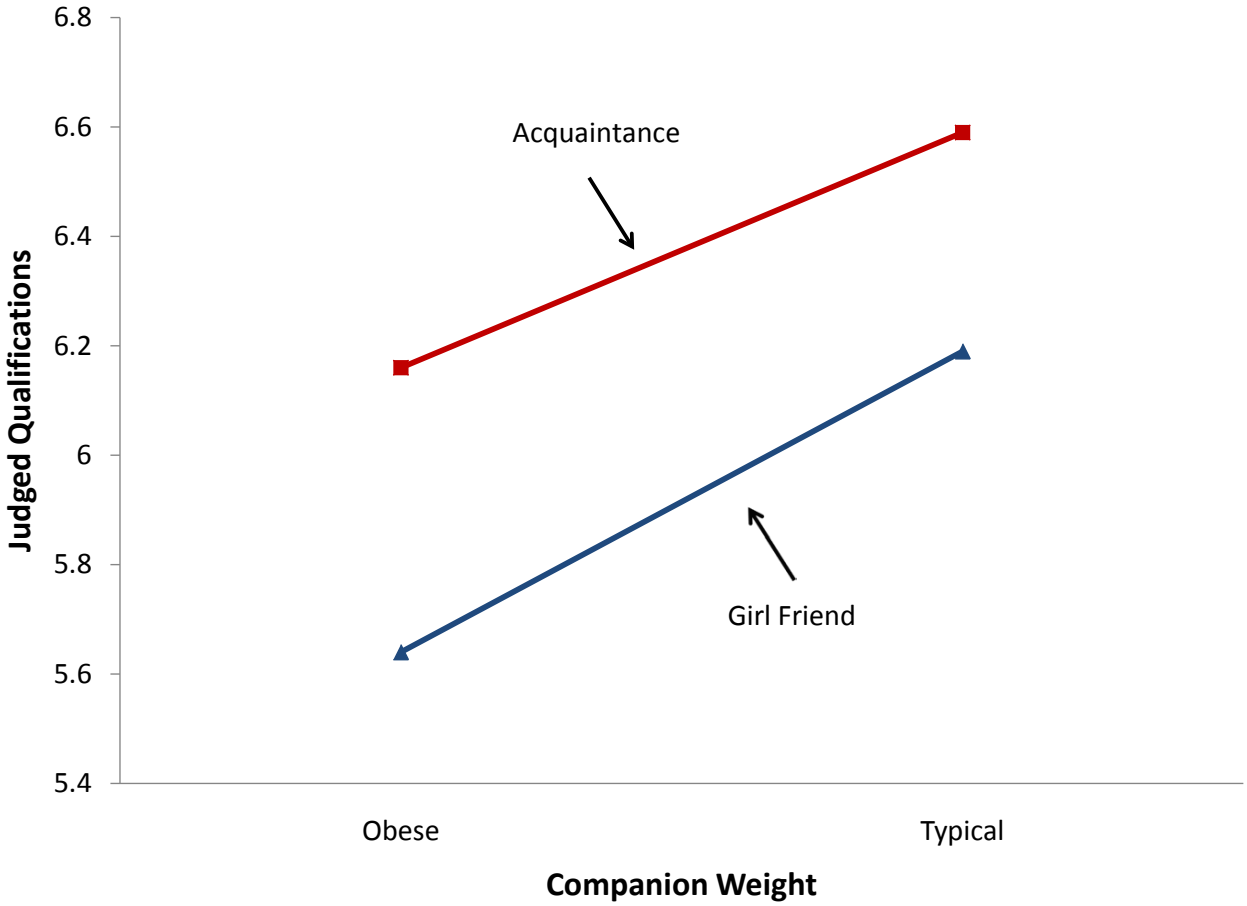


Figure 1. An interaction plot.

If you have three or more levels on the X-axis, you should not use lines unless there is some numeric ordering to the levels. If your variable on the X-axis is a qualitative variable, you can use a plot such as the one in Figure 2. However, as discussed in the section on bar charts, it would be better to replace each bar with a *box plot*.

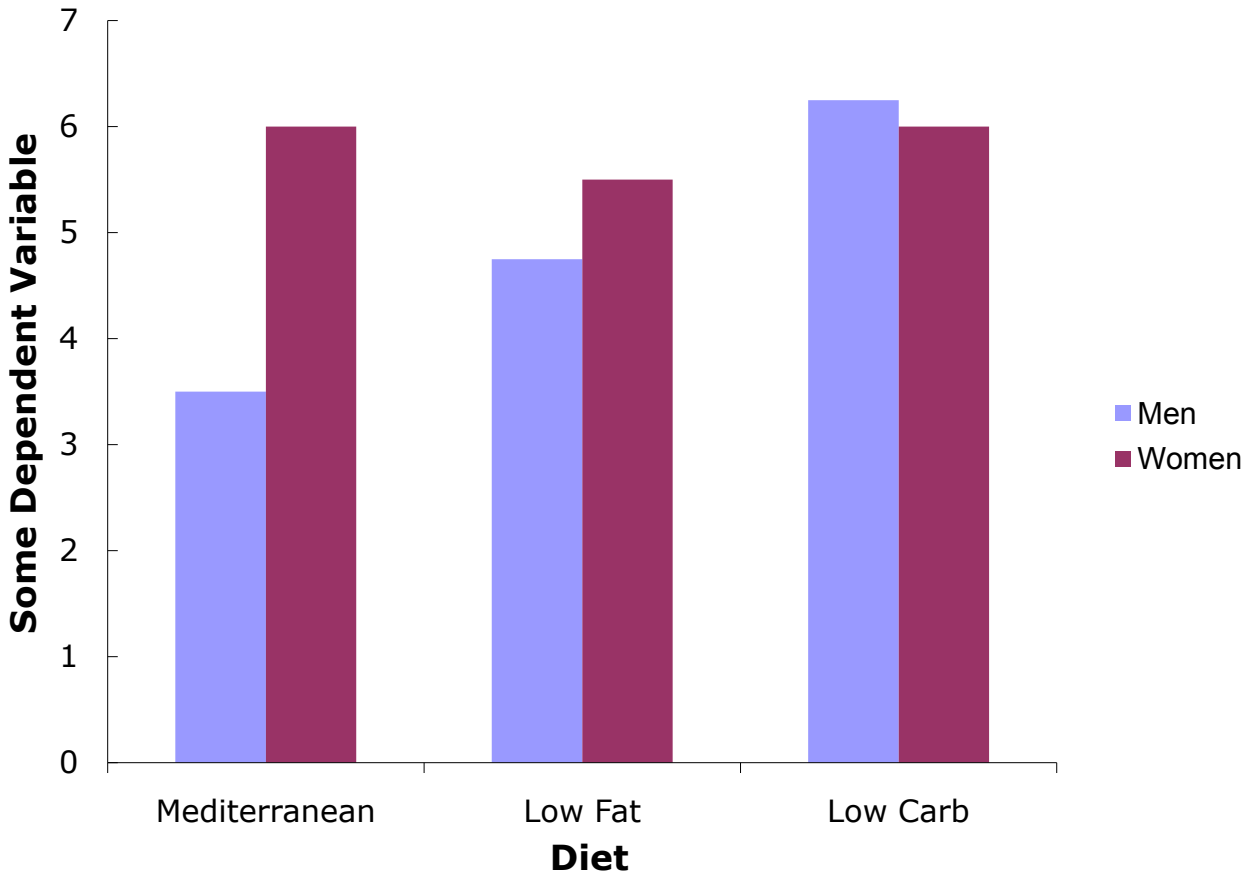


Figure 2. Plot with a qualitative variable on the X-axis.

Figure 3 shows such a plot. Notice how it contains information about the medians, quantiles, and minimums and maximums not contained in Figure 2. Most important, you get an idea about how much the distributions overlap from Figure 3 which you do not get from Figure 2.

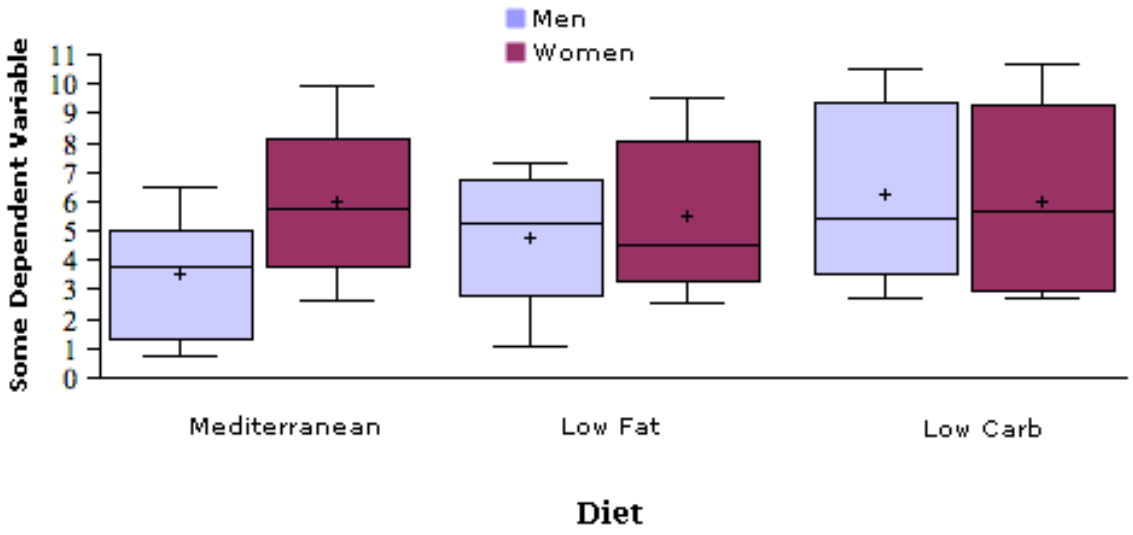


Figure 3. Box plots.

Line graphs are a good option when there are more than two levels of a numeric variable. Figure 4 shows an example. A line graph has the advantage of showing the pattern of interaction clearly. Its disadvantage is that it does not convey the distributional information contained in box plots.

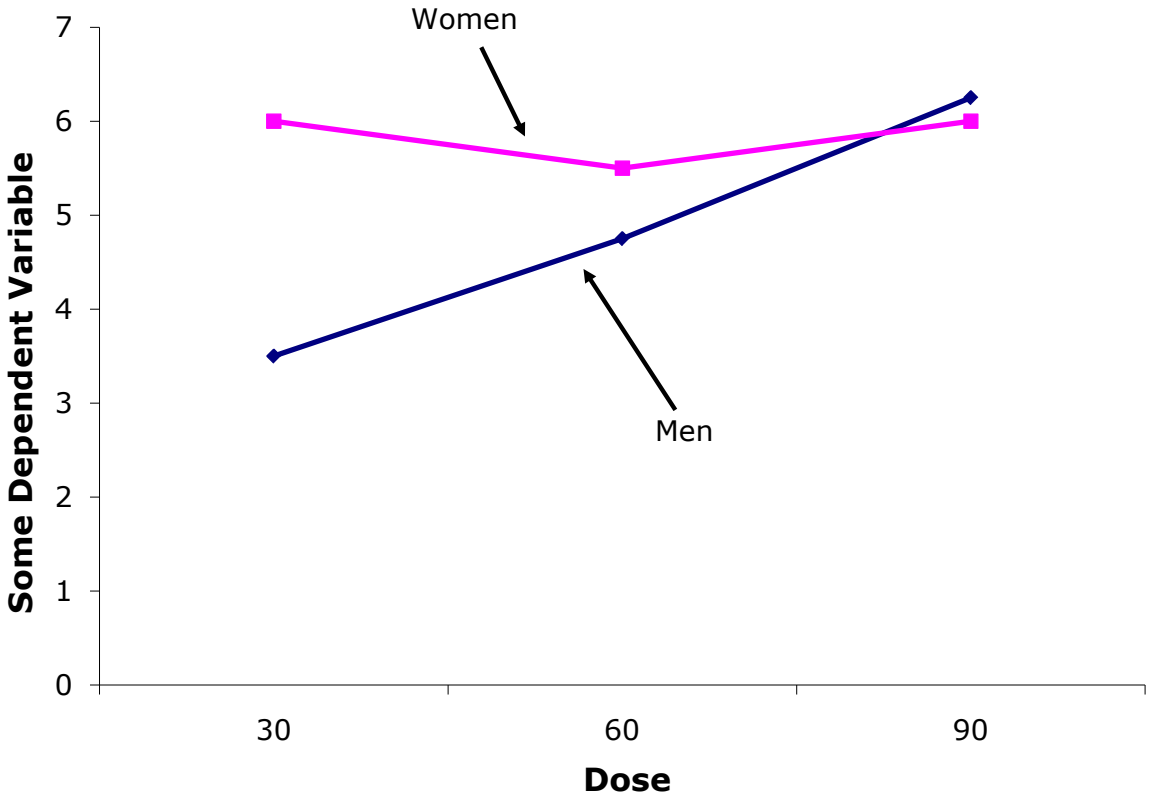


Figure 4. Plot with a quantitative variable on the X-axis.

An Example with Interaction

The following example was presented in the section on specific comparisons among means. It is also relevant here.

This example uses the made-up data from a hypothetical experiment shown in Table 4. Twelve subjects were selected from a population of high-self-esteem subjects and an additional 12 subjects were selected from a population of low-self-esteem subjects. Subjects then performed on a task and (independent of how well they really did) half in each esteem category were told they succeeded and the other half were told they failed. Therefore, there were six subjects in each of the four esteem/outcome combinations and 24 subjects in all.

After the task, subjects were asked to rate (on a 10-point scale) how much of their outcome (success or failure) they attributed to themselves as opposed to being due to the nature of the task.

Table 4. Data from Hypothetical Experiment on Attribution

		Esteem	
		High	Low
Outcome	Success	7	6
		8	5
		7	7
		8	4
		9	5
		5	6
	Failure	4	9
		6	8
		5	9
		4	8
		7	7
		3	6

The ANOVA Summary Table for these data is shown in Table 5.

Table 5. ANOVA Summary Table for Made-Up Data

Source	df	SSQ	MS	F	p
Outcome	1	0.0417	0.0417	0.0256	0.8744
Esteem	1	2.0417	2.0417	1.2564	0.2756
O x E	1	35.0417	35.0417	21.5641	0.0002
Error	20	32.5000	1.6250		
Total	23	69.6250			

As you can see, the only significant effect is the Outcome x Esteem (O x E) interaction. The form of the interaction can be seen in Figure 5.

