Section 1.2: Quantifying Sources of Variation

Section 1.2 Learning Goals:

- Partitioning variation in the response variable into variation explained by the model and unexplained variation
- Measuring percentage of variation explained
- Understanding effect size and practical significance

Introduction

In the previous section, we saw how study design impacted sources of variation and revisited how to connect sources of variation with a statistical model. In this section we will build on this intuitive idea of sources of variation, by exploring how to quantify how much variation is explained by various sources. To do this, we will first need to establish mathematical ways to quantify variation, and then see how to standardize these approaches across studies to have consistent ways to talk about contributions of sources of variation across studies. Along the way we will also talk about how to begin to talk about whether a source of variation is meaningful.

Keep in mind the following notation:

- y represents the response variable
- y_i represents the i^{th} observation of the response variable
- $\bar{y}$ represents the mean of the response variable outcomes
- $\bar{y}_j$ represents the mean of the response variable outcomes in the j^{th} group
- n represents the overall sample size in the study
- n_j represents the group size of the j^{th} group

Example 1.2: Scents and Consumer Behavior continued

Recall the Scents and Consumer Behavior study from Example 1.1, which examined students' ratings of a store depending on whether or not a scent was used while they were in the store. Our hypothesized Sources of Variation diagram for this study is shown again in Figure 1.2.1.

Figure 1.2.1: Hypothesized Sources of Variation diagram for Scents/Consumer Behavior Study

Observed Variation in: Favorability ratings (1-7)	**Sources of explained variation**	**Sources of unexplained variation**
	• Scent or not	• Attitude • Scent sensitivity • Understanding of questions • Unknown
Inclusion criteria • Background (Business majors) • Age (20-21 years old) *Design* • Store environment • Time in store		

We said in Example 1.1 that the scent group model explained some of the variation in the favorability ratings because the standard error of the residuals (1.10) was smaller than the

standard error of the residuals from the model that ignored the treatment group assignments (1.27). Let's look at these calculations in a bit more detail.

Sum of Squares Total

Recall that before taking the scent exposure groups into account we attempted to predict the favorability rating from the overall mean:

$predicted\ rating = 4.48$,
standard deviation of ratings = standard error of residuals = 1.27

As discussed in the Preliminaries, the standard error of the residuals for this model is just the standard deviation of the response variable. Another way to think of this value is related to the *total of the squared residuals* when using the overall mean to predict the response for each subject.

Sum of squares total

$$SD\ of\ ratings = \sqrt{\frac{Sum\ of\ all\ squared\ residuals}{n-1}} = \sqrt{\frac{\sum_{all\ obs}(observed\ value - 4.48)^2}{n-1}} = \frac{75.98}{47} = 1.27$$

Definition: The numerator of this calculation is called the ***sum of squares total***, or ***SSTotal***.

$$SSTotal = Sum\ of\ all\ squared\ residuals = \sum_{all\ obs}(observed\ value - overall\ mean)^2$$

Note that we use the symbol $\Sigma_{all\ obs}$ to mean "sum over all observations."

Dividing the Sum of Squares Total (*SSTotal*) by *n* – 1 and taking the square root equals the standard deviation of the response variable. So, why are we discussing *SSTotal*? It turns out that the sum of squares total has some nice properties that will be useful in the future as we seek to understand different sources of variation in the response.

You may recall from your first statistics class that in the SD calculation we divide by *n* – 1 = 47 instead of *n* = 48. This is because these data are considered a sample from some ongoing random process, and the mean that we are comparing each observation to, 4.48, was estimated from the same data. This implies that once we know 47 of the values, the 48th value is "determined" so that the mean of all the values is 4.48. For this reason, we have 47 "independent" pieces of information, and we say there are 47 ***degrees of freedom*** in this calculation. To find the "*average* squared deviation from the mean" we divide by 47 instead of 48.

Definition: The ***degrees of freedom (df)*** for a sum of squares calculation represents the number of "independent" values in the sum.

Sum of Squared Errors for the Separate Means Model

In Section 1.1, we found that a model using the two different group means to make predictions, was better than the single mean model in the sense that the typical prediction error was smaller, 1.10 vs. 1.27.

$$predicted\ rating = \begin{cases} 5.13\ if\ exposed\ to\ scent \\ 3.83\ if\ not\ exposed\ scent \end{cases}$$, standard error of residuals = 1.10 points.

When using the group means to predict each observation (in other words, when using the "separate means" or "scent group" model), the standard error of the residuals is computed as:

$$\sqrt{\frac{\sum_{scent\ group}(scent\ group\ values - 5.13)^2 + \sum_{no\ scent\ group}(no\ scent\ group\ values - 3.83)^2}{n-2}}$$

Sum of squared errors

$$\sqrt{\frac{55.96}{48-2}} = 1.10$$

Definition: The numerator of this calculation is called the ***sum of squared errors***, or ***SSError***. The *SSError* is the sum of the squared prediction errors (residuals) for a particular statistical model.

$$SSError = \sum_{all\ obs}(observed\ value - predicted\ value)^2$$
$$= \sum_{all\ obs} residuals^2$$

When comparing groups, the *SSError* is computed by comparing each observed value to its group mean and captures the variation *within* the two scent exposure groups, or the variation leftover after knowing which scent exposure group the observation came from. Another description for this is the variation that is unexplained by scent group.

Hopefully you are asking: When computing the standard error of these residuals, why do we now divide by 46 instead of 47 or 48? Once we know the mean of each scent exposure group, there are (24 – 1) = 23 independent pieces of information within each group, or 48 – 2 = 46 degrees of freedom in this calculation. Thus, to find the "*average* squared deviation from the group means" we divide by 46. You can also think of the degrees of freedom in terms of the sample size minus the number of estimated parameters in model (e.g., 1 mean vs. the 2 separate means).

Taking the square root of this "average squared deviation" gives us a measure of the average prediction error for the model. When the sample sizes are equal, this is equivalent to averaging the two group variances and taking the square root (see HW exercise). Note that this value will differ slightly from the standard deviation of the residuals, which divides by $n - 1$; that's why we called it the standard *error* instead.

Variation Explained by the Scent Groups

Now, let's examine one more sum of squares value. The *SSTotal* and the *SSError* capture the variation in the observed response from either the overall mean (*SSTotal*) or the treatment group mean (*SSError*). But how much variation is there *between* the treatment groups themselves? In other words, we will measure how much variation there is in the group means by comparing each to the overall mean. First, let's introduce a new term, but with a warning: this new term, *effect*, will be used in this course and in statistics in general, with slightly different variations and meanings.

Definition: The ***effect*** of a group or a treatment is the difference between the mean response in the group and the overall mean response.