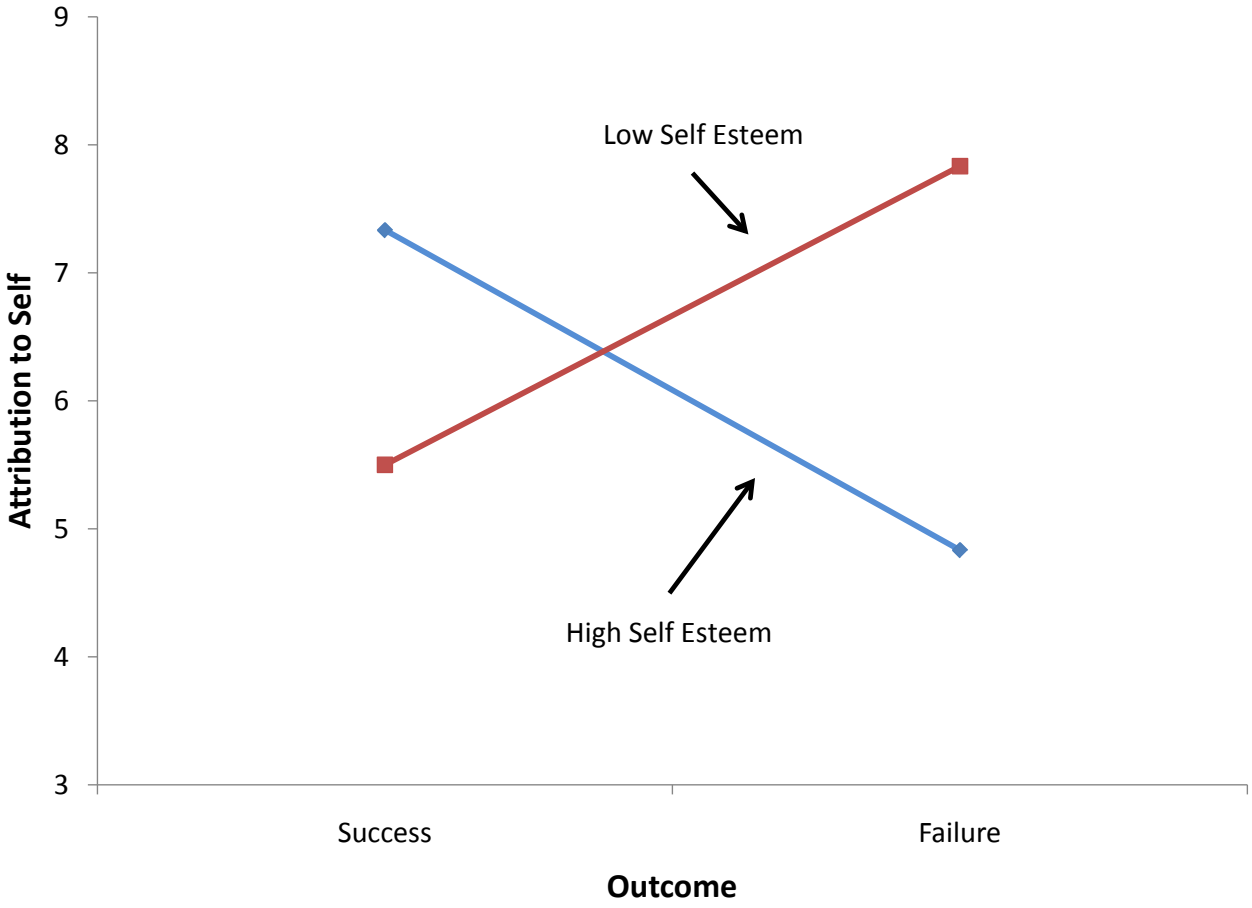


Figure 5. Interaction plot for made-up data.

Clearly the effect of "Outcome" is different for the two levels of "Esteem": For subjects high in self-esteem, failure led to less attribution to oneself than did

success. By contrast, for subjects low in self-esteem, failure led to more attribution to oneself than did success. Notice that the two lines in the graph are not parallel. Nonparallel lines indicate interaction. The significance test for the interaction determines whether it is justified to conclude that the lines in the population are not parallel. Lines do not have to cross for there to be an interaction.

Three-Factor Designs

Three-factor designs are analyzed in much the same way as two-factor designs. Table 6 shows the analysis of a study described by Franklin and Cooley (2002) investigating three factors on the strength of industrial fans: (1) Hole Shape (Hex or Round), (2) Assembly Method (Staked or Spun), and (3) Barrel Surface (Knurled or Smooth). The dependent variable, Breaking Torque, was measured in foot-pounds. There were eight observations in each of the eight combinations of the three factors.

As you can see in Table 6, there are three main effects, three two-way interactions, and one three-way interaction. The degrees of freedom for the main effects are, as in a two-factor design, equal to the number of levels of the factor minus one. Since all the factors here have two levels, all the main effects have one degree of freedom. The interaction degrees of freedom is always equal to the product of the degrees of freedom of the component parts. This holds for the three-factor interaction as well as for the two-factor interactions. The error degrees of freedom is equal to the number of observations (64) minus the number of groups (8) and equals 56.

Table 6. ANOVA Summary Table for Fan Data

Source	df	SSQ	MS	F	p
Hole	1	8258.27	8258.27	266.68	<0.0001
Assembly	1	13369.14	13369.14	431.73	<0.0001
H x A	1	2848.89	2848.89	92.00	<0.0001
Barrel	1	35.0417	35.0417	21.5641	<0.0001
H x B	1	594.14	594.14	19.1865	<0.0001
A x B	1	135.14	135.14	4.36	0.0413
H x A x B	1	1396.89	1396.89	45.11	<0.0001

Error	56	1734.12	30.97		
Total	63	221386.91			

A three-way interaction means that the two-way interactions differ as a function of the level of the third variable. The usual way to portray a three-way interaction is to plot the two-way interactions separately. Figure 6 shows the Barrel (Knurled or Smooth) x Assembly (Staked or Spun) separately for the two levels of Hole Shape (Hex or Round). For the Hex Shape, there is very little interaction with the lines being close to parallel with a very slight tendency for the effect of Barrel to be bigger for Staked than for Spun. The two-way interaction for the Round Shape is different: The effect of Barrel is bigger for Spun than for Staked. The finding of a significant three-way interaction indicates that this difference in two-way interactions is significant.

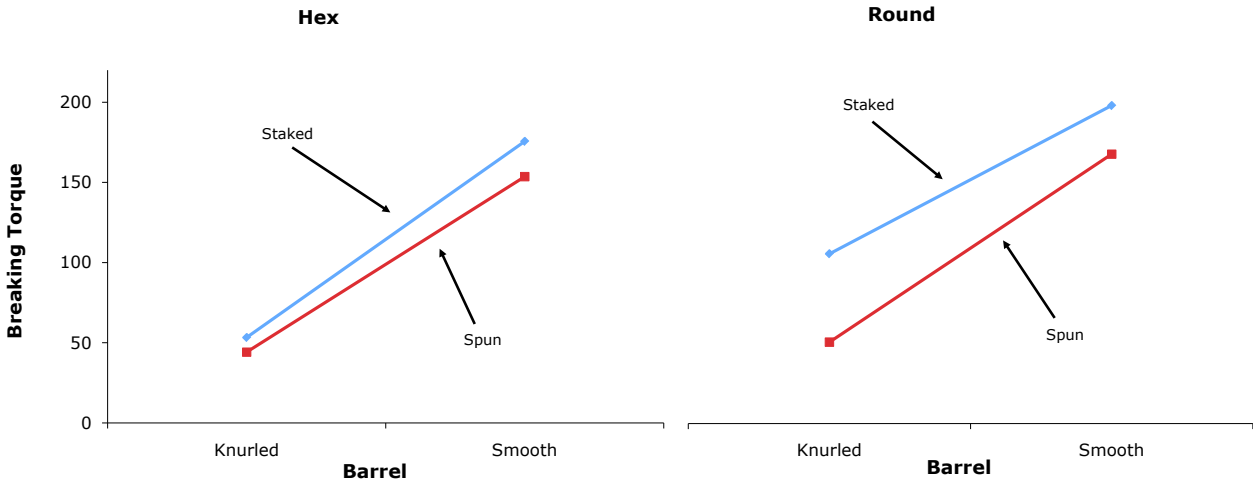


Figure 6. Plot of the three-way interaction.

Formatting Data for Computer Analysis

The data in Table 4 have been reformatted in Table 7. Note how there is one column to indicate the level of outcome and one column to indicate the level of esteem. The coding is as follows:

```
High-self-esteem:1

Low self-esteem: 2
```

```
Success: 1

Failure: 2
```

Table 7. Attribution Data Reformatted

outcome	esteem	attrib
1	1	7
1	1	8
1	1	7
1	1	8
1	1	9
1	1	5
1	2	6
1	2	5
1	2	7
1	2	4
1	2	5
1	2	6
2	1	4
2	1	6
2	1	5
2	1	4
2	1	7
2	1	3
2	2	9
2	2	8
2	2	9
2	2	8
2	2	7
2	2	6

Unequal Sample Sizes

by David M. Lane

Prerequisites

- Chapter 15: ANOVA Designs
- Chapter 15: Multi-Factor Designs

Learning Objectives

1. State why unequal n can be a problem
2. Define confounding
3. Compute weighted and unweighted means
4. Distinguish between Type I and Type III sums of squares
5. Describe why the cause of the unequal sample sizes makes a difference in the interpretation

The Problem of Confounding

Whether by design, accident, or necessity, the number of subjects in each of the conditions in an experiment may not be equal. For example, the sample sizes for the "Bias Against Associates of the Obese" case study are shown in Table 1. Although the sample sizes were approximately equal, the "Acquaintance Typical" condition had the most subjects. Since n is used to refer to the sample size of an individual group, designs with unequal sample sizes are sometimes referred to as designs with unequal n.

Table 1. Sample Sizes for "Bias Against Associates of the Obese" Study.

		Companion Weight	
		Obese	Typical
Relationship	Girl Friend	40	42
	Acquaintance	40	54

We consider an absurd design to illustrate the main problem caused by unequal n. Suppose an experimenter were interested in the effects of diet and exercise on cholesterol. The sample sizes are shown in Table 2.

Table 2. Sample Sizes for "Diet and Exercise" Example.

		Exercise	
		Moderate	**None**
Diet	Low Fat	5	0
	High Fat	0	5

What makes this example absurd is that there are no subjects in either the "Low-Fat No-Exercise" condition or the "High-Fat Moderate-Exercise" condition. The hypothetical data showing change in cholesterol are shown in Table 3.

Table 3. Data for "Diet and Exercise" Example.

		Exercise		
		Moderate	**None**	**Mean**
Diet	Low Fat	-20 -25 -30 -35 -15		-25
	High Fat		-20 6 -10 -6 5	-5
	Mean	-25	-5	-15

The last column shows the mean change in cholesterol for the two diet conditions, whereas the last row shows the mean change in cholesterol for the two Exercise conditions. The value of -15 in the lower-right-most cell in the table is the mean of all subjects.

We see from the last column that those on the low-fat diet lowered their cholesterol an average of 25 units, whereas those on the high-fat diet lowered theirs by only an average of 5 units. However, there is no way of knowing whether the difference is due to diet or to exercise since every subject in the low-fat

condition was in the moderate-exercise condition and every subject in the high-fat condition was in the no-exercise condition. Therefore, Diet and Exercise are completely *confounded*. The problem with unequal n is that it causes confounding.

Weighted and Unweighted Means

The difference between *weighted* and *unweighted means* is a difference critical for understanding how to deal with the confounding resulting from unequal n.

Weighted and unweighted means will be explained using the data shown in Table 4. Here, Diet and Exercise are confounded because 80% of the subjects in the low-fat condition exercised as compared to 20% of those in the high-fat condition. However, there is not complete confounding as there was with the data in Table 3.

The weighted mean for "Low Fat" is computed as the mean of the "Low-Fat Moderate-Exercise" mean and the "Low-Fat No-Exercise" mean, weighted in accordance with sample size. To compute a weighted mean, you multiply each mean by its sample size and divide by N, the total number of observations. Since there are four subjects in the "Low-Fat Moderate-Exercise" condition and one subject in the "Low-Fat No-Exercise" condition, the means are weighted by factors of 4 and 1 as shown below, where M_w is the weighted mean.

$$M_w = \frac{(4)(-27.5) + (1)(-20)}{5} = -26$$

The weighted mean for the low-fat condition is also the mean of all five scores in this condition. Thus if you ignore the factor "Exercise," you are implicitly computing weighted means.

The unweighted mean for the low-fat condition (M_u) is simply the mean of the two means.

$$M_u = \frac{-27.5 - 20}{2} = -23.75$$

Table 4. Data for Diet and Exercise with Partial Confounding Example

		Exercise			
		Moderate	**None**	**Weighted Mean**	**Unweighted Mean**
Diet	Low Fat	-20 -20 -30 -35	-20	-26	-23.750
		M=-27.5	M=-20.0		
	High Fat	-15	6 6 5 -10	-4	-8.125
		M=-15.0	M=-1.25		
	Weighted Mean	-25	-5		
	Unweighted Mean	-21.25	-10.625		

One way to evaluate the *main effect* of Diet is to compare the weighted mean for the low-fat diet (-26) with the weighted mean for the high-fat diet (-4). This difference of -22 is called "the effect of diet ignoring exercise" and is misleading since most of the low-fat subjects exercised and most of the high-fat subjects did not. However, the difference between the unweighted means of -15.625 (-23.75 minus -8.125) is not affected by this confounding and is therefore a better measure of the main effect. In short, weighted means ignore the effects of other variables (exercise in this example) and result in confounding; unweighted means control for the effect of other variables and therefore eliminate the confounding.

Statistical analysis programs use different terms for means that are computed controlling for other effects. SPSS calls them *estimated marginal means*, whereas SAS and SAS JMP call them *least squares means*.

Types of Sums of Squares

When there is unequal n, the sum of squares total is not equal to the sum of the sums of squares for all the other sources of variation. This is because the confounded sums of squares are not apportioned to any source of variation. For the data in Table 4, the sum of squares for Diet is 390.625, the sum of squares for Exercise is 180.625, and the sum of squares confounded between these two factors is 819.375 (the calculation of this value is beyond the scope of this introductory text). In the ANOVA Summary Table shown in Table 5, this large portion of the sums of squares is not apportioned to any source of variation and represents the "missing" sums of squares. That is, if you add up the sums of squares for Diet, Exercise, D x E, and Error, you get 902.625. If you add the confounded sum of squares of 819.375 to this value, you get the total sum of squares of 1722.000. When confounded sums of squares are not apportioned to any source of variation, the sums of squares are called *Type III sums of squares*. Type III sums of squares are, by far, the most common and if sums of squares are not otherwise labeled, it can safely be assumed that they are Type III.

Table 5. ANOVA Summary Table for Type III SSQ

Source	df	SSQ	MS	F	p
Diet	1	390.625	390.625	7.42	0.034
Exercise	1	180.625	180.625	3.43	0.113
D x E	1	15.625	15.625	0.30	0.605
Error	6	315.750	52.625		
Total	9	1722.000			

When all confounded sums of squares are apportioned to sources of variation, the sums of squares are called *Type I sums of squares*. The order in which the confounded sums of squares are apportioned is determined by the order in which the effects are listed. The first effect gets any sums of squares confounded between it and any of the other effects. The second gets the sums of squares confounded between it and subsequent effects, but not confounded with the first effect, etc. The Type I sums of squares are shown in Table 6. As you can see, with Type I sums of squares, the sum of all sums of squares is the total sum of squares.

Table 6. ANOVA Summary Table for Type I SSQ

Source	df	SSQ	MS	F	p
Diet	1	1210.000	1210.000	22.99	0.003
Exercise	1	180.625	180.625	3.43	0.113
D x E	1	15.625	15.625	0.30	0.605
Error	6	315.750	52.625		
Total	9	1722.000			

In *Type II sums of squares*, sums of squares confounded between main effects are not apportioned to any source of variation, whereas sums of squares confounded between main effects and interactions are apportioned to the main effects. In our example, there is no confounding between the D x E interaction and either of the main effects. Therefore, the Type II sums of squares are equal to the Type III sums of squares.

Which Type of Sums of Squares to Use (optional)

Type I sums of squares allow the variance confounded between two main effects to be apportioned to one of the main effects. Unless there is a strong argument for how the confounded variance should be apportioned (which is rarely, if ever, the case), Type I sums of squares are not recommended.

There is not a consensus about whether Type II or Type III sums of squares is to be preferred. On the one hand, if there is no interaction, then Type II sums of squares will be more powerful for two reasons: (1) variance confounded between the main effect and interaction is properly assigned to the main effect and (2) weighting the means by sample sizes gives better estimates of the effects. To take advantage of the greater power of Type II sums of squares, some have suggested that if the interaction is not significant, then Type II sums of squares should be used. Maxwell and Delaney (2003) caution that such an approach could result in a Type II error in the test of the interaction. That is, it could lead to the conclusion that there is no interaction in the population when there really is one. This, in turn, would increase the Type I error rate for the test of the main effect. As a result, their general recommendation is to use Type III sums of squares.

Maxwell and Delaney (2003) recognized that some researchers prefer Type II sums of squares when there are strong theoretical reasons to suspect a lack of

interaction and the p value is much higher than the typical α level of 0.05. However, this argument for the use of Type II sums of squares is not entirely convincing. As Tukey (1991) and others have argued, it is doubtful that any effect, whether a main effect or an interaction, is exactly 0 in the population. Incidentally, Tukey argued that the role of significance testing is to determine whether a confident conclusion can be made about the direction of an effect, not simply to conclude that an effect is not exactly 0.

Finally, if one assumes that there is no interaction, then an ANOVA model with no interaction term should be used rather than Type II sums of squares in a model that includes an interaction term. (Models without interaction terms are not covered in this book).

There are situations in which Type II sums of squares are justified even if there is strong interaction. This is the case because the hypotheses tested by Type II and Type III sums of squares are different, and the choice of which to use should be guided by which hypothesis is of interest. Recall that Type II sums of squares weight cells based on their sample sizes whereas Type III sums of squares weight all cells the same. Consider Figure 1 which shows data from a hypothetical A(2) x B(2) design. The sample sizes are shown numerically and are represented graphically by the areas of the endpoints.

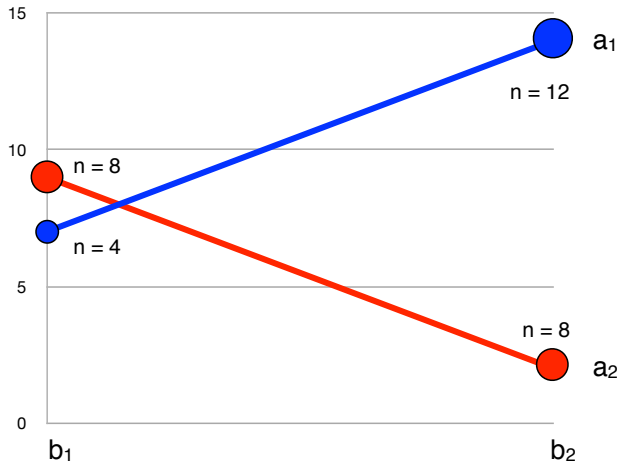


Figure 1. An interaction plot with unequal sample sizes.

First, let's consider the hypothesis for the main effect of B tested by the Type III sums of squares. Type III sums of squares weight the means equally and, for these data, the marginal means for b_1 and b_2 are equal:

$(b_1a_1 + b_1a_2)/2 = (7 + 9)/2 = 8.$

$(b_2a_1 + b_2a_2)/2 = (14 + 2)/2 = 8.$

Thus, there is no main effect of B when tested using Type III sums of squares.

For Type II sums of squares, the means are weighted by sample size. For b_1:

$(4 \times b_1a_1 + 8 \times b_1a_2)/12 =$

$(4 \times 7 + 8 \times 9)/12 = 8.33$

For b_2:

$(12 \times b_2a_1 + 8 \times b_2a_2)/20 =$

$(12 \times 14 + 8 \times 2)/20 = 9.2.$

Since the weighted marginal mean for b_2 is larger than the weighted marginal mean for b_1, there is a main effect of B when tested using Type II sums of squares.

The Type II and Type III analyses are testing different hypotheses. First, let's consider the case in which the differences in sample sizes arise because in the sampling of intact groups, the sample cell sizes reflect the population cell sizes (at least approximately). In this case, it makes sense to weight some means more than others and conclude that there is a main effect of B. This is the result obtained with Type II sums of squares. However, if the sample size differences arose from random assignment, and there just happened to be more observations in some cells than others, then one would want to estimate what the main effects would have been with equal sample sizes and, therefore, weight the means equally. With the means weighted equally, there is no main effect of B, the result obtained with Type III sums of squares.

Causes of Unequal Sample Sizes

None of the methods for dealing with unequal sample sizes are valid if the experimental treatment is the source of the unequal sample sizes. Imagine an experiment seeking to determine whether publicly performing an embarrassing act would affect one's anxiety about public speaking. In this imaginary experiment, the experimental group is asked to reveal to a group of people the most embarrassing

thing they have ever done. The control group is asked to describe what they had at their last meal. Twenty subjects are recruited for the experiment and randomly divided into two equal groups of 10, one for the experimental treatment and one for the control. Following their descriptions, subjects are given an attitude survey concerning public speaking. This seems like a valid experimental design. However, of the 10 subjects in the experimental group, four withdrew from the experiment because they did not wish to publicly describe an embarrassing situation. None of the subjects in the control group withdrew. Even if the data analysis were to show a significant effect, it would not be valid to conclude that the treatment had an effect because a likely alternative explanation cannot be ruled out; namely, subjects who were willing to describe an embarrassing situation differed from those who were not. Thus, the differential dropout rate destroyed the *random assignment* of subjects to conditions, a critical feature of the experimental design. No amount of statistical adjustment can compensate for this flaw.

References

Maxwell, S. E., & Delaney, H. D. (2003) *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Second Edition, Lawrence Erlbaum Associates, Mahwah, New Jersey.

Tukey, J. W. (1991) The philosophy of multiple comparisons, *Statistical Science*, 6, 110-116.

Tests Supplementing ANOVA

by David M. Lane

Prerequisites

- Chapter 15: One-Factor ANOVA, Multi-Factor ANOVA
- Chapter 15: Pairwise Comparisons Among Means
- Chapter 15: Specific Comparisons Among Means

Learning Objectives

1. Compute Tukey HSD test
2. Describe an interaction in words
3. Describe why one might want to compute simple effect tests following a significant interaction

The *null hypothesis* tested in a one-factor ANOVA is that all the population means are equal. Stated more formally,

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$

where H_0 is the null hypothesis and k is the number of conditions. When the null hypothesis is rejected, all that can be said is that at least one population mean is different from at least one other population mean. The methods for doing more specific tests described in "All Pairwise Comparisons among Means" and in "Specific Comparisons" apply here. Keep in mind that these tests are valid whether or not they are preceded by an ANOVA.

Main Effects

As will be seen, significant *main effects* in multi-factor designs can be followed up in the same way as significant effects in one-way designs. Table 1 shows the data from an imaginary experiment with three *levels* of Factor A and two levels of Factor B.

Table 1. Made-Up Example Data.

	A1	A2	A3	Marginal Means
B1	5 4 6 5	9 8 7 8	5 9 9 8	7.08
	Mean = 5	Mean = 8	Mean = 8.25	
B2	4 3 6 8	8 6 8 5	8 9 7 6	6.50
	Mean = 5.25	Mean = 6.75	Mean = 7.50	
Marginal Means	5.125	7.375	7.875	6.79

Table 2 shows the ANOVA Summary Table for these data. The *significant* main effect of A indicates that, in the population, at least one of the *marginal means* for A is different from at least one of the others.

Table 2. ANOVA Summary Table for Made-Up Example Data.

Source	df	SSQ	MS	F	p
A	2	34.333	17.17	9.29	0.002
B	1	2.042	2.04	1.10	0.307
A x B	2	2.333	1.167	0.63	0.543
Error	18	33.250	1.847		
Total	23	71.958			

The Tukey HSD test can be used to test all pairwise comparisons among means in a one-factor ANOVA as well as comparisons among marginal means in a multi-factor ANOVA. The formula for the equal-sample-size case is shown below.

$$Q = \frac{M_i - M_j}{\sqrt{\frac{MSE}{n}}}$$

where M_i and M_j are marginal means, MSE is the mean square error from the ANOVA, and n is the number of scores each mean is based upon. For this example, MSE = 1.847 and n = 8 because there are eight scores at each level of A. The probability value can be computed using the Studentized Range Calculator. The degrees of freedom is equal to the degrees of freedom error. For this example, df = 18. The results of the Tukey HSD test are shown in Table 3. The mean for A_1 is significantly lower than the mean for A_2 and the mean for A_3. The means for A_2 and A_3 are not significantly different.

Table 3. Pairwise Comparisons Among Marginal Means for A.

Comparison	Mi - Mj	Q	p
A_1 - A_2	-2.25	-4.68	0.010
A_1 - A_3	-2.75	-5.73	0.002
A_2 - A_3	-0.50	-1.04	0.746

Specific comparisons among means are also carried out much the same way as shown in the relevant section on testing means. The formula for L is

$$L = \sum c_i M_i$$

where c_i is the coefficient for the i^{th} marginal mean and M_i is the i^{th} marginal mean. For example, to compare A_1 with the average of A_2 and A_3, the coefficients would be 1, -0.5, -0.5. Therefore,

```
L = (1)(5.125) + (-0.5)(7.375) + (-0.5)(7.875)

= -2.5.
```

To compute t, use:

$$t = \frac{L}{\sqrt{\frac{\sum c_i^2 MSE}{n}}}$$

$$= -4.25$$

where MSE is the mean square error from the ANOVA and n is the number of scores each marginal mean is based on (eight in this example). The degrees of freedom is the degrees of freedom error from the ANOVA and is equal to 18. Using the Online Calculator, we find that the two-tailed probability value is 0.0005. Therefore, the difference between A_1 and the average of A_2 and A_3 is significant.

Important issues concerning multiple comparisons and *orthogonal comparisons* are discussed in the Specific Comparisons section in the Testing Means chapter.

Interactions

The presence of a significant interaction makes the interpretation of the results more complicated. Since an interaction means that the *simple effects* are different, the main effect as the mean of the simple effects does not tell the whole story. This section discusses how to describe interactions, proper and improper uses of simple effects tests, and how to test components of interactions.

Describing Interactions

A crucial first step in understanding a significant interaction is constructing an *interaction plot*. Figure 1 shows an interaction plot from data presented in the section on Multi-Factor ANOVA.

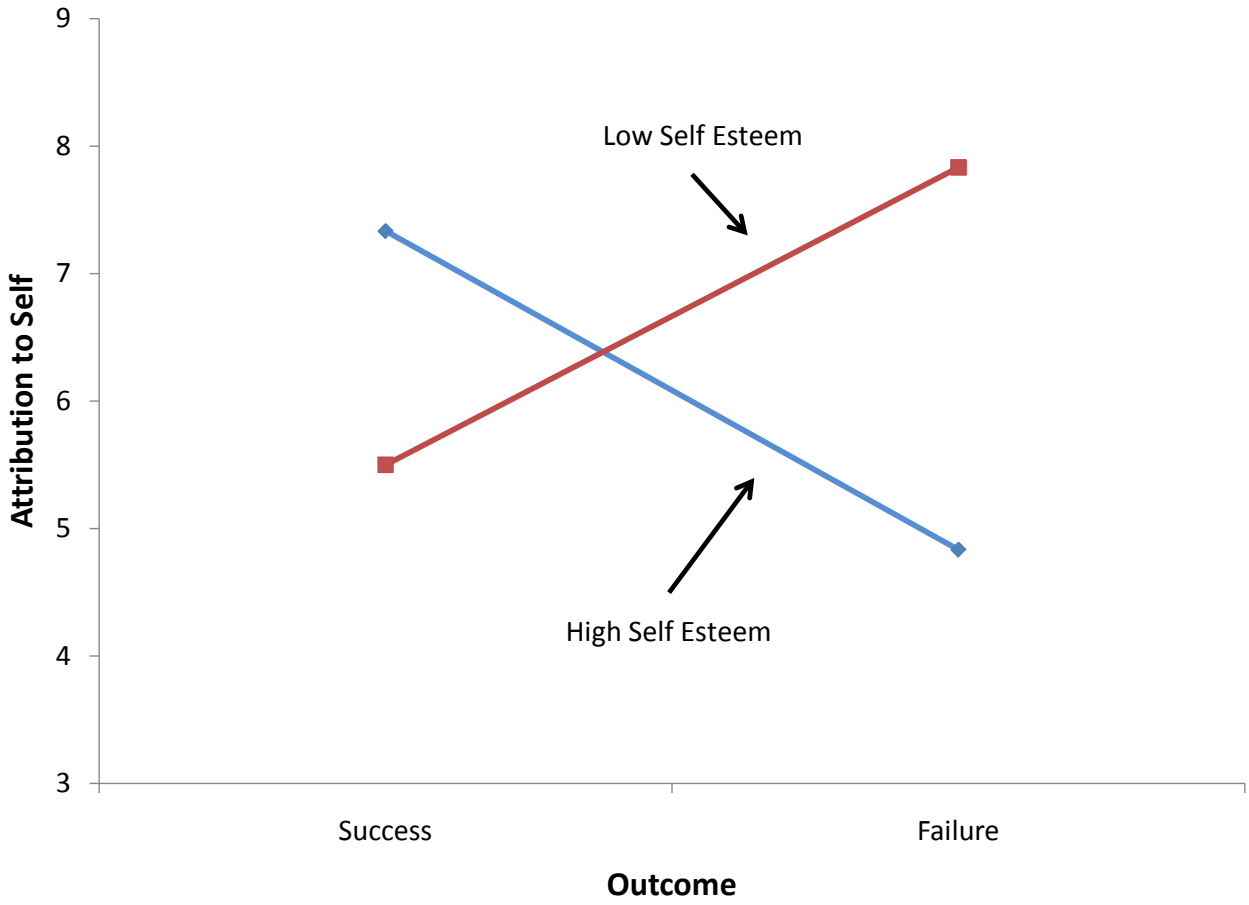


Figure 1. Interaction plot for made-up data.