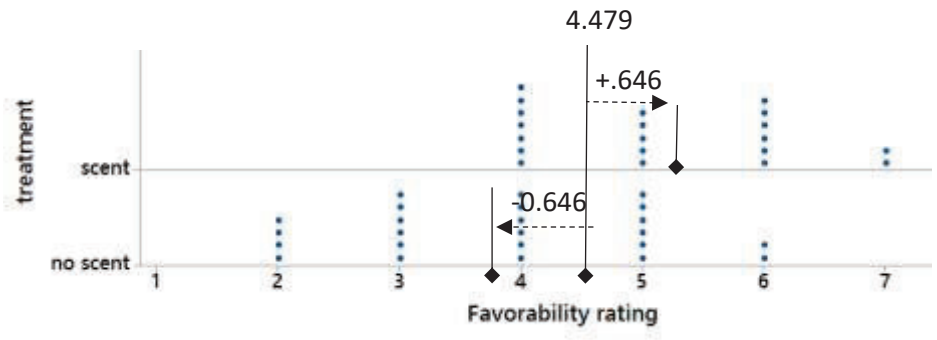
Think about it: How would you calculate the "scent" effect and the "no scent" effect for this study?

In Section 1.1, we found the treatment means to be 5.125 and 3.833. Comparing these group means to the overall mean, we find the *effect* of being in the scent group to be 5.125 – 4.479 = 0.646 points and the *effect* of being in the no scent group to be 3.833 – 4.479 = -0.646 points.

Think about it: What do these two values (effects) tell us?

On average, students in the scent group tend to rate their experience 0.65 points above the overall average, and on average, students in the no scent group tend to rate their experience 0.65 points below the overall average. Figure 1.2.2 displays the data from study with the overall means, the treatment means, and the effects illustrated.

Figure 1.2.2: Dotplots comparing overall ratings given to the store for the two groups



Now that we know what an *effect* is, we can re-write the scent group model in terms of an overall mean and the effects of the scent groups. (See HW exercise discussion how this is equivalent to the earlier version.)

$$Predicted\ Favorability\ rating = 4.48 + \begin{cases} 0.65\ if\ scent\ group \\ -0.65\ if\ no\ scent\ group \end{cases}'$$

SE of residuals = 1.10 points

In other words, each response outcome is modeled as the overall mean + treatment effect + random error. Notice that because we have the same sample size in each treatment group, the effects sum to zero. See the Calculation Details at the end of this section for a slight variation to this calculation when the group sizes are not equal.

Think about it: What will the value of each of the effects be if scent groups do not explain any of the variation in the response variable?

If the effects of being in the two scent groups were both 0, the group means would be identical to the overall mean, and, thus, the scent groups would not explain any variation in the favorability ratings; the two group means would be the same.

Key Idea: The larger the *effects* (in absolute value), the larger the differences between the groups.

To measure how much these scent group means vary from each other (the "between group" variation), we need a measure like the standard deviation of the group means. The numerator

will sum the differences between the group means and the overall mean and the denominator will convey the degrees of freedom of that sum.

Definition: The ***sum of squares for the model***, or ***SSModel***, measures the variation in the group means from the overall mean. For each observation in the data set, we find the difference between that observation's group mean and the overall mean, then sum the squared differences. Because each observation within the same group has the same difference between the group mean and the overall mean, we can simplify the formula to focus on the squared effects and the number of observations in each group (group size).

$$SSModel = \sum_{all\ observations} (group\ mean - overall\ mean)^2 = \sum_{all\ groups} (group\ size) \times (effect)^2$$

For the "scent model," the sum of squares for the model (or the *SSscent*) is

$$SSModel = 24(-0.646)^2 + 24(0.646)^2 = 20.03.$$

The degrees of freedom of this calculation will be 1. This is because once we know the sum of the effects is zero, if we know the effect for one of the groups, we know what the other effect has to be.

To summarize these calculations for this example, we have:

Overall variation in data, *SSTotal* = 75.98, df = 47
Unexplained variation with separate means model, *SSError* = 55.96, df = 46
Variation in the group means, *SSModel* = 20.03, df = 1

Think about it: What relationship do you notice between the three sums of squares, *SSTotal*, *SSError*, and *SSModel*? What about the degrees of freedom?

It's no coincidence that *SSModel* and *SSError* add up to the *SSTotal*- that's one reason we call it *SSTotal*! In other words: 20.03 + 55.96 = 75.99 (any difference is due to rounding). This relationship, and the corresponding one for the degrees of freedom, will always be true.

Key Idea: The variation in the response (as measured by *SSTotal*) can be split up (partitioned) into the variability of interest (as measured by *SSModel*) and the unexplained variation (*SSError*).

SSTotal = *SSModel* + *SSError*

Also, df total = df model + df error.

In general, we are *partitioning* sources for the observed variation in the response variable (measured by *SSTotal*) into two categories:
(1) the source of the variation of interest (the explanatory variable), measured by *SSModel*
(2) the sources of the unexplained variation or the variation which remains within each of the treatment groups, measured by *SSError*.

So how much of the variation in the favorability ratings have we explained using the presence or absence of scent?

One approach to quantify how the variation in the group means or the explained variation (*SSModel*) compares to the overall variation (*SSTotal*), is to compute the "percentage of variation explained."

Definition: R^2 (also known as the ***coefficient of determination***) tells us the proportion of the total variation in the response variable which is explained by the source(s) of interest specified by the model,

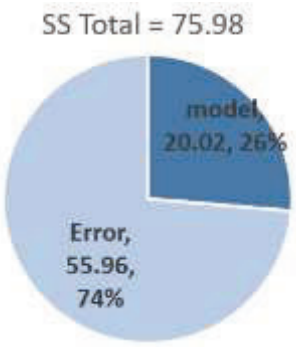
$$R^2 = \frac{SSModel}{SSTotal} = 1 - \frac{SSError}{SSTotal}$$

Note that $0 \leq R^2 \leq 1$. Larger values of R^2 are better because that indicates more of the variation in the response is explained by the source of variation of interest (explanatory variable).

Think about it: How would you compute the R^2 value for the Scents and Consumer Behavior study?

The R^2 for this study is 20.03/75.98 = 0.264. We say that this model (that is, accounting for the presence or absence of scent) explains about 26% of the variation in the observed favorability ratings, whereas about 74% of the variation is not explained by the scents model (55.96/75.98 = 0.7365). R^2 is often reported in decimal form or directly as a percentage. Pie charts can also be useful to visualize the R^2 for a study.

Figure 1.2.3: Partitioning the total variability in the favorability ratings into variation explained by the scent group model and unexplained variation



Think about it: Is 26.4% a "good" R^2 value?