The second step is to describe the interaction in a clear and understandable way. This is often done by describing how the *simple effects* differed. Since this should be done using as little jargon as possible, the expression "simple effect" need not appear in the description. An example is as follows:

> The effect of Outcome differed depending on the subject's self-esteem. The difference between the attribution to self following success and the attribution to self following failure was larger for high-self-esteem subjects (mean difference = 2.50) than for low-self-esteem subjects (mean difference = -2.33).

No further analyses are helpful in understanding the interaction since the interaction means only that the simple effects differ. The interaction's significance indicates that the simple effects differ from each other, but provides no information about whether they differ from zero.

Simple Effect Tests

It is not necessary to know whether the simple effects differ from zero in order to understand an interaction because the question of whether simple effects differ from zero has nothing to do with interaction except that if they are both zero there is no interaction. It is not uncommon to see research articles in which the authors report that they analyzed simple effects in order to explain the interaction. However, this is not a valid approach since an interaction does not depend on the analysis of the simple effects.

However, there is a reason to test simple effects following a significant interaction. Since an interaction indicates that simple effects differ, it means that the main effects are not general. In the made-up example, the main effect of Outcome is not very informative, and the effect of outcome should be considered separately for high- and low-self-esteem subjects.

As will be seen, the simple effects of Outcome are significant and in opposite directions: Success significantly increases attribution to self for high-self-esteem subjects and significantly lowers attribution to self for low-self-esteem subjects. This is a very easy result to interpret.

What would the interpretation have been if neither simple effect had been significant? On the surface, this seems impossible: How can the simple effects both be zero if they differ from each other significantly as tested by the interaction? The answer is that a non-significant simple effect does not mean that the simple effect is zero: the null hypothesis should not be accepted just because it is not rejected.

(See section on Interpreting Non-Significant Results)

If neither simple effect is significant, the conclusion should be that the simple effects differ, and that at least one of them is not zero. However, no conclusion should be drawn about which simple effect(s) is/are not zero.

Another error that can be made by mistakenly accepting the null hypothesis is to conclude that two simple effects are different because one is significant and the other is not. Consider the results of an imaginary experiment in which the researcher hypothesized that addicted people would show a larger increase in brain activity following some treatment than would non-addicted people. In other words, the researcher hypothesized that addiction status and treatment would interact. The results shown in Figure 2 are very much in line with the hypothesis. However, the test of the interaction resulted in a *probability value* of 0.08, a value not quite low enough to be significant at the conventional 0.05 level. The proper conclusion is

that the experiment supports the researcher's hypothesis, but not strongly enough to allow a confident conclusion.

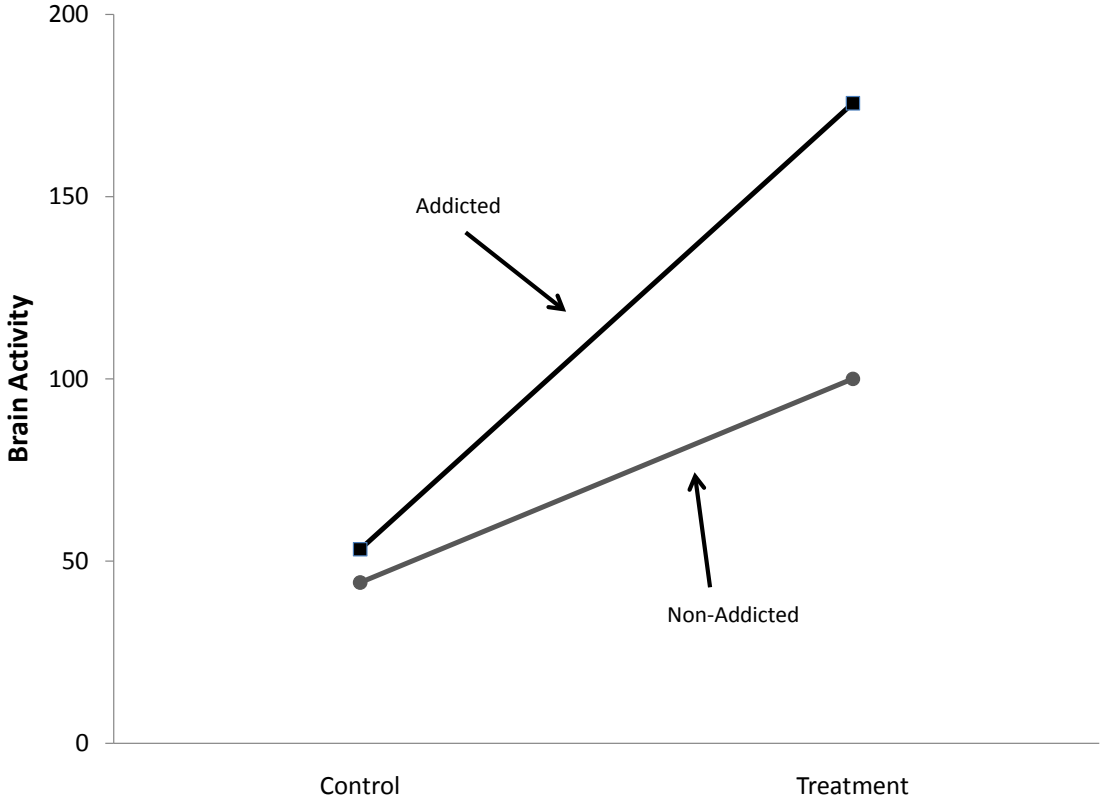


Figure 2. Made-up data with one significant simple effect.

Unfortunately, the researcher was not satisfied with such a weak conclusion and went on to test the simple effects. It turned out that the effect of Treatment was significant for the Addicted group ($p = 0.02$) but not significant for the Non-Addicted group ($p = 0.09$). The researcher then went on to conclude that since there is an effect of Treatment for the Addicted group but not for the Non-Addicted group, the hypothesis of a greater effect for the former than for the latter group is demonstrated. This is faulty logic, however, since it is based on accepting the null hypothesis that the simple effect of Treatment is zero for the Non-Addicted group just because it is not significant.

Components of Interaction (optional)

Figure 3 shows the results of an imaginary experiment on diet and weight loss. A control group and two diets were used for both overweight teens and overweight adults.

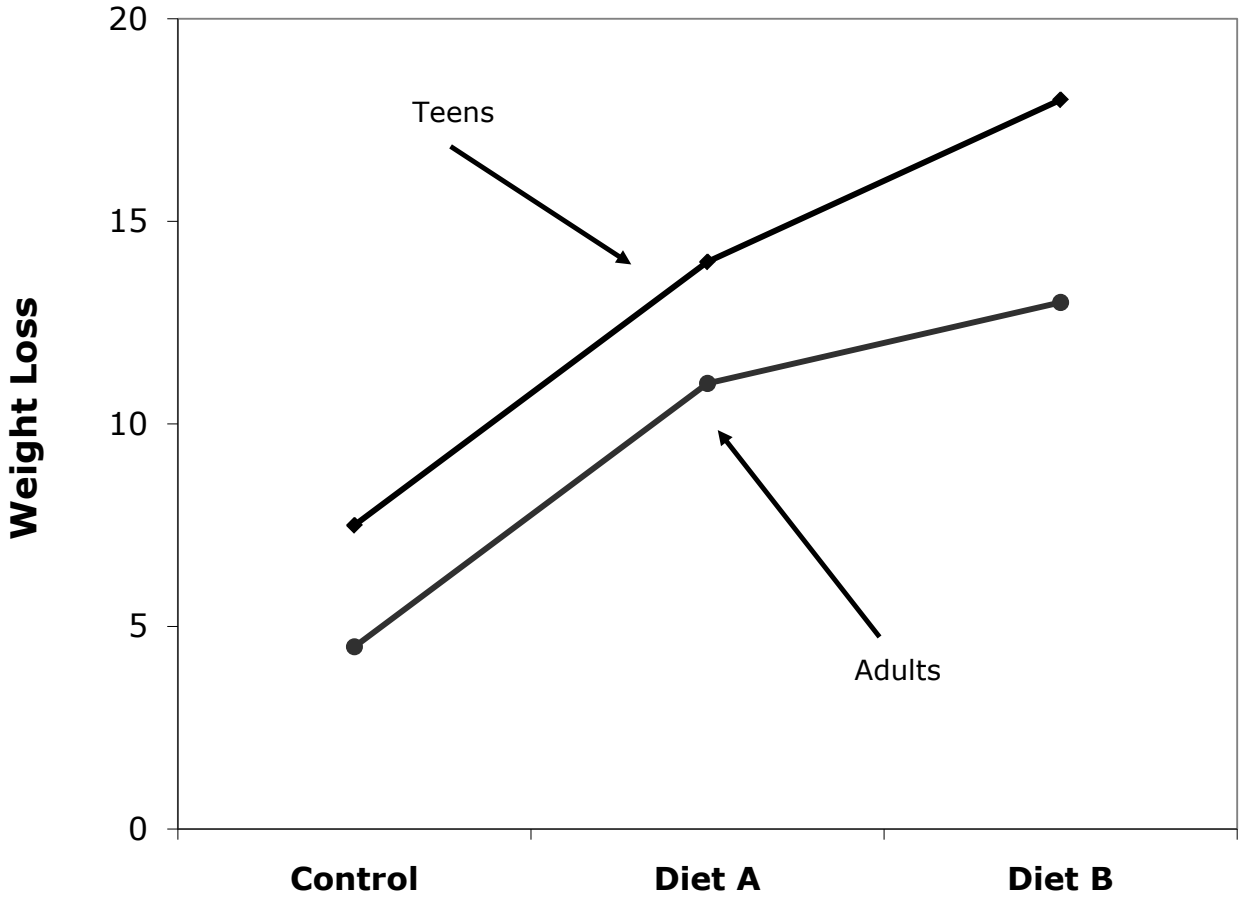


Figure 3. Made-up Data for Diet Study.

The difference between Diet A and the Control diet was essentially the same for teens and adults, whereas the difference between Diet B and Diet A was much larger for the teens than it was for the adults. Over one portion of the graph the lines are parallel, whereas over another portion they are not. It is possible to test these portions or components of interactions using the method of specific comparisons discussed previously. The test of the difference between Teens and Adults on the difference between Diets A and B could be tested with the coefficients shown in Table 4. Naturally, the same consideration regarding multiple comparisons and *orthogonal comparisons* that apply to other comparisons among means also apply to comparisons involving components of interactions.

Table 4. Coefficients for a Component of the Interaction.

Age Group	Diet	Coefficient
Teen	Control	0
Teen	A	1
Teen	B	-1
Adult	Control	0
Adult	A	-1
Adult	B	1

Within-Subjects ANOVA

by David M. Lane

Prerequisites

- Chapter 12: Difference Between Two Means (Correlated Pairs)
- Chapter 15: Additional Measures of Central Tendency
- Chapter 15: Introduction to ANOVA
- Chapter 15: ANOVA Designs, Multi-Factor ANOVA

Learning Objectives

1. Define a within-subjects factor
2. Explain why a within-subjects design can be expected to have more power than a between-subjects design
3. Be able to create the Source and df columns of an ANOVA summary table for a one-way within-subjects design
4. Explain error in terms of interaction
5. Discuss the problem of carryover effects
6. Be able to create the Source and df columns of an ANOVA summary table for a design with one between-subjects and one within-subjects variable
7. Define sphericity
8. Describe the consequences of violating the assumption of sphericity
9. Discuss courses of action that can be taken if sphericity is violated

Within-subjects factors involve comparisons of the same subjects under different conditions. For example, in the "ADHD Treatment" study, each child's performance was measured four times, once after being on each of four drug doses for a week. Therefore, each subject's performance was measured at each of the four *levels* of the *factor* "Dose." Note the difference from *between-subjects factors* for which each subject's performance is measured only once and the comparisons are among different groups of subjects. A within-subjects factor is sometimes referred to as a *repeated-measures factor* since repeated measurements are taken on each subject. An experimental design in which the independent variable is a *within-subjects factor* is called a within-subjects design.

An advantage of within-subjects designs is that individual differences in subjects' overall levels of performance are controlled. This is important because subjects invariably will differ from one another. In an experiment on problem

solving, some subjects will be better than others regardless of the condition they are in. Similarly, in a study of blood pressure some subjects will have higher blood pressure than others regardless of the condition. Within-subjects designs control these individual differences by comparing the scores of a subject in one condition to the scores of the same subject in other conditions. In this sense each subject serves as his or her own control. This typically gives within-subjects designs considerably more *power* than between-subjects designs.

One-Factor Designs

Let's consider how to analyze the data from the "ADHD Treatment" case study. These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. For now, we will be concerned only with testing the difference between the mean in the placebo condition (the lowest dosage, D0) and the mean in the highest dosage condition (D60). The details of the computations are relatively unimportant since they are almost universally done by computers. Therefore we jump right to the ANOVA Summary table shown in Table 1.

Table 1. ANOVA Summary Table

Source	df	SSQ	MS	F	p
Subjects	23	5781.98	251.39		
Dosage	1	295.02	295.02	10.38	0.004
Error	23	653.48	28.41		
Total	47	6730.48			

The first source of variation, "Subjects," refers to the differences among subjects. If all the subjects had exactly the same mean (across the two dosages), then the sum of squares for subjects would be zero; the more subjects differ from each other, the larger the sum of squares subjects.

Dosage refers to the differences between the two dosage levels. If the means for the two dosage levels were equal, the sum of squares would be zero. The larger the difference between means, the larger the sum of squares.

The error reflects the degree to which the effect of dosage is different for different subjects. If subjects all responded very similarly to the drug, then the error would be very low. For example, if all subjects performed moderately better

with the high dose than they did with the placebo, then the error would be low. On the other hand, if some subjects did better with the placebo while others did better with the high dose, then the error would be high. It should make intuitive sense that the less consistent the effect of dosage, the larger the dosage effect would have to be in order to be significant. The degree to which the effect of dosage differs depending on the subject is the Subjects x Dosage interaction. Recall that an interaction occurs when the effect of one variable differs depending on the level of another variable. In this case, the size of the error term is the extent to which the effect of the variable "Dosage" differs depending on the level of the variable "Subjects." Note that each subject is a different level of the variable "Subjects."

Other portions of the summary table have the same meaning as in between-subjects ANOVA. The F for dosage is the mean square for dosage divided by the mean square error. For these data, the F is significant with $p = 0.004$. Notice that this F test is equivalent to the t test for correlated pairs, with $F = t^2$.