Larger R^2 values indicate less unexplained variation in the response variable and more precise predictions. However, it is unusual to find an R^2 value of exactly 1 (that would mean you had explained all of the variation in the response) or exactly 0 (meaning you had explained none of the variability in the response). An R^2 of 0.264 is certainly not zero, but is not all that large either. There is no set "cut-off" value for what makes an R^2 value meaningful, it is going to differ by context. In a study of physical laws (e.g., the law of gravity), an R^2 below 0.90 would probably indicate a problem. In a study trying to change customer behavior (a complex task for sure), the researchers might be quite pleased at explaining more than 25% of the variation in consumer favorability ratings. In a study trying to understand a new, complex disease, an R^2 of 0.05 might be meaningful as a first step in beginning to understand the disease process. We could compare our R^2 value to other studies on consumer behavior or to models using other

explanatory variables to help decide. We will refer to these types of consideration as evaluating the ***practical significance*** of the study results.

Definition: ***Practical significance*** refers to whether the group differences are large enough to be of value in context. It can be difficult to evaluate practical significance without subject matter knowledge and/or something to which to compare the group differences.

For example, is the effect of scent, ± 0.65, enough to make a store manager take notice? The R^2 value gives us one way to consider practical significance, as it puts the results on a common 0-1 scale that can be compared across models and studies.

Another common comparison is to consider the size of the effects in conjunction with the standard error of the residuals.

Definition: An ***effect size*** measure compares differences in group means to the standard error of the residuals.

You will actually see various measures of effect size in the literature. The key is comparing the variation in your group means to the "natural" variation in the data. We can use the standard error of the residuals as a measure of the natural variation (what's leftover after accounting for the group differences). In this study, the standard error of the residuals, after accounting for the different scent conditions, was 1.10 points. So the difference in group means (1.30) is larger than one residual standard error, which also seems meaningful in this context.

These numerical values (e.g., R^2, effect size) still need to be evaluated in combination with subject matter knowledge to help evaluate practical significance. It's also important to consider the generalizability of the study results and that this was a simulated retail environment. The authors of this article did declare "The presence of an inoffensive scent in a store is an inexpensive and effective way to enhance consumer reactions to the store and its merchandise" but also recommended "careful consideration of cost" as there was some flexibility (found in another study) in which scents were used to achieve the same results.

But, what if a researcher was unhappy with his/her effect size or R^2 value? One consideration, among many, would be to consider improving the statistical model by considering other variables (this is something we did in the Preliminaries – we'll revisit this idea in later chapters). Other options are to do other things to reduce variation in the response variable as we discussed in the previous section – like using inclusion criteria.

Exploration 1.2: Starry Navigation

The movements of dung beetles have fascinated observers for thousands of years. Some species of dung beetles, known as "rollers," find a pile of dung which they form into a ball, and then immediately roll away from the source in order to prevent other beetles from stealing it. The goal is for the beetles to move the ball away as fast as possible. The nocturnal African dung beetle (*Scarabaeus satyrus*) is known to use celestial objects (e.g., sun, moon) to help it move along straighter (quicker) paths so its dung doesn't get stolen. But, what, if it's the middle of the night and the moon isn't out (new moon); can the beetles navigate their way using just the stars? Dacke, Baird, Byrne, Scholtz, and Warrant ("Dung Beetles Use the Milky Way for Orientation," *Current Biology*, 23, 2013) report on several experiments they ran to document whether these dung beetles use stars to navigate. In one of their studies, beetles were placed on top of a dung ball at the center of a circular wooden platform (10 cm in diameter) and the researchers timed how long it took each beetle to reach the edge of the platform (another way of determining how straight a path was taken). Some of the beetles were given a small, black cardboard 'cap' which obscured their view of the sky (up) but not of the edge of the platform (out), while others were given a transparent cap. (Why?). On a moonless, starry night beetles wearing the transparent cap took an average of 40.1 seconds to reach the edge, compared to an average time of 124.5 seconds for beetles wearing the black cardboard cap.

STEP 1: Ask a research question.

1. Summarize the researchers' conjecture in collecting these data.

STEP 2: Design a study and collect data.

2. Explain how this is an experiment rather than an observational study. Identify the response variable and the explanatory variable. What are the treatments?

3. Identify at least one component of the study protocol that was important in ensuring consistent and accurate measurements across the beetles.

STEP 3: Explore the data.

One hypothesized Sources of Variation diagram for this study is shown in Figure 1.2.4.

Figure 1.2.4: Possible Sources of Variation diagram for Starry Nights study

Observed Variation in: Time to reach edge (sec)	**Sources of explained variation** • Type of cap	**Sources of unexplained variation** • Age of beetle • Gender of beetle • Unknown
Inclusion criteria • Beetle species *Design* • Size of platform		

4. Based on the averages provided (124.5 seconds with black cap and 40.1 seconds with transparent cap), are you convinced that obscuring the beetles' vision of the night sky causes them to have more trouble moving in a direct line away from the starting position? If not, what other information about the data would you like to know?

The difference in means 124.5 – 40.1 = 84.4 seconds sure seems large, but we need to know more about how much variation there is from beetle to beetle. If the longest time a beetle takes to reach the edge of the circle is more than 500 seconds, then 84 seconds might not seem so large.

In the dataset ***DungBeetles***, we provide data for 18 beetles (9 which wore the black cap and 9 which wore the transparent cap). **Note**: The researchers did not provide the exact data in their publication, this file contains simulated data similar to what the researchers observed.

The Single Mean Model

As we saw in the previous section, before taking the type of cap (black or clear) into account, we can predict the dung ball rolling time using the overall mean (the *single mean model*).

Load the data into the **Multiple Variables** applet and drag the *time* variable into the **Response** variable box. Check the **Show descriptive** and **Show residuals** boxes.

5. Record the overall mean and standard deviation for the times. Use these values to write out a "single mean" statistical model for predicting the time to reach the edge.

Prediction equation:

Standard error of residuals:

The standard error of the residuals from this "empty" model is the standard deviation of the times themselves:

$$SD\ of\ times = \sqrt{\frac{Sum\ of\ all\ squared\ residuals}{n-1}} = \sqrt{\frac{\sum_{all\ obs}(observed\ value - 84.66)^2}{18-1}} = \frac{37441.6}{17} = 46.93 \text{ sec}$$

Definition: The numerator of this calculation is called the ***sum of squares total*****,** or ***SSTotal***.

$$SSTotal = Sum\ of\ all\ squared\ residuals = \sum_{all\ obs} (observed\ value - overall\ mean)^2$$

We will use the *SSTotal* as one representation of the total variation in the residuals from the single mean model or just the total variation in the response variable. Note that we divide this sum by 17 rather than 18 because we are using the sample mean in the same calculation, so once we know 17 of the values, we know what the 18th must be. So we say this calculation has 17 ***degrees of freedom***. Note that we use the symbol $\Sigma_{all\ obs}$ to mean "sum over all observations."