Table 2 shows the ANOVA Summary Table when all four doses are included in the analysis. Since there are now four dosage levels rather than two, the df for dosage is three rather than one. Since the error is the Subjects x Dosage interaction, the df for error is the df for "Subjects" (23) times the df for Dosage (3) and is equal to 69.

Table 2. ANOVA Summary Table

Source	df	SSQ	MS	F	p
Subjects	23	9065.49	394.15		
Dosage	3	557.61	185.87	5.18	0.003
Error	69	2476.64	35.89		
Total	95	12099.74			

Carryover Effects

Often performing in one condition affects performance in a subsequent condition in such a way as to make a within-subjects design impractical. For example, consider an experiment with two conditions. In both conditions subjects are presented with pairs of words. In Condition A, subjects are asked to judge whether the words have similar meaning whereas in Condition B, subjects are asked to judge whether they sound similar. In both conditions, subjects are given a surprise

memory test at the end of the presentation. If Condition were a within-subjects variable, then there would be no surprise after the second presentation and it is likely that the subjects would have been trying to memorize the words.

Not all carryover effects cause such serious problems. For example, if subjects get fatigued by performing a task, then they would be expected to do worse on the second condition they were in. However, as long as the order of presentation is counterbalanced so that half of the subjects are in Condition A first and Condition B second, the fatigue effect itself would not invalidate the results, although it would add noise and reduce power. The carryover effect is symmetric in that having Condition A first affects performance in Condition B to the same degree that having Condition B first affects performance in Condition A.

Asymmetric carryover effects cause more serious problems. For example, suppose performance in Condition B were much better if preceded by Condition A, whereas performance in Condition A was approximately the same regardless of whether it was preceded by Condition B. With this kind of carryover effect, it is probably better to use a between-subjects design.

One Between- and One Within-Subjects Factor

In the "Stroop Interference" case study, subjects performed three tasks: naming colors, reading color words, and naming the ink color of color words. Some of the subjects were males and some were females. Therefore, this design had two factors: gender and task. The ANOVA Summary Table for this design is shown in Table 3.

Table 3. ANOVA Summary Table for Stroop Experiment

Source	df	SSQ	MS	F	p
Gender	1	83.32	83.32	1.99	0.165
Error	45	1880.56	41.79		
Task	2	9525.97	4762.99	228.06	<0.001
Gender x Task	2	55.85	27.92	1.34	0.268
Error	90	1879.67	20.89		

The computations for the sums of squares will not be covered since computations are normally done by software. However, there are some important things to learn

from the summary table. First, notice that there are two error terms: one for the between-subjects variable Gender and one for both the within-subjects variable Task and the interaction of the between-subjects variable and the within-subjects variable. Typically, the mean square error for the between-subjects variable will be higher than the other mean square error. In this example, the mean square error for Gender is about twice as large as the other mean square error.

The degrees of freedom for the between-subjects variable is equal to the number of levels of the between-subjects variable minus one. In this example, it is one since there are two levels of gender. Similarly, the degrees of freedom for the within-subjects variable is equal to the number of levels of the variable minus one. In this example, it is two since there are three tasks. The degrees of freedom for the interaction is the product of the degrees of freedom for the two variables. For the Gender x Task interaction, the degrees of freedom is the product of degrees of freedom Gender (which is 1) and the degrees of freedom Task (which is 2) and is equal to 2.

Assumption of Sphericity

Within-subjects ANOVA makes a restrictive assumption about the variances and the correlations among the dependent variables. Although the details of the assumption are beyond the scope of this book, it is approximately correct to say that it is assumed that all the correlations are equal and all the variances are equal. Table 4 shows the correlations among the three dependent variables in the Stroop Interference case study.

Table 4. Correlations Among Dependent Variables

	word reading	color naming	interference
word reading	1	0.7013	0.1583
color naming	0.7013	1	0.2382
interference	0.1583	0.2382	1

Note that the correlation between the word reading and the color naming variables of 0.7013 is much higher than the correlation between either of these variables with the interference variable. Moreover, as shown in Table 5, the variances among the variables differ greatly.

Table 5. Variances.

Variable	Variance
word reading	15.77
color naming	13.92
interference	55.07

Naturally the assumption of sphericity, like all assumptions, refers to populations not samples. However, it is clear from these sample data that the assumption is not met in the population.

Consequences of Violating the Assumption of Sphericity

Although ANOVA is robust to most violations of its assumptions, the assumption of sphericity is an exception: Violating the assumption of sphericity leads to a substantial increase in the Type I error rate. Moreover, this assumption is rarely met in practice. Although violations of this assumption had at one time received little attention, the current consensus of data analysts is that it is no longer considered acceptable to ignore them.

Approaches to Dealing with Violations of Sphericity

If an effect is highly significant, there is a conservative test that can be used to protect against an inflated Type I error rate. This test consists of adjusting the degrees of freedom for all within-subjects variables as follows: The degrees of freedom numerator and denominator are divided by the number of scores per subject minus one. Consider the effect of Task shown in Table 3. There are three scores per subject and therefore the degrees of freedom should be divided by two. The adjusted degrees of freedom are:

```
(2)(1/2) = 1 for the numerator and

(90)(1/2 = 45 for the denominator
```

The probability value is obtained using the F probability calculator with the new degrees of freedom parameters. The probability of an F of 228.06 of larger with 1 and 45 degrees of freedom is less than 0.001. Therefore, there is no need to worry about the assumption violation in this case.

Possible violation of sphericity does make a difference in the interpretation of the analysis shown in Table 2. The probability value of an F or 5.18 with 1 and 23 degrees of freedom is 0.032, a value that would lead to a more cautious conclusion than the p value of 0.003 shown in Table 2.

The correction described above is very conservative and should only be used when, as in Table 3, the probability value is very low. A better correction, but one that is very complicated to calculate, is to multiply the degrees of freedom by a quantity called ε (the Greek letter epsilon). There are two methods of calculating ε. The correction called the Huynh-Feldt (or H-F) is slightly preferred to the one called the Greenhouse Geisser (or G-G), although both work well. The G-G correction is generally considered a little too conservative.

A final method for dealing with violations of sphericity is to use a multivariate approach to within-subjects variables. This method has much to recommend it, but it is beyond the scope of this text.

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 15: Multi-factor ANOVA

A research design to compare three drugs for the treatment of Alzheimer's disease is described here. For the first two years of the study, researchers will follow the subjects with scans and memory tests.

What do you think?

Assume the data were analyzed as a two-factor design with pre-post testing as one factor and the three drugs as the second factor. What term in an ANOVA would reflect whether the pre-post change was different for the three drugs??

> It would be the interaction of the two factors since the question is whether the effect of one factor (pre-post) differs as a function of the level of a second factor (drug).

Exercises

Prerequisites

- All material presented in the ANOVA Chapter

1. What is the null hypothesis tested by analysis of variance?

2. What are the assumptions of between-subjects analysis of variance?

3. What is a between-subjects variable?

4. Why not just compute t-tests among all pairs of means instead computing an analysis of variance?

5. What is the difference between "N" and "n"?

6. How is it that estimates of variance can be used to test a hypothesis about means?

7. Explain why the variance of the sample means has to be multiplied by "n" in the computation of $MS_{between}$.

8. What kind of skew does the F distribution have?

9. When do $MS_{between}$ and MS_{error} estimate the same quantity?

10. If an experiment is conducted with 5 conditions and 6 subjects in each condition, what are df_n and df_e?

11. How is the shape of the F distribution affected by the degrees of freedom?

12. What are the two components of the total sum of squares in a one-factor between-subjects design?

13. How is the mean square computed from the sum of squares?

14. An experimenter is interested in the effects of two independent variables on self-esteem. What is better about conducting a factorial experiment than conducting two separate experiments, one for each independent variable?

15. An experiment is conducted on the effect of age (5 yr, 10 yr and 15 yr) and treatment condition (experimental versus control) on reading speed. Which statistical term (main effect, simple effect, interaction, specific comparison) applies to each of the descriptions of effects.

    a. The effect of the treatment was larger for 15-year olds than it was for 5- or 10-year olds.

    b. Overall, subjects in the treatment condition performed faster than subjects in the control condition.

    c. The age effect was significant under the treatment condition.

    d. The difference between the 15- year olds and the average of the 5- and 10-year olds was significant.

    e. As they grow older, children read faster.

16. An A(3) x B(4) factorial design with 6 subjects in each group is analyzed. Give the source and degrees of freedom columns of the analysis of variance summary table.

17. The following data are from a hypothetical study on the effects of age and time on scores on a test of reading comprehension. Compute the analysis of variance summary table.

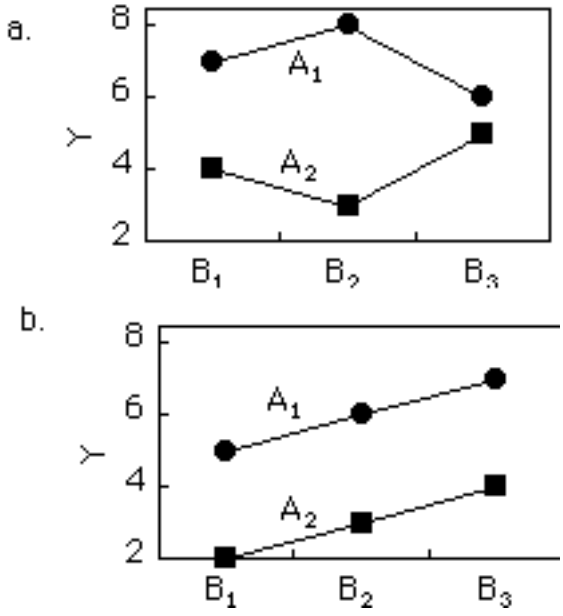
	12-year olds	**16-year olds**
30 minutes	66 68 59 72 46	74 71 67 82 76
60 minutes	69 61 69 73 61	95 92 95 98 94

18. Define “Three-way interaction”

19. Define interaction in terms of simple effects.

20. Plot an interaction for an A(2) x B(2) design in which the effect of B is greater at A1 than it is at A2. The dependent variable is "Number correct." Make sure to label both axes.

21. Following are two graphs of population means for 2 x 3 designs. For each graph, indicate which effect(s) (A, B, or A x B) are nonzero.



22. The following data are from an A(2) x B(4) factorial design.

	B1	B2	B3	B4
A1	1	2	3	4
	3	2	4	5
	4	4	2	6
	5	5	6	8
A2	1	2	4	8
	1	3	6	9
	2	2	7	9
	2	4	8	8

a. Compute an analysis of variance.

b. Test differences among the four levels of B using the Bonferroni correction.

c. Test the linear component of trend for the effect of B.

d. Plot the interaction.

e. Describe the interaction in words.

23. Why are within-subjects designs usually more powerful than between-subjects design?

24. What source of variation is found in an ANOVA summary table for a within-subjects design that is not in in an ANOVA summary table for a between-subjects design. What happens to this source of variation in a between-subjects design?

25. The following data contain three scores from each of five subjects. The three scores per subject are their scores on three trials of a memory task.

4	6	7
3	7	7
2	8	5
1	4	7
4	6	9

a. Compute an ANOVA

b. Test all pairwise differences between means using the Bonferroni test at the .01 level.

c. Test the linear and quadratic components of trend for these data.

26. Give the source and df columns of the ANOVA summary table for the following experiments:

a. Twenty two subjects are each tested on a simple reaction time task and on a choice reaction time task.

b. Twelve male and 12 female subjects are each tested under three levels of drug dosage: 0 mg, 10 mg, and 20 mg.

c. Twenty subjects are tested on a motor learning task for three trials a day for two days.

d. An experiment is conducted in which depressed people are either assigned to a drug therapy group, a behavioral therapy group, or a control group. Ten

subjects are assigned to each group. The level of measured once a month for four months.

Questions from Case Studies

Stroop Interference (S) case study

27. (S) The dataset Stroop Interference has the scores (times) for males and females on each of three tasks.
    a. Do a Gender (2) x Task (3) analysis of variance.
    b. Plot the interaction.

ADHD Treatment (AT) case study

28. (AT) The dataset ADHD Treatment has four scores per subject. a. Is the design between-subjects or within-subjects? b. Create an ANOVA summary table.

29. (AT) Using the Anger Expression Index from the Angry Moods study as the dependent variable, perform a 2x2 ANOVA with gender and sports participation as the two factors. Do athletes and non-athletes differ significantly in how much anger they express? Do the genders differ significantly in Anger Expression Index? Is the effect of sports participation significantly different for the two genders?

Weapons and Aggression (WA) case study

30. (WA) Using the Weapons and Aggression data, Compute a 2x2 ANOVA with the follow- ing two factors: prime type (was the first word a weapon or not?) and word type (was the second word aggressive or non-aggressive?). Consider carefully whether the variables are between-subject or within-subjects variables.

"Smiles and Leniency" (SL) case study

31. (SL) Compute the ANOVA summary table for the smiles and leniency data.

16. Transformations

A. Log

B. Tukey's Ladder of Powers

C. Box-Cox Transformations

D. Exercises

The focus of statistics courses is the exposition of appropriate methodology to analyze data to answer the question at hand. Sometimes the data are given to you, while other times the data are collected as part of a carefully-designed experiment. Often the time devoted to statistical analysis is less than 10% of the time devoted to data collection and preparation. If aspects of the data preparation fail, then the success of the analysis is in jeopardy. Sometimes errors are introduced into the recording of data. Sometimes biases are inadvertently introduced in the selection of subjects or the mis-calibration of monitoring equipment.

In this chapter, we focus on the fact that many statistical procedures work best if individual variables have certain properties. The measurement scale of a variable should be part of the data preparation effort. For example, the correlation coefficient does not require the variables have a normal shape, but often relationships can be made clearer by re-expressing the variables. An economist may choose to analyze the logarithm of prices if the relative price is of interest. A chemist may choose to perform a statistical analysis using the inverse temperature as a variable rather than the temperature itself. But note that the inverse of a temperature will differ depending on whether it is measured in °F, °C, or °K.

The introductory chapter covered linear transformations. These transformations normally do not change statistics such as Pearson's r, although they do affect the mean and standard deviation. The first section here is on log transformations which are useful to reduce skew. The second section is on Tukey's ladder of powers. You will see that log transformations are a special case of the ladder of powers. Finally, we cover the relatively advanced topic of the Box-Cox transformation.