Definition: The ***degrees of freedom*** for a sum of squares calculations represents how many "independent" values are being summed over.

6. Confirm that $(n - 1) \times (SD\ of\ times)^2 = SSTotal$ (reported by the applet under the histogram).

Sum of Squared Errors for the Separate Means ("Cap") Model

Now drag the treatment variable into the **Subset By** box. Report the means of the treatment groups.

7. Visually, does the type of cap appear to explain variation in the times?

Check the box to **Show residuals** and note the standard error of the residuals. Include a screen capture of the results.

8. Write out the statistical model using the group means to predict the times. How does the standard error of the residuals for the "cap model" compare to the single mean model?

The standard error of the residuals for this model, taking into account the type of cap, is calculated by comparing each observation to its group mean, that is, the residual from using the group mean to predict an observation in that group.

$$SE\ of\ residuals = \sqrt{\frac{\sum_{clear}(observed\ value - 42.78)^2 + \sum_{black}(observed\ value - 126.55)^2}{18 - 2}}$$

Definition: The numerator of this calculation is called the ***sum of squared errors*****,** or ***SSError***. The *SSError* represents totaling the squared predictions errors (residuals) for a particular statistical model.

$$SSError = \sum_{all\ obs} (observed\ value - predicted\ value)^2$$
$$= \sum_{all\ obs} residuals^2$$

The *SSError* represents the leftover variation in the response variable after conditioning on the treatment group, that is the unexplained variation within the treatment groups. Notice that this time we are dividing by 18 – 2 = 16. This reflects that we have used both of the group means in our calculation. Previously we only used the overall mean and divided by 18 – 1. See the Example for more details on these ***degrees of freedom*** values.

Key Idea: The ***degrees of freedom*** for a sum of squares calculation will be the sample size minus the number of estimated parameters in the model.

Taking the square root of this "average squared deviation" gives us a measure of a typical prediction error for the model. When the sample sizes are equal, this is equivalent to averaging the two group variances and taking the square root (see HW exercise). In other words, it is the "pooled" (across the groups) "within group variation." Another phrasing for this is the variation unexplained by the type of cap. Note that this value will differ slightly from the standard deviation of the residuals which divides by $n - 1$, that's why we called it the standard *error* instead.

9. Verify that $(n - 2) \times (SE\ residuals)^2 \approx SSError$ (given in the pie chart of the applet).

Variation Explained by the Cap Groups

Let's examine one more sum of squares value. Rather than computing the difference between the observed response and what we predict based on the treatment group, we will compare what we predict based on the treatment group to what we would predict if we ignored the treatment group. In other words, we will measure how much variation there is in the group means by comparing each to the overall mean. First, let's introduce a new term, but with a warning: this new term, *effect*, will be used in this course and in statistics in general, with slightly different variations and meanings.

Definition: The ***effect*** of each treatment is the difference of the mean response in the treatment group from the overall mean response.

10. Calculate the cap effect and the "no cap" effect. **Note:** When computing effects make sure that you subtract the overall mean from the group mean in both cases. How do the two effects compare to each other?

When the effects are defined this way, they will always sum to zero (except possibly for round off error). See the Calculation Details at the end of this section for a slight variation to the calculation for unequal group sizes.

11. Using the overall mean and these treatment effects, suggest another way we can write out the statistical model.

To measure how much these gap group means vary from each other (the "between group variation"), we need a measure like the standard deviation of the group means. The numerator

will sum the differences of the group means to the overall mean and the denominator will convey the degrees of freedom of that sum.

Definition: The ***sum of squares for the model***, or ***SSModel***, measures the variation in the group means from the overall mean. For each observation in the data set, we find the difference between that observation's group mean and the overall mean, then sum the squared differences. Because each observation within the same group has the same difference between the group mean and the overall mean, we can simplify the formula to focus on the squared effects and the number of observations in each group.

$$SSModel = \sum_{all\ observations} (group\ mean - overall\ mean)^2 = \sum_{all\ groups} (group\ size) \times (effect)^2$$

12. Calculate the *SSModel* (or "*SScap*") for these data. (*Hint*: What is the group size in each group?)

These sums of squares calculations have a very special property.

Key Idea: The variation in the response (as measured by *SSTotal*) is partitioned into the variability of interest (as measured by *SSModel*) and the unexplained variation (*SSError*).

SSTotal = *SSModel* + *SSError*

Also, df total = df model + df error

13. Verify these two identities for our data.

The *SSModel* is interpreted as a measure of the "variation in the response explained by the model." So we have *partitioned* the total variation in the times (*SSTotal*) into variation explained by the model (*SSModel*, from knowing the treatment) and the variation left unexplained (*SSError*).

14. Calculate the ***percentage of variation explained*** for these data.

$$\left(\frac{SSModel}{SSTotal} = 1 - \frac{SSError}{SSTotal}\right) \times\ 100\%$$

Definition: R^2 (also known as the ***coefficient of determination***) tells us the proportion of the total variation in the response variable which is explained by the source(s) of interest specified by the model. The maximum value of R^2 is 1 and larger values are better (more of the variation in the response is explained by the variable of interest).

$$R^2 = \frac{SSModel}{SSTotal} = 1 - \frac{SSE}{SSTotal}$$

15. Write a one-sentence interpretation of this value, in context.

Pie charts can also be useful to visualize the R^2 for a study.

16. Copy and paste the pie chart from the applet. Notice that one slice of the "pie" represents the variation due to the model (cap type), and the remaining slice represents the "unexplained variation." The size of the "cap type" slice divided by the *SSTotal* gives the R^2.

In the next section, we will look at methods for deciding whether this amount of variation explained is *statistically significant*. For now, we will consider whether this research result is ***practically significant.***

Definition: *Practical significance* refers to whether the treatment effects and group differences are large enough to be of value in context. It can be difficult to evaluate practical significance without subject matter knowledge and/or something to compare to.

One way to assess practical significance is to compare the difference between the groups to the "leftover" or unexplained variation.

Definition: An ***effect size*** measure compares differences in group means to the standard error of the residuals.

17. Calculate the difference in the two treatment means divided by the standard error of the residuals. Is this larger than one or two? [Often values larger than one or two are considered noteworthy...]

STEP 5: Formulate conclusions.

18. Summarize what you have learned so far from this study, in context. For example, do you find the difference in times impressive? How are you deciding? Do you think there could be

any confounding variables or alternative explanations for why the beetles traveled faster with the clear cap?