Log Transformations

by David M. Lane

Prerequisites

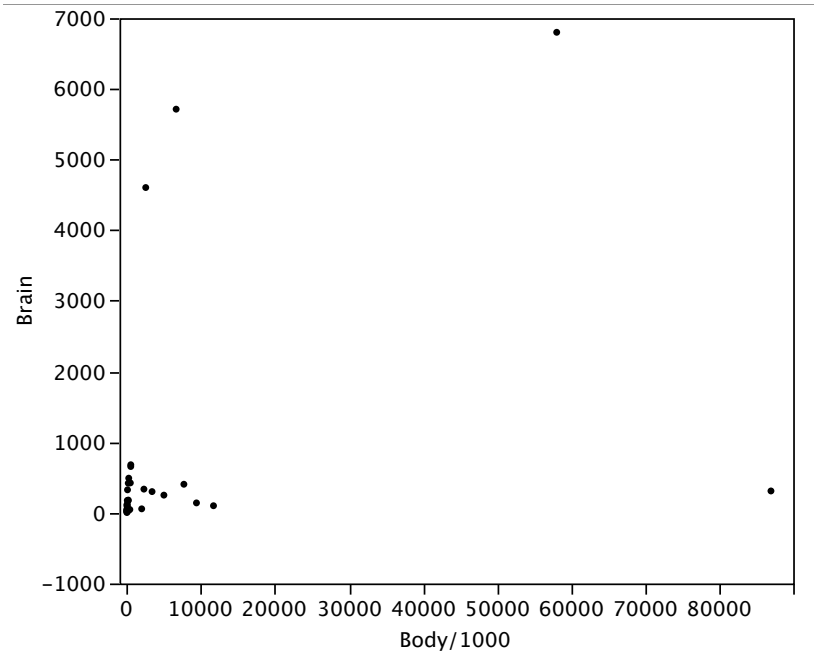
- Chapter 1: Logarithms
- Chapter 1: Shapes of Distributions
- Chapter 3: Additional Measures of Central Tendency
- Chapter 4: Introduction to Bivariate Data

Learning Objectives

1. State how a log transformation can help make a relationship clear
2. Describe the relationship between logs and the geometric mean

The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.

Figure 1 shows an example of how a log transformation can make patterns more visible. Both graphs plot the brain weight of animals as a function of their body weight. The raw weights are shown in the upper panel; the log-transformed weights are plotted in the lower panel.



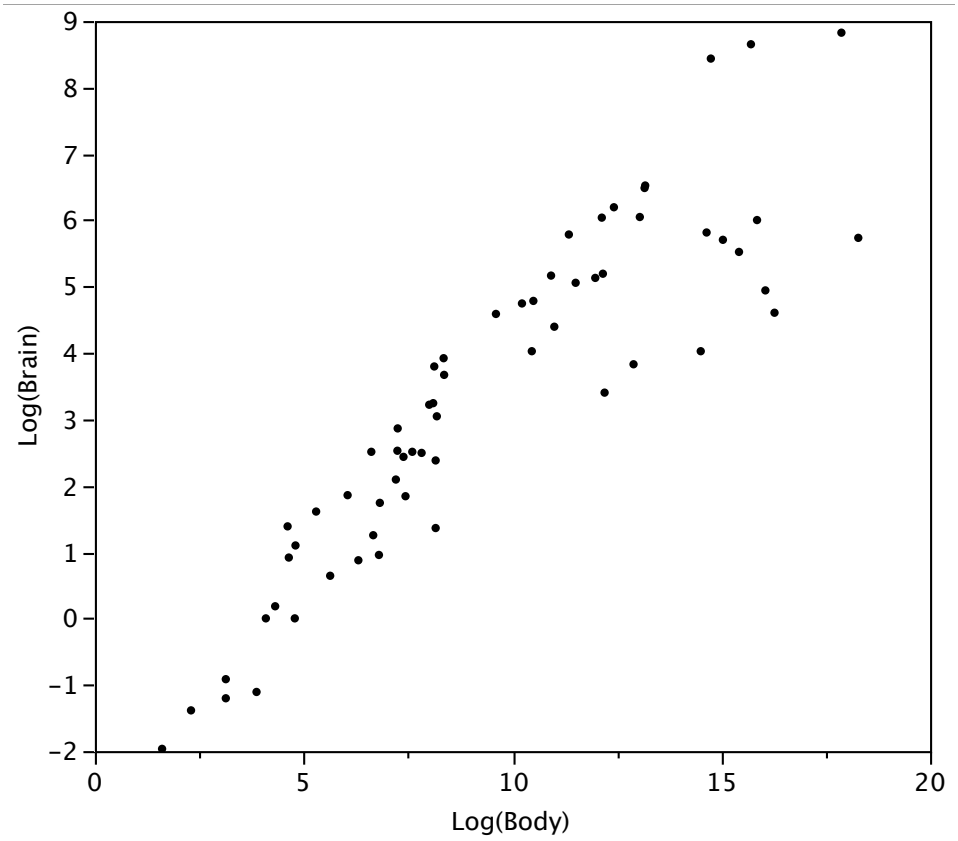


Figure 1. Scatter plots of brain weight as a function of body weight in terms of both raw data (upper panel) and log-transformed data (lower panel).

It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel.

The comparison of the means of log-transformed data is actually a comparison of geometric means. This occurs because, as shown below, the anti-log of the arithmetic mean of log-transformed values is the geometric mean.

Table 1 shows the logs (base 10) of the numbers 1, 10, and 100. The arithmetic mean of the three logs is

$$(0 + 1 + 2)/3 = 1$$

The anti-log of this arithmetic mean of 1 is:

$$10^1 = 10$$

which is the geometric mean:

$$(1 \times 10 \times 100)^{.3333} = 10.$$

Table 1. Logarithms.

X	$Log_{10}(X)$
1	0
10	1
100	2

Therefore, if the arithmetic means of two sets of log-transformed data are equal then the geometric means are equal.

Tukey Ladder of Powers

by David W. Scott

Prerequisites

- Chapter 1: Logarithms
- Chapter 4: Bivariate Data
- Chapter 4: Values of Pearson Correlation
- Chapter 12: Independent Groups t Test
- Chapter 13: Introduction to Power
- Chapter 16: Tukey Ladder of Powers

Learning Objectives

1. Give the Tukey ladder of transformations
2. Find a transformation that reveals a linear relationship
3. Find a transformation to approximate a normal distribution

Introduction

We assume we have a collection of bivariate data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

and that we are interested in the relationship between variables x and y. Plotting the data on a scatter diagram is the first step. As an example, consider the population of the United States for the 200 years before the Civil War. Of course, the decennial census began in 1790. These data are plotted two ways in Figure 1. Malthus predicted that geometric growth of populations coupled with arithmetic growth of grain production would have catastrophic results. Indeed the US population followed an exponential curve during this period.

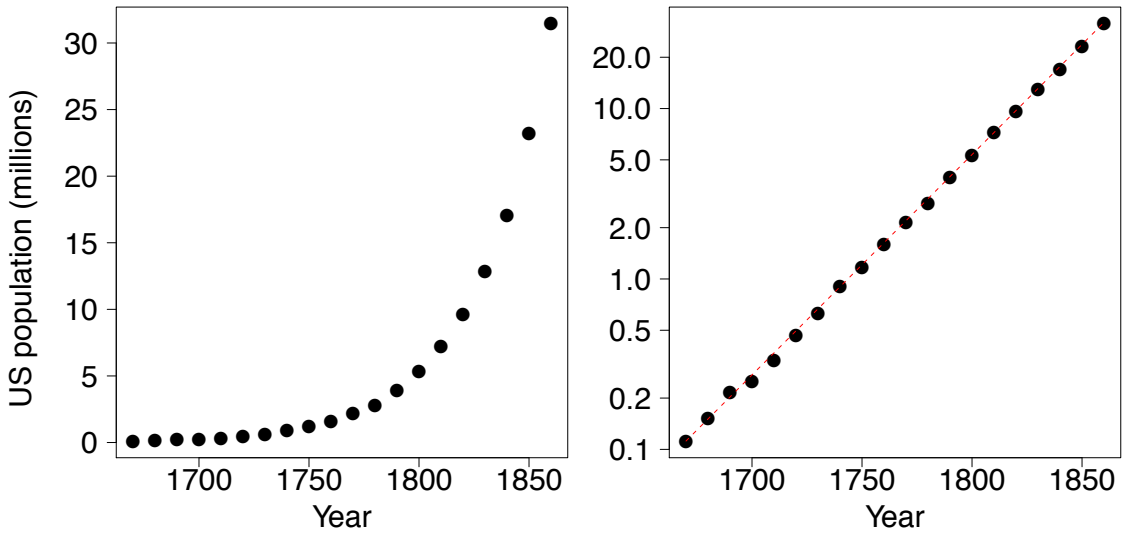


Figure 1. The US population from 1670 - 1860. The Y-axis on the right panel is on a log scale.

Tukey's Transformation Ladder

Tukey (1977) describes an orderly way of re-expressing variables using a power transformation. You may be familiar with polynomial regression (a form of multiple regression) in which the simple linear model $y = b_0 + b_1X$ is extended with terms such as $b_2x^2 + b_3x^3 + b_4x^4$. Alternatively, Tukey suggests exploring simple relationships such as

$$y = b_0 + b_1X^{\lambda} \text{ or } y^{\lambda} = b_0 + b_1X \quad \text{(Equation 1)}$$

where λ is a parameter chosen to make the relationship as close to a straight line as possible. Linear relationships are special, and if a transformation of the type x^{λ} or y^{λ} works as in Equation (1), then we should consider changing our measurement scale for the rest of the statistical analysis.

There is no constraint on values of λ that we may consider. Obviously choosing $\lambda = 1$ leaves the data unchanged. Negative values of λ are also reasonable. For example, the relationship

$$y = b_0 + b_1/x$$

would be represented by $\lambda = -1$. The value $\lambda = 0$ has no special value, since $X^0 = 1$, which is just a constant. Tukey (1977) suggests that it is convenient to simply define the transformation when $\lambda = 0$ to be the logarithm function rather than the

constant 1. We shall revisit this convention shortly. The following table gives examples of the Tukey ladder of transformations.

Table 1. Tukey's Ladder of Transformations

λ	-2	-1	-1/2	0	1/2	1	2
Xfm	$\frac{1}{x^2}$	$\frac{1}{x}$	$\frac{1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

If x takes on negative values, then special care must be taken so that the transformations make sense, if possible. We generally limit ourselves to variables where $x > 0$ to avoid these considerations. For some dependent variables such as the number of errors, it is convenient to add 1 to x before applying the transformation.

Also, if the transformation parameter λ is negative, then the transformed variable x^{λ} is reversed. For example, if x is increasing, then 1/x is decreasing. We choose to redefine the Tukey transformation to be $-(x^{\lambda})$ if $\lambda < 0$ in order to preserve the order of the variable after transformation. Formally, the Tukey transformation is defined as

$$\tilde{x}_{\lambda} = \begin{cases} x^{\lambda} & \text{if } \lambda > 0 \\ \log x & \text{if } \lambda = 0 \\ -(x^{\lambda}) & \text{if } \lambda < 0 \end{cases} \qquad (2)$$

In Table 2 we reproduce Table 1 but using the modified definition when $\lambda < 0$.

Table 2. Modified Tukey's Ladder of Transformations

λ	-2	-1	-1/2	0	1/2	1	2
Xfm	$\frac{-1}{x^2}$	$\frac{-1}{x}$	$\frac{-1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

The Best Transformation for Linearity

The goal is to find a value of λ that makes the scatter diagram as linear as possible. For the US population, the logarithmic transformation applied to y makes the relationship almost perfectly linear. The red dashed line in the right frame of Figure 1 has a slope of about 1.35; that is, the US population grew at a rate of about 35% per decade.

The logarithmic transformation corresponds to the choice λ = 0 by Tukey's convention. In Figure 2, we display the scatter diagram of the US population data for λ = 0 as well as for other choices of λ.

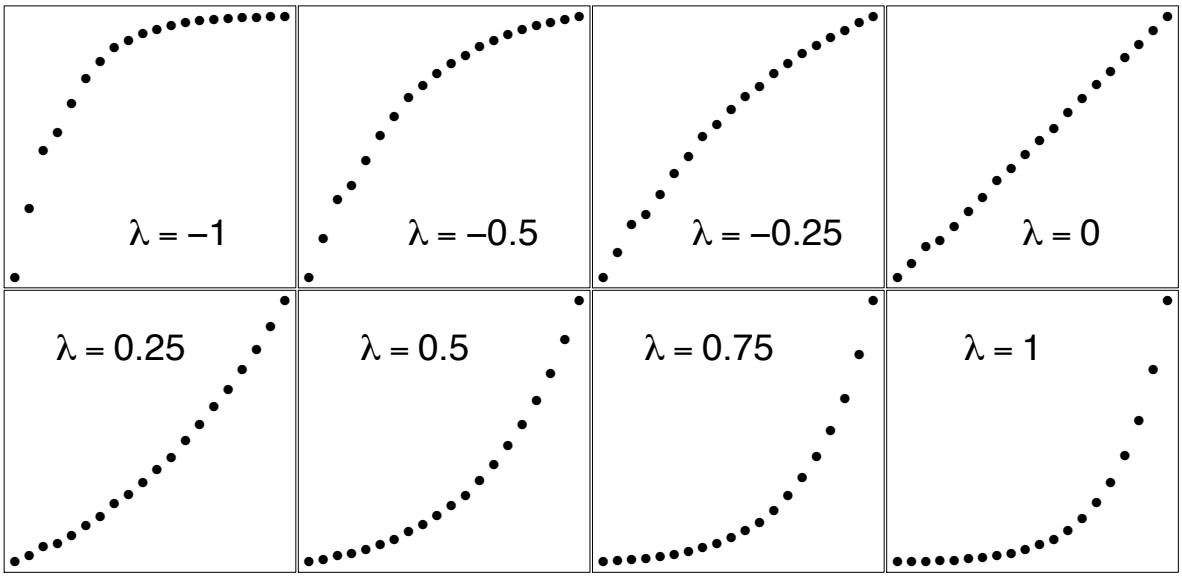


Figure 2. The US population from 1670 to 1860 for various values of λ.

The raw data are plotted in the bottom right frame of Figure 2 when λ = 1. The logarithmic fit is in the upper right frame when λ = 0. Notice how the scatter diagram smoothly morphs from convex to concave as λ increases. Thus intuitively there is a unique best choice of λ corresponding to the “most linear” graph.

One way to make this choice objective is to use an objective function for this purpose. One approach might be to fit a straight line to the transformed points and try to minimize the residuals. However, an easier approach is based on the fact that the correlation coefficient, r, is a measure of the linearity of a scatter diagram. In particular, if the points fall on a straight line then their correlation will be r = 1. (We need not worry about the case when r = −1 since we have defined the Tukey transformed variable x_λ to be positively correlated with x itself.)

In Figure 3, we plot the correlation coefficient of the scatter diagram $(x, \tilde{y}_\lambda)$ as a function of λ. It is clear that the logarithmic transformation (λ = 0) is nearly optimal by this criterion.