STEP 6: Look back and ahead.

19. Suggest at least one way you would improve this study if you were to carry it, or a follow-up study, out yourself.

Section 1.2 Summary

It's typically not enough to simply identify sources of variation in a study. In most cases, researchers also want to be able to quantify how much variation is explained by different sources and how much is still unexplained. The standard error of the residuals is one way to begin to quantify unexplained variation. In the simplest case, the single mean model, the residuals reflect what would happen if we simply used the mean response to predict each data value. When considering other sources of variation, we will want to assess how much smaller they make the standard error of the residuals compared to this simple model. Reporting R^2, which tells us the proportion of all the variation (e.g., the Sum of Squares Total) is explained by the variable(s) of interest (Sum of Squares Model), gives us a quantity that can be readily compared across models and studies. In this section, we discussed some (debatable) ways to decide whether the R^2 value for a particular study is of practice importance, a key consideration. But, *practical significance* is not the only consideration. As you may remember from your first statistics course, *statistical significance* is also important. We'll dig into statistical significance in the next section.

Section 1.2 Calculation Details

The dataset analyzed in Exploration 1.2, ***DungBeetles***, claimed that there were 9 beetles in each group. In the actual study, there were 13 beetles wearing transparent caps and 6 beetles wearing the cardboard cap. Does having unequal group sizes impact our calculations? For the most part the answer is no. For example, in calculating the standard error of the residuals, we already allowed the comparisons to the group means to sum over different numbers of observations.

$$SE\ of\ residuals = \sqrt{\frac{\sum_{clear}(observed\ value - 42.78)^2 + \sum_{black}(observed\ value - 126.55)^2}{18-2}}$$

However, we will calculate the *effects* slightly differently in this case. Recall that *effect* was defined as the difference between the group mean and the overall mean. The two group means were 42.78 and 126.55. When the two groups have the same sample size, then the overall mean was equal to the average of the two groups means $(9 \times 42.78 + 9 \times 126.5)/18 = (42.78 + 126.5)/2 = 84.66 = \bar{y}$. But when the two groups do not have the same sample size, $(13 \times 42.78 + 6 \times 126.5)/19 = 69.22 \neq (42.78 + 126.5)/2 = 84.66$. The "weighted" average is much closer to the 42.78 average of the larger group. If your primary goals is to use the "mean response" to predict a beetle time, then you might prefer to use the average that you consider more precise because it is based on a larger number of observations, rather than an average that treats the two group means equally. However, in defining effects, partly so they always sum to zero, we will use the unweighted mean, called the "least squares mean."

Definition: With unequal group sizes, the ***least squares mean*** of the response variable is still $(\bar{y}_1 + \bar{y}_2)/2$ and the *effects* are the group means compared to this value.

The effects version of our prediction equation is

$$Predicted\ time = 84.66 + \begin{cases} 41.9\ if\ obscured \\ -41.9\ if\ sees\ sky \end{cases}$$

which is still equivalent to using the group means to make the prediction for each group. With this definition for effects, our calculations proceed as before, but we can no longer take the "effects squared" shortcut in calculating *SSModel* (but we can still take the *SSTotal* – *SSError* shortcut).

$$SSModel = \sum_{all\ groups} (group\ size) \times (group\ mean - overall\ mean)^2$$

$$\neq \sum_{all\ groups} (group\ size) \times (effect)^2$$

Dacke et al. (2013) report that the path lengths of beetles rolling with the black caps had mean 124.5 sec and standard deviation 30.76 sec (n = 6). The path lengths of beetles using the clear caps was 40.1 sec, with standard deviation 15.3 sec (n = 13). From this information, we can find:

$\bar{y} = (6(124.5) + 13(40.1))/19 = 66.75$
SSError = $(6\text{-}1)(30.76^2) + (13\text{-}1)(15.3^2) \approx 7540$ (df = 17)
SSModel = $6(124.5 - 66.75)^2 + 13(40.1 - 66.75)^2 \approx 29{,}243$ (df = 1)

Section 1.3: Is the Variation Explained Statistically Significant?

Section 1.3 Learning Goals:
- Assess the statistical significance of a two-group comparison
- Carry out and evaluate a randomization test comparing two groups on a quantitative response variable
- Apply two-sample *t*-procedures for tests of significance and confidence intervals

Introduction

Sections 1.1 and 1.2 were all about describing and quantifying the variation in the study and using a statistical model to make predictions and provide some measure of the accuracy of those predictions (Steps 1-3 of the six-step process). Section 1.2 ended by thinking of ways we could decide whether we have explained "a lot" of variation by considering the *practical significance* of the effects. Another consideration centers on the idea of *statistical significance*. In other words, is chance a plausible explanation for the data we observed? Is the difference not only practically meaningful but also beyond what we would expect to happen by chance? Can we say that an observed difference is unlikely to have happened 'just by chance'? In this section, we will look at some different strategies to answer this question.

Terms we assume you saw in your previous statistics course include:
- *Parameter vs. statistic*: Numerical characteristics (e.g., mean or proportion) of the population and sample respectively
- *Null and alternative hypotheses*: Two competing claims about the population or underlying process. Using the null hypothesis is the uninteresting case (e.g., no effect, no difference) and the alternative hypothesis is usually what the researchers are hoping to provide evidence for
- *p-value*: The probability of observing a statistic at least as extreme as the value observed in the actual study when the null hypothesis is true
- *Statistical significance*: When the observed statistic is different enough from the null hypothesis that we don't think it happened by random chance alone
- *Confidence intervals*: An interval of plausible values of the parameter based on the observed statistic

Example 1.3: Scents and Consumer Behavior cont.

Recall the Odor and Consumer Behavior study from Example 1.1. We found that *SSTotal* was approximately 76.0, *SSGroups* ≈ 20.0, and the *SSError* ≈ 56.0. With effects of $\pm$ 0.65, the model using scent exposure to explain variation in ratings found that scent exposure explained about 26.4% of the observed variation in favorability ratings, leaving about 73.6% unexplained. Furthermore, this was a randomized experiment, so we have reason to believe the scent exposure treatment may be *causing* this variation. But, is there another plausible explanation?

Think about it: Is it possible that a difference in means of 1.30 points and an R^2 value of 0.264 could result not because the scent has an impact on ratings, but because we just happened, by chance alone, to have been a little "unlucky" in the random assignment and ended up with more of the students who tend to give out higher favorability ratings in general

in the scent group. Or is 26.4% large enough that we have trouble believing an R^2 value this large could have happened purely by "luck of the draw"?

Recall that in Example 1.1 we did Steps 1–3 of the 6-step statistical investigation method, so we'll just dig into Step 4 here.