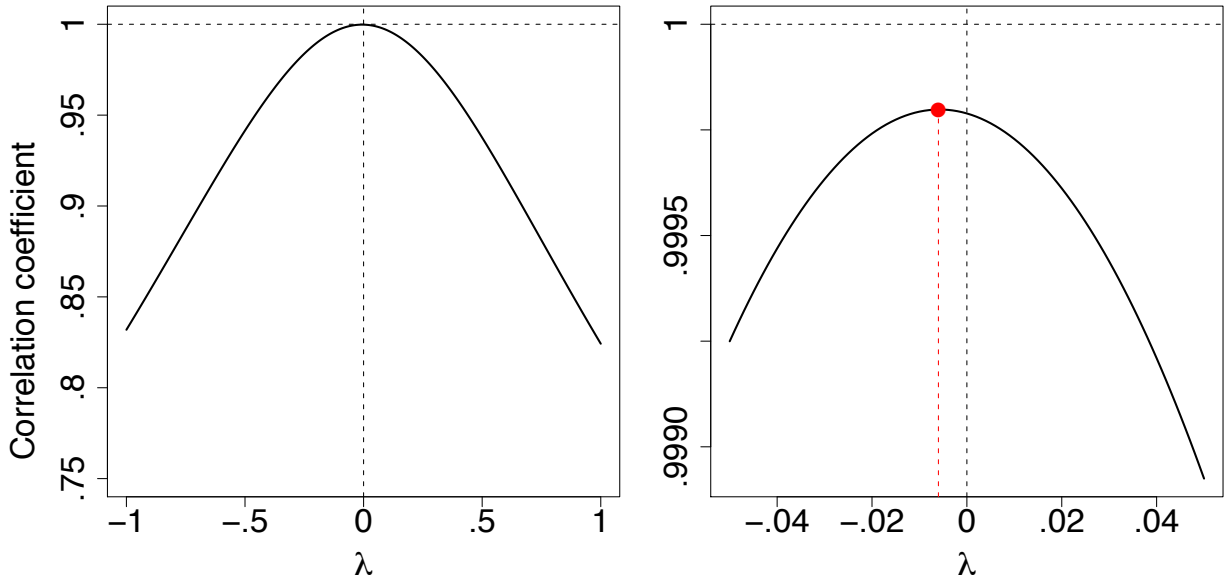


Figure 3. Graph of US population correlation coefficient as function of λ.

Is the US population still on the same exponential growth pattern? In Figure 4 we display the US population from 1630 to 2000 using the transformation and fit used in the right frame of Figure 1. Fortunately, the exponential growth (or at least its rate) was not sustained into the Twentieth Century. If it had, the US population in the year 2000 would have been over 2 billion (2.07 to be exact), larger than the population of China.

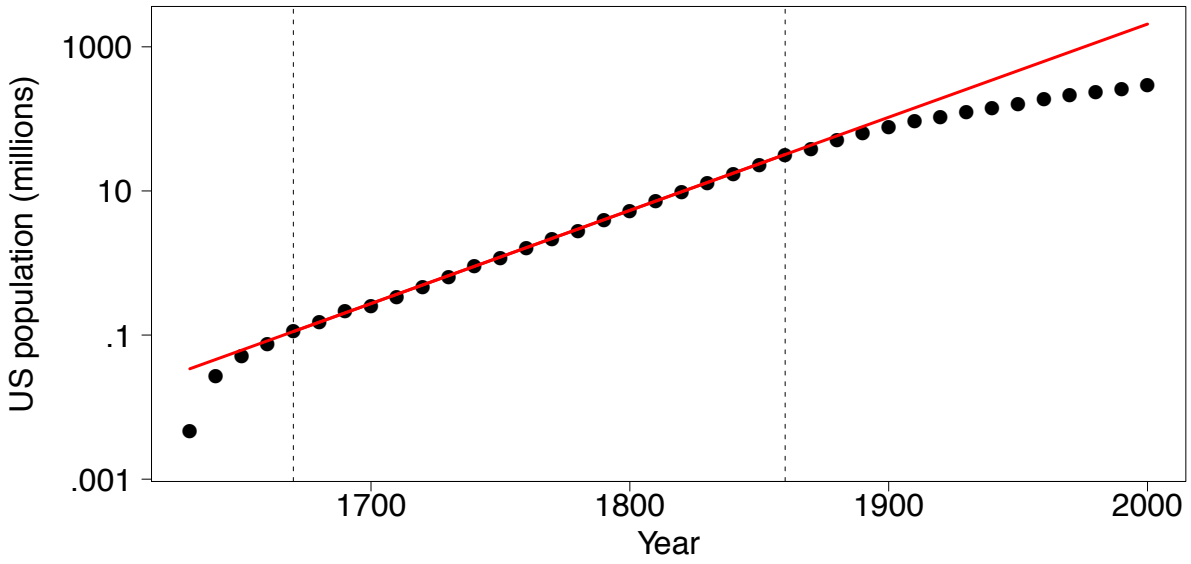


Figure 4. Graph of US population 1630-2000 with $\lambda = 0$.

We can examine the decennial census population figures of individual states as well. In Figure 5 we display the population data for the state of New York from 1790 to 2000, together with an estimate of the population in 2008. Clearly something unusual happened starting in 1970. (This began the period of mass migration to the West and South as the rust belt industries began to shut down.) Thus, we compute the best λ value using the data from 1790-1960 in the middle frame of Figure 5. The right frame displays the transformed data, together with the linear fit for the 1790-1960 period. The value of $\lambda = 0.41$ is not obvious and one might reasonably choose to use $\lambda = 0.50$ for practical reasons.

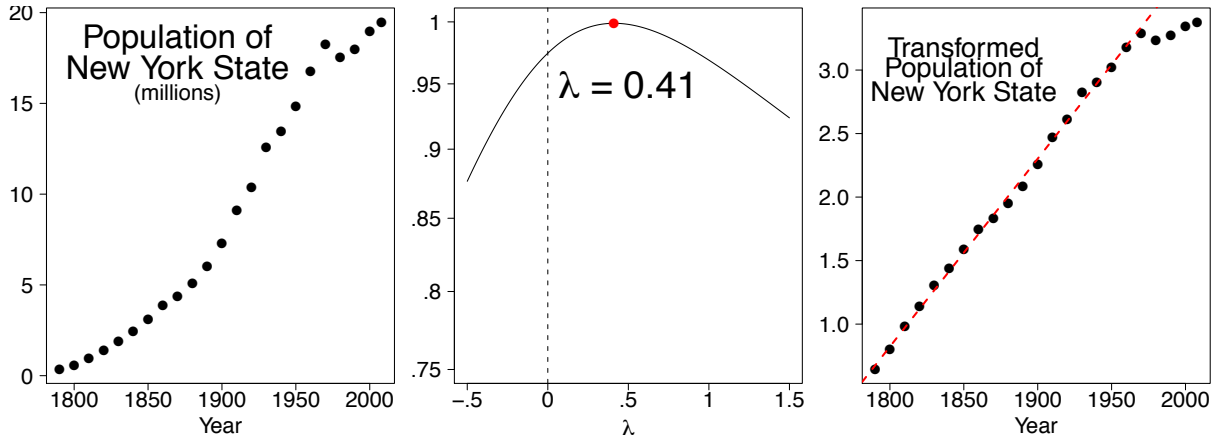


Figure 5. Graphs related to the New York state population 1790-2008.

If we look at one of the younger states in the West, the picture is different. Arizona has attracted many retirees and immigrants. Figure 6 summarizes our findings. Indeed, the growth of population in Arizona is logarithmic, and appears to still be logarithmic through 2005.

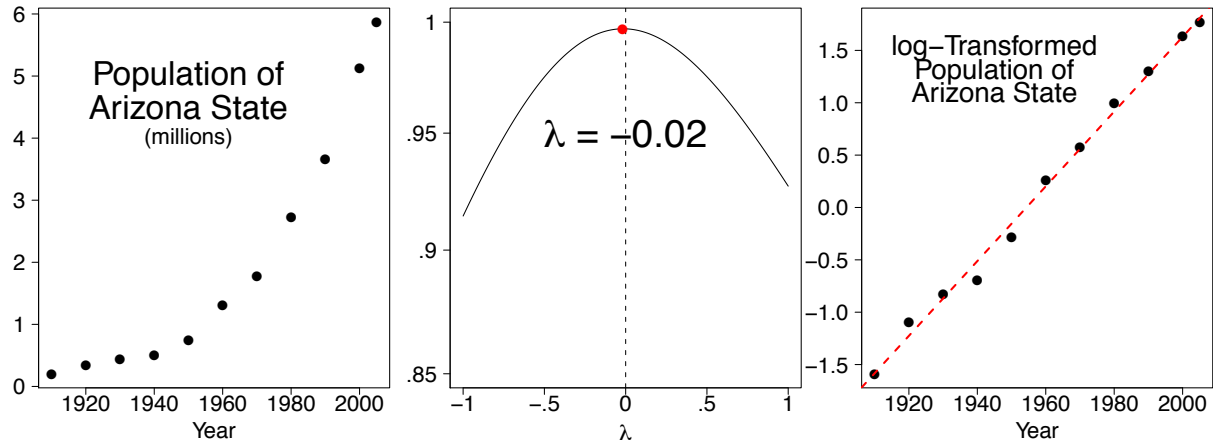


Figure 6. Graphs related to the Arizona state population 1910-2005.

Reducing Skew

Many statistical methods such as t tests and the analysis of variance assume normal distributions. Although these methods are relatively robust to violations of normality, transforming the distributions to reduce skew can markedly increase their power.

As an example, the data in the "Stereograms" case study is very skewed. A t test of the difference between the two conditions using the raw data results in a p value of 0.056, a value not conventionally considered significant. However, after a log transformation ($\lambda = 0$) that reduces the skew greatly, the p value is 0.023 which is conventionally considered significant.

The demonstration in Figure 7 shows distributions of the data from the Stereograms case study as transformed with various values of λ. Decreasing λ makes the distribution less positively skewed. Keep in mind that $\lambda = 1$ is the raw data. Notice that there is a slight positive skew for $\lambda = 0$ but much less skew than found in the raw data ($\lambda = 1$). Values of below 0 result in negative skew.

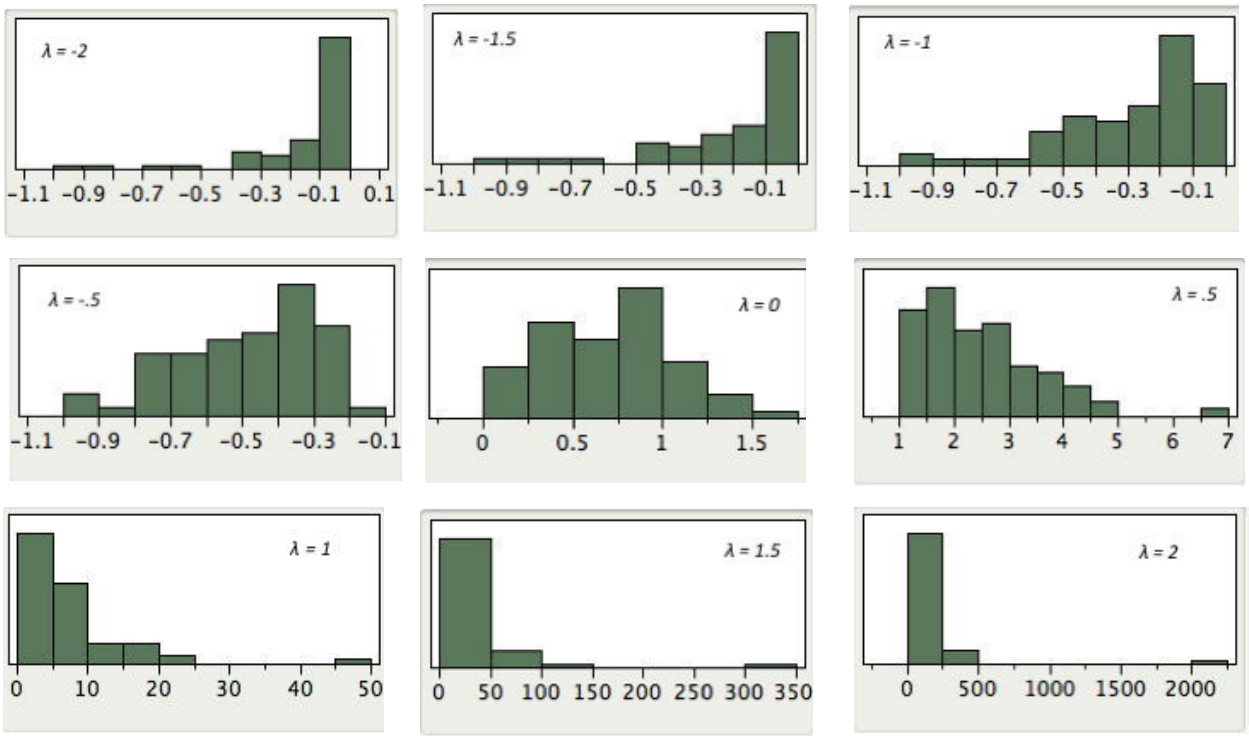


Figure 7. Distribution of data from the Stereogram case study for various values of λ.

Box-Cox Transformations

by David Scott

Prerequisites

This section assumes a higher level of mathematics background than most other sections of this work.

- Chapter 1: Logarithms
- Chapter 3: Additional Measures of Central Tendency (Geometic Mean)
- Chapter 4: Bivariate Data
- Chapter 4: Values of Pearson Correlation
- Chapter 16: Tukey Ladder of Powers

George Box and Sir David Cox collaborated on one paper (Box, 1964). The story is that while Cox was visiting Box at Wisconsin, they decided they should write a paper together because of the similarity of their names (and that both are British). In fact, Professor Box is married to the daughter of Sir Ronald Fisher.

The Box-Cox transformation of the variable x is also indexed by λ, and is defined as

$$x'_\lambda = \frac{x^\lambda - 1}{\lambda} \,. \qquad \texttt{(Equation 1)}$$

At first glance, although the formula in Equation (1) is a scaled version of the Tukey transformation x^λ, this transformation does not appear to be the same as the Tukey formula in Equation (2). However, a closer look shows that when $\lambda < 0$, both x_λ and x'_λ change the sign of x^λ to preserve the ordering. Of more interest is the fact that when $\lambda = 0$, then the Box-Cox variable is the indeterminate form 0/0. Rewriting the Box-Cox formula as

$$x'_\lambda = \frac{e^{\lambda \, \log(x)} - 1}{\lambda} \approx \frac{\left(1 + \lambda \, \log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + \cdots\right) - 1}{\lambda} \rightarrow \log(x)$$

as $\lambda \rightarrow 0$. This same result may also be obtained using l'Hôpital's rule from your calculus course. This gives a rigorous explanation for Tukey's suggestion that the

log transformation (which is not an example of a polynomial transformation) may be inserted at the value $\lambda = 0$.

Notice with this definition of x'_λ that $x = 1$ always maps to the point $x'_\lambda = 0$ for all values of λ. To see how the transformation works, look at the examples in Figure 1. In the top row, the choice $\lambda = 1$ simply shifts x to the value $x-1$, which is a straight line. In the bottom row (on a semi-logarithmic scale), the choice $\lambda = 0$ corresponds to a logarithmic transformation, which is now a straight line. We superimpose a larger collection of transformations on a semi-logarithmic scale in Figure 2.

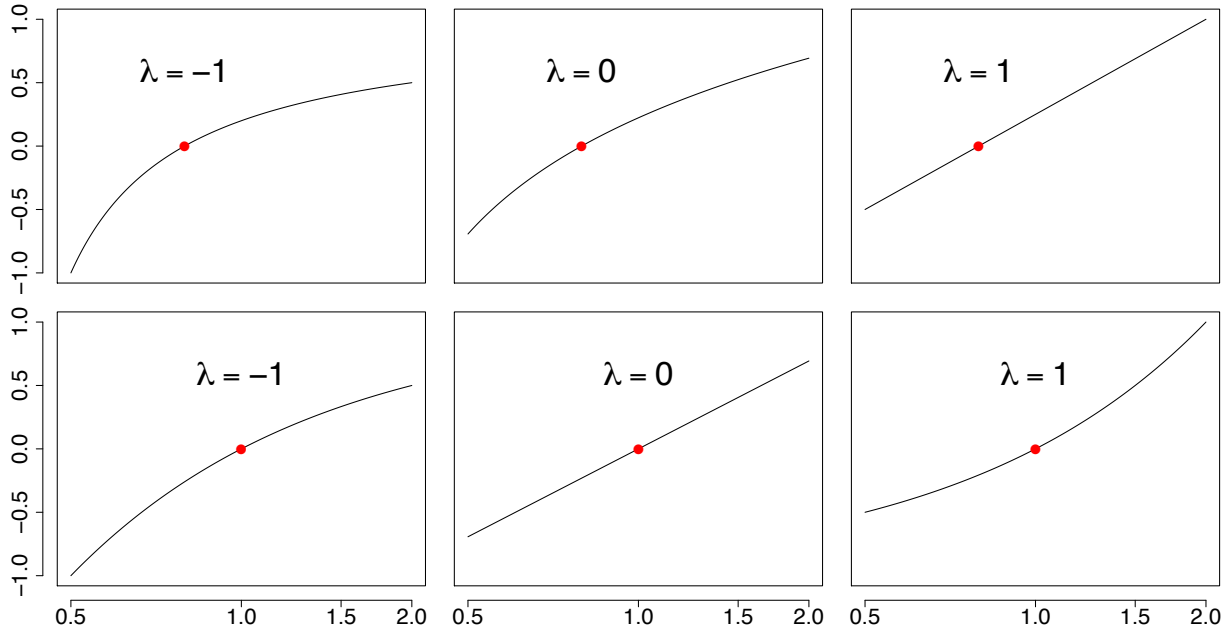


Figure 1. Examples of the Box-Cox transformation x'_λ versus x for $\lambda = -1, 0, 1$. In the second row, x'_λ is plotted against log(x). The red point is at $(1, 0)$.

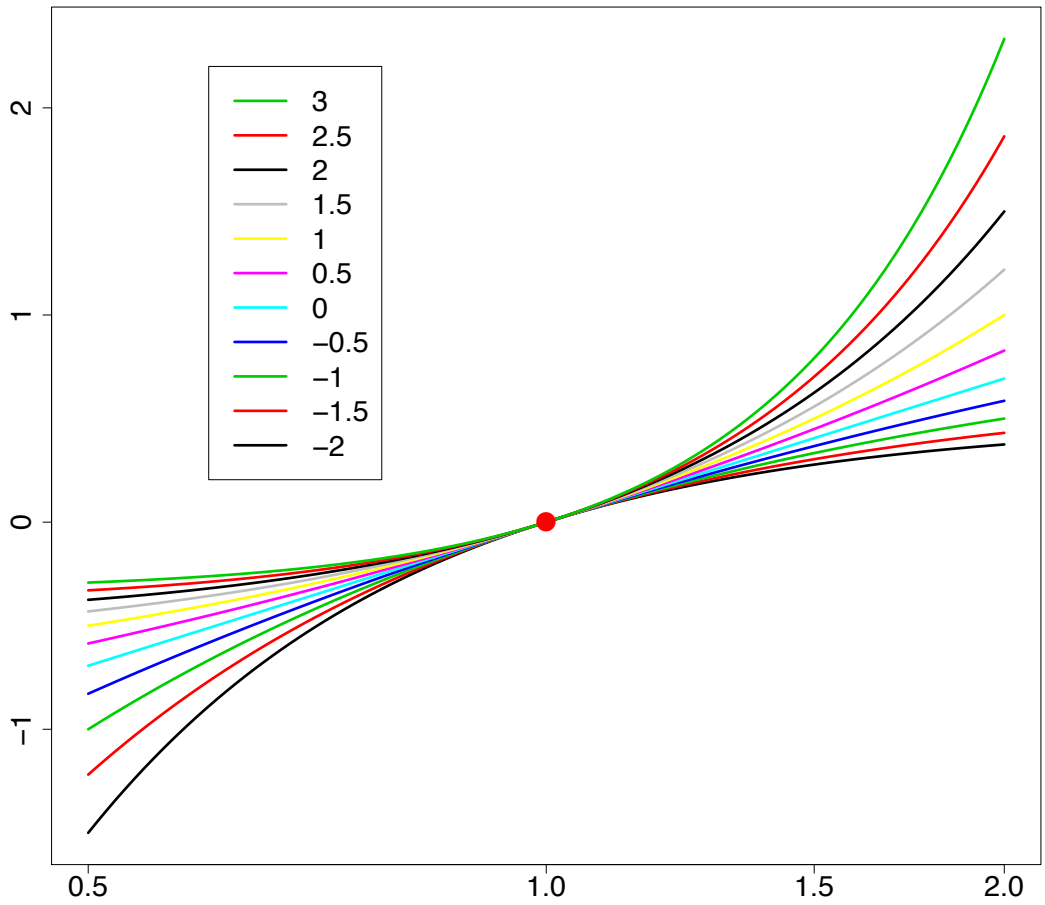


Figure 2. Examples of the Box-Cox transformation versus log(x) for $-2 < \lambda < 3$. The bottom curve corresponds to $\lambda = -2$ and the upper to $\lambda = 3$.

Transformation to Normality

Another important use of variable transformation is to eliminate skewness and other distributional features that complicate analysis. Often the goal is to find a simple transformation that leads to normality. In the article on q-q plots, we discuss how to assess the normality of a set of data,

$$x_1, x_2, \ldots, x_n.$$

Data that are normal lead to a straight line on the q-q plot. Since the correlation coefficients maximized when a scatter diagram is linear, we can use the same approach above to find the most normal transformation.
Specifically, we form the n pairs

$$\left(\Phi^{-1} \left(\frac{i - 0.5}{n} \right), x_{(i)} \right) , \qquad \text{for } i = 1, 2, \ldots, n ,$$

where Φ^{-1} is the inverse CDF of the normal density and $x_{(i)}$ denotes the i^{th} sorted value of the data set. As an example, consider a large sample of British household incomes taken in 1973, normalized to have mean equal to one (n = 7,125). Such data are often strongly skewed, as is clear from Figure 3. The data were sorted and paired with the 7125 normal quantiles. The value of λ that gave the greatest correlation (r = 0.9944) was λ = 0.21.

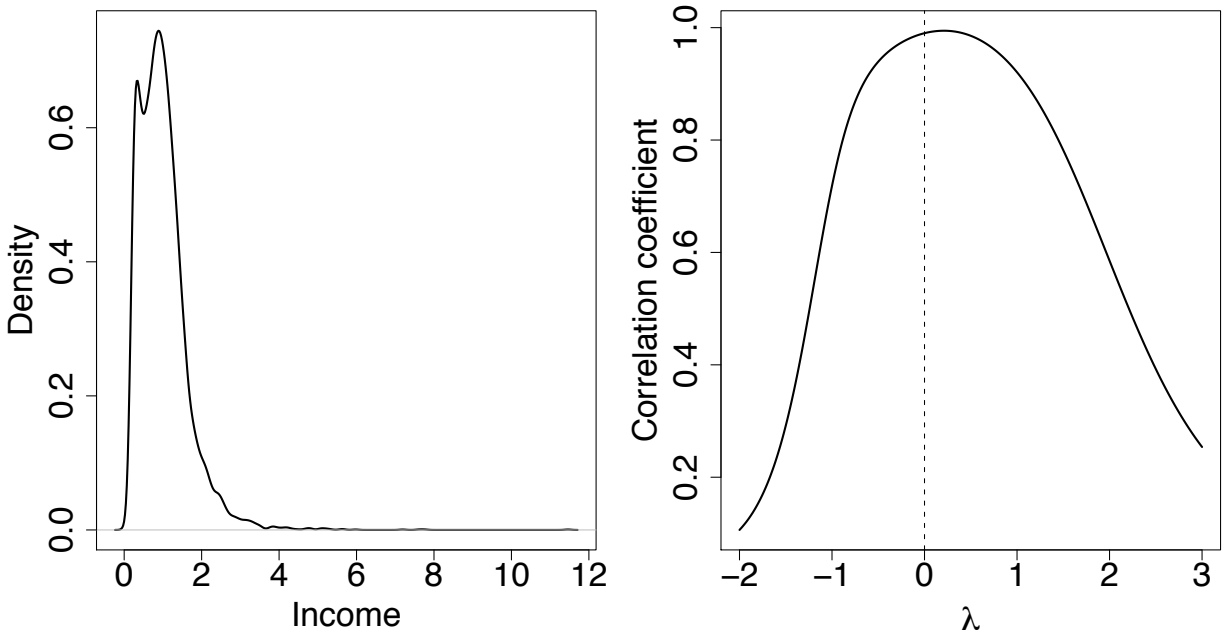


Figure 3. (L) Density plot of the 1973 British income data. (R) The best value of λ is 0.21.

The kernel density plot of the optimally transformed data is shown in the left frame of Figure 4. While this figure is much less skewed than in Figure 3, there is clearly an extra "component" in the distribution that might reflect the poor. Economists often analyze the logarithm of income corresponding to λ = 0; see Figure 4. The correlation is only r = 0.9901 in this case, but for convenience, the log-transform probably will be preferred.

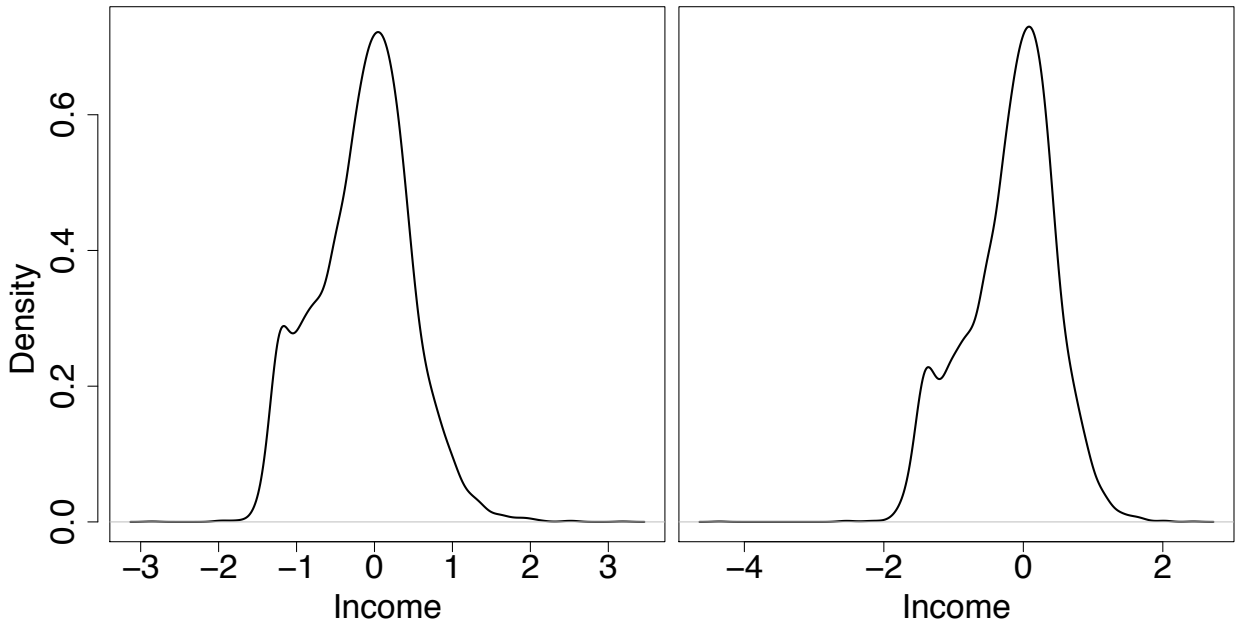


Figure 4. (L) Density plot of the 1973 British income data transformed with $\lambda = 0.21$. (R) The log-transform with $\lambda = 0$.

Other Applications

Regression analysis is another application where variable transformation is frequently applied. For the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

and fitted model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \, ,$$

each of the predictor variables x_j can be transformed. The usual criterion is the variance of the residuals, given by

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \, .$$

Occasionally, the response variable y may be transformed. In this case, care must be taken because the variance of the residuals is not comparable as λ varies. Let $\bar{g}_y$ represent the geometric mean of the response variables.

$$\bar{g}_y = \left(\prod_{i=1}^{n} y_i \right)^{1/n}$$

Then the transformed response is defined as

$$y'_\lambda = \frac{y^\lambda - 1}{\lambda \cdot \bar{g}_y^{\lambda - 1}}$$

When λ = 0 (the logarithmic case),

$$y'_0 = \bar{g}_y \cdot \log(y)$$

For more examples and discussions, see Kutner, Nachtsheim, Neter, and Li (2004).

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 16: Logarithms

Many financial web pages give you the option of using a linear or a logarithmic Y-axis. An example from Google Finance is shown below.

What do you think?

To get a straight line with the linear option chosen, the price would have to go up the same amount every time period. What would result in a straight line with the logarithmic option chosen?

> The price would have to go up the same proportion every time period. For example, go up 0.1% every day.

References

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society*, Series B, 26, 211-252.

Kutner, M., Nachtsheim, C., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models*, McGraw-Hill/Irwin, Homewood, IL.

Tukey, J. W. (1977) Exploratory Data Analysis. Addison-Wesley, Reading, MA.

Exercises

Prerequisites

All Content in This Chapter

1. When is a log transformation valuable?
2. If the arithmetic mean of $\log_{10}$ transformed data were 3, what would be the geometric mean?
3. Using Tukey's ladder of transformation, transform the following data using a λ of 0.5: 9, 16, 25
4. What value of λ in Tukey's ladder decreases skew the most?
5. What value of λ in Tukey's ladder increases skew the most?
6. In the ADHD case study, transform the data in the placebo condition (D0) with λ's of .5, 0, -.5, and -1. How does the skew in each of these compare to the skew in the raw data. Which transformation leads to the least skew?