STEP 4: Draw inferences beyond the data.

To help answer the question of whether chance is a reasonable explanation, we want to consider how the study could have turned out if there really was no actual "scent effect" and we randomly assigned the treatment groups.

Key Idea: In assessing statistical significance, we typically define null and alternative hypotheses.

- The *null hypothesis* (H_0) is the "by chance alone" explanation for the observed results
- The *alternative hypothesis* (H_a) typically corresponds to the research conjecture (e.g., the imposed treatments explain variation in the response variable)

For this study, we can state the null hypothesis a couple of different ways. For example:

H_0: there is no underlying association between scent condition and favorability ratings
H_a: there is an underlying association (our research conjecture)

If we assume that the only impact of imposing the scent condition is to shift up/down students' ratings, these hypotheses are equivalent to:

H_o: the underlying treatment means are equal ($\mu_{\text{no scent}} - \mu_{\text{scent}} = 0$ where μ_{scent} refers to the underlying treatment mean for someone in the scent condition and $\mu_{\text{no scent}}$ refers to the underlying treatment mean for someone in the no scent condition)
H_a: the underlying treatment means are not equal ($\mu_{\text{no scent}} - \mu_{\text{scent}} \neq 0$)

Note: We have specified a *two-sided alternative* here, saying "there is a difference," rather than a *one-sided alternative* which would predict which treatment mean was larger. Also note that the null hypothesis says our "single mean" model is adequate, whereas the alternative hypothesis includes our "separate means" model.

We never get to observe the "true" difference in treatment means, we only get to estimate it from the sample data, such as the observed difference in group means. In this study, we found a difference of 1.30, but what if the underlying treatment means were actually equal (no genuine association between scent condition and favorability ratings) and the random assignment had turned out differently, might we still see a value as large as 1.30? In other words, is 1.30 a typical outcome for the difference in the group means if the treatment means were actually the same?

You may recall from your first course that a *parameter* summarizes the population or process, but a *statistic* is what we calculate from the observed sample data.

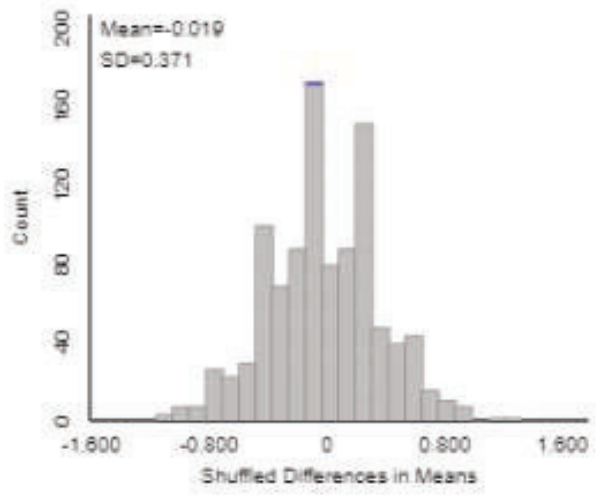
For example, here we could consider 1.30 (the difference in group means) or 0.65 (absolute value of effect) or 0.264 (R^2) as a statistic. The null hypothesis was about the "underlying treatment means" ($\mu_{\text{no scent}} - \mu_{\text{scent}}$) which we could consider the parameter of interest.

To see what could happen if the null hypothesis was true, we will assume that each student in the study would have given the same favorability rating of the store no matter which treatment they had been assigned to (the underlying difference in treatment means is zero). Then randomly mix or "shuffle" the students into two groups of 24 and find the new difference in group means – see what types of values of the statistic we might see if the random assignment worked out differently. In such a "could have been" situation, when we compare the two "simulated" group means, we know that the difference between the groups is simply due to random chance in the random assignment.

Definition: A ***randomization test*** assumes the null hypothesis to be true and examines all possible re-random assignments of the observed responses among the groups, recalculating the statistic each time. Instead of finding all possible arrangements, we can repeat the process a large number of times to approximate the ***null distribution*** of the statistic.

Using the **Multiple Groups** applet, we can randomly shuffle the observed favorability ratings, reassigning them to two groups of 24, and calculate a new difference in means each time. Figure 1.3.1 shows the null distribution for 1,000 such *differences in group means.* Keep in mind that these simulated statistics are generated assuming there is nothing special about being in the scented group or the no scent group with regard to how one would rate the store (the pleasant scent doesn't impact ratings, on average). In this null distribution, any variation in the group means is due entirely to the re-randomization process.

Figure 1.3.1: Null distribution of 1,000 *differences in means* for Scents and Consumer Behavior study

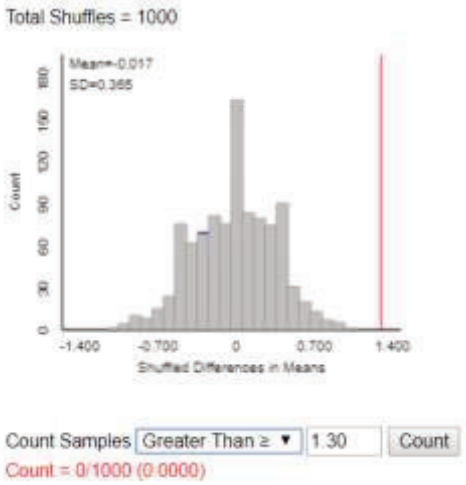


Think about it: What do you learn from the graph Figure 1.3.1?

This distribution is fairly symmetric, bell shaped and centered near zero. This makes sense because in the long run, which group the response value is assigned does not change the response value, so we expect group effects of about 0, and *differences in means* close to 0, and the random shuffling is equally likely to give us a difference in means above 0 as below 0. But the main thing we learn from this distribution is how large the difference in means values can be when we know the null hypothesis is true. A difference of 1.30, as the researchers found in their

study, appears rather unlikely to happen by chance or an unlucky random assignment alone, because 1.30 is far in the tail of the null distribution. One way to quantify how unusual such a value is, is to determine how many simulated values were just as extreme, or even more extreme (even stronger evidence against the null hypothesis) than what the researchers observed. Figure 1.3.2 shows the null distribution counting the number of shuffles with difference in means $\geq$ 1.30.

Figure 1.3.2: Null distribution for 1,000 differences in means, counting how many are 1.30 or larger



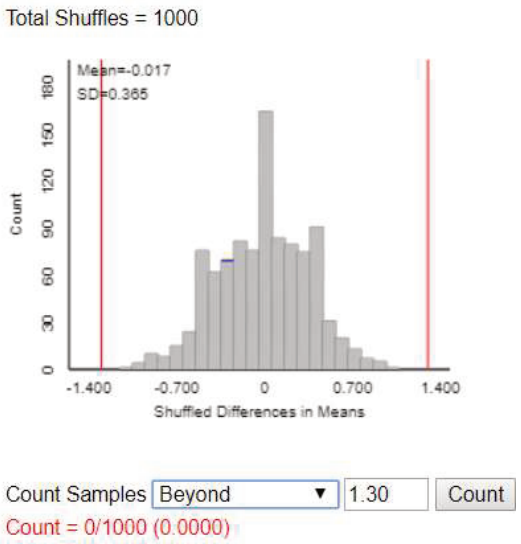
None of the 1,000 shuffles returned a difference in means value at least as large as 1.30.

Key Idea: The *p-value* for a randomized experiment is how often random assignment alone (assuming the null hypothesis is true) could have produced a statistic at least as extreme as the statistic found in the actual study. A small p-value (e.g., below 0.05) constitutes strong evidence against the null hypothesis of "random chance alone," with even smaller values (e.g., below 0.01 or 0.001) providing even stronger and stronger evidence against the null hypothesis. When the p-value is small, we say that the observed difference is *statistically significant*, meaning the observed statistic is unlikely to have happened by random chance alone.

Recall that we specified a two-sided alternative. A corresponding *two-sided p-value* would look for results at least as extreme as 1.30 or -1.30, just as far from zero (what we expect under the null hypothesis) in the other direction (Figure 1.3.3).

For this simulation, we still don't find any values as extreme as the observed statistic. It's possible that if we had done many, many more re-randomizations, we might occasionally see a re-random assignment with a difference in means of 1.30 or larger or -1.30 or smaller. But, since we didn't find such values in these 1,000 repetitions, we estimate this happens less than 1 in 1,000 shuffles. So, our estimated (two-sided) p-value is < 0.001. Because this value is so small (e.g., smaller than a cut-off like 0.05), we conclude that we have strong evidence that the observed difference in means did not arise through the random assignment process alone.

Figure 1.3.3: Null distribution for 1,000 differences in means, estimating a two-sided p-value



We will call this the ***3S Strategy for measuring strength of evidence***:

1. **Statistic:** Compute a statistic from the observed sample data which measures the comparison of interest (e.g., difference in group means).
2. **Simulate:** Identify a "by-chance-alone" explanation for the data (the null hypothesis). Then use a computer to repeatedly simulate values of the statistic, *mirroring the randomness of the study design*, that could have happened if the chance explanation is true.
3. **Strength of evidence:** If the observed statistic is unlikely to have occurred when the chance explanation is true, then we say we have "strong evidence" against the reasonableness of chance alone as an explanation for the study results.

Note that the values we consider "at least as extreme" as the observed statistic in determining the p-value will depend on the direction of the alternative hypothesis and whether it is one-sided or two-sided.

Other Choices of Statistics

The randomization test we just conducted can easily be carried out with other statistics as well. For example, we could look at the R^2 value. But, recall that R^2 values can be difficult to evaluate and assess on their own. However, there are some statistics (***standardized statistics***) that are helpful in judging *statistical significance*, even before looking at the p-value.

A very handy result about using the difference in means as the statistic is we can predict the shuffle-to-shuffle variation in this statistic in advance, without carrying out the simulations.

Definition: We can predict shuffle-to-shuffle variation in the difference in means $(\bar{y}_1 - \bar{y}_2)$ without doing the shuffling. In particular, $\text{SE}(\bar{y}_1 - \bar{y}_2) \approx (SE\ of\ residuals) \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

This SE formula predicts a standard deviation of $1.10 \times \sqrt{2/24} \approx 0.32$, in the ballpark of our simulation results in Figure 1.3.3. (See HW Exercise about why it's not even closer in this case.)

Notice this standard deviation predicts the shuffle to shuffle variation in the difference in group means. In comparison, *SE of residuals* estimates the person to person variation in ratings after adjusting for the treatment group. We can use this result to standardize the observed difference in means.

Definition: A ***standardized statistic*** considers the random variation in the statistic arising from the randomness in the data collection process, which will depend on the natural variation in the data and the sample sizes.

Standardized statistic = statistic / variation in statistic

When you are willing to assume the population standard deviations of the two groups are the same, the ***pooled t-statistic***, assuming no difference in the underlying treatment means, is

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{(SE\ of\ residuals) \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Another formula you may have seen in your previous statistics course is the "unpooled *t*-statistic":

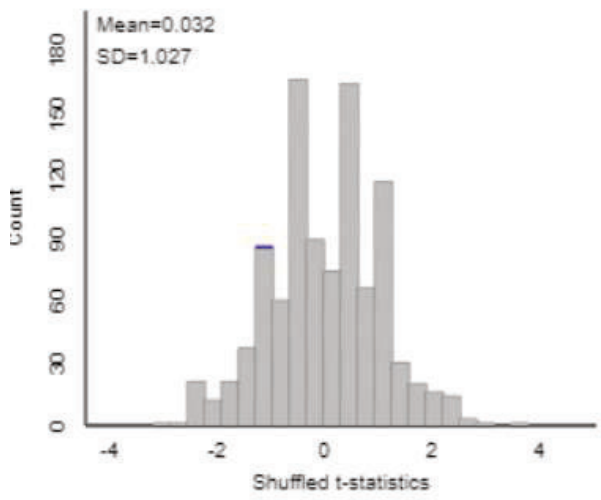
$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This formula is preferred when you don't want to assume the population standard deviations are equal, but the pooled version generalizes more easily to comparing several groups as we will see in the next section. The point is, when standardizing, we are dividing not just by the unexplained variation in the data but also taking the sample sizes into account.

Think about it: What would we change in the 3S strategy to assess the statistical significance of the pooled *t*-statistic? Predict how the new null distribution will behave.

To assess the statistical significance of the pooled *t*-statistic, all we need to change is calculating the *t*-statistic after each shuffle (rather than the difference in group means). Figure 1.3.4 shows the null distribution for 1,000 shuffled pooled *t*-statistics.

Figure 1.3.4: Null distribution for 1,000 pooled *t*-statistics



This distribution looks similar to the distribution of the difference in group means statistic; it is also centered at zero, but the variability is quite different. To estimate the strength of evidence against the null hypothesis, we first need to calculate the observed *t*-statistic for the study data.

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{(std\ error\ of\ residuals)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{1.29}{1.10\sqrt{\frac{1}{24}+\frac{1}{24}}} = 4.03.$$

Reviewing Figure 1.3.4, we see that this value of 4.03 is once again in the far right tail of the null distribution, again estimating a p-value < 0.001.

Think about it: What additional information/value does the standardized statistic provide?

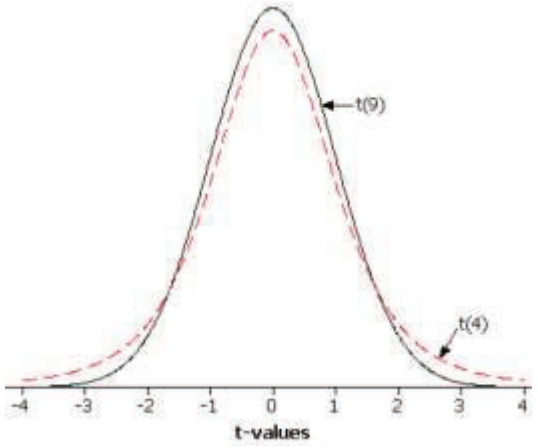
One advantage of a standardized statistic is that it reflects the sample sizes in the study. It also, like R^2, gives you a "unitless" measure that can be compared across studies. For example, a *t*-statistic larger than 2 is generally going to give you a p-value below 0.05; and a *t*-statistic larger than 3 generally gives a p-value less than 0.001. Another advantage is that when certain ***validity conditions*** are met, the null distribution can be approximated by a mathematical model – meaning that you don't have to do a simulation (e.g., lots of shuffles) to approximate the null distribution – you can use mathematics to predict what would happen if you were to shuffle.

Validity Conditions: For the pooled *t*-statistic, when comparing two population means, if
(1) the samples are independent of each other,
(2) the sample standard deviations are roughly equal (e.g., the larger SD is not more than twice the size of the smaller), and
(3) the sample distributions are roughly symmetric or both sample sizes are at least 20 without strong skewness or outliers in the distributions,
then we can approximate the null distribution of the *t*-statistic with a ***t-distribution*** with (*total sample size* – 2) *degrees of freedom.* (This probability distribution was "discovered" by W. S. Gosset and published under the name "Student" in 1908.)

To check these conditions for a particular data set, we need to
(1) consider the data collection method: This study used random assignment so we consider this condition met,
(2) examine the two sample standard deviations: 0.947 and 1.239, which seem similar (we can consider this condition met if the ratio of larger to smaller is less than 2),
(3) examine graphs of the sample data: Figure 1.1.2 showed distributions that were discrete (had spaces between the possible values) but relatively symmetric with no extreme outliers, so we will consider this condition met. Looking at these graphs is equivalent to examining a graph of the residuals and seeing that the overall shape of the distribution of the residuals is approximately normal (See Figure 1.1.3).

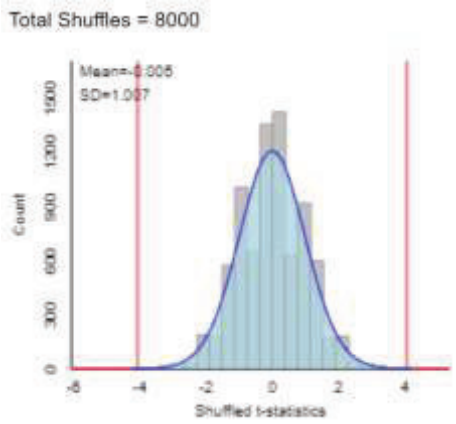
So it makes sense to use the *t-distribution* with 48 – 2 = 46 degrees of freedom as a mathematical approximation to the null distribution of the *t*-statistics for this study. It's no coincidence that this degrees of freedom value matches the degree of freedom for the *SSError* calculation. In this case, the degrees of freedom tells you which *t* distribution to use, because there is actually a family of *t*-distributions, characterized by their *degrees of freedom.*

Figure 1.3.5: Example *t*-distributions with 4 and 9 degrees of freedom



Once we have an appropriate mathematical approximation for the null distribution, the p-value is approximated as the area under the *t*-distribution for the *t*-values at least as extreme as the observed *t*-statistic. We will often refer to such a p-value as a ***theory-based*** p-value to distinguish it from the simulated p-value based on the simulation of the null distribution.

Figure 1.3.6: Theoretical *t*-distribution overlaid on shuffled *t*-statistics, shading the area representing the theory-based p-value



The theory-based (two-sided) p-value is 0.0002, similar to our simulation results. In this case, we would interpret this p-value as how often we would get a *t*-statistic at least as extreme as 4.06 if there really was no underlying treatment effect from the scent ($\mu_{\text{no scent}} - \mu_{\text{scent}} = 0$). All that has changed in our interpretation is the choice of statistic, which should be noted, but does not often change the magnitude of the p-value by a large amount. The simulated null distribution may vary in shape and/or spread, but the process for finding the p-value is always the same, whether using simulation or the mathematical distribution. However, a standardized statistic will also incorporate information about the sample sizes involved. For instance, an R^2 value of 26.4% would be considered more impressive with large sample sizes and less impressive with small sample sizes. Statistical significance considers the amount of natural variation in the response, the sample sizes involved in the study, and the randomness imposed by the study design. Remember that with very large sample sizes, almost any result may be considered "statistically significant," but you should also consider the "practical significance" of your result as well.

Estimating the Size of the Difference

Another advantage of the theory-based approach is that we can also calculate *confidence intervals*.

You may recall: A *confidence interval* estimates the parameter with an indication of the accuracy of that estimate (*margin of error*) and the reliability of our method (*confidence level*). Confidence intervals typically have the form: statistic ± (multiplier)(standard error of statistic).

In this study, the parameter we are trying to estimate is the underlying difference in the treatment means ($\mu_{\text{no scent}} - \mu_{\text{scent}}$).The statistic is the observed difference in group means ($\bar{y}_1 - \bar{y}_2$). Confidence intervals typically have the form: *statistic ± (multiplier)(standard error of statistic)*. Earlier, we saw that we could estimate the standard error of the statistic (SE($\bar{y}_1 - \bar{y}_2$) ≈ $(std\ error\ of\ residuals) \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$). The multiplier will depend on the level of confidence (e.g., roughly 2 for 95% confidence). As we saw above, when the validity conditions are met, we can find the multiplier *t** based on the theoretical *t* distribution.

Definition: A (pooled) ***two-sample t-confidence interval*** for the difference in two population or treatment means, assuming the population standard deviations are equal:

$$(\bar{y}_1 - \bar{y}_2) \pm t^* \times (std\ error\ of\ residuals)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the ***t**** ***critical value*** comes from the *t* distribution with $n_1 + n_2 - 2$ degrees of freedom. For a 95% confidence interval, this multiplier will be close to 2.

Think about it: How would you calculate and interpret a 95% confidence interval for the scent study?

If we use df = 46 and 95% confidence, technology gives us a *t** value of 2.013. So the 95% confidence interval is 1.29 ± 2.103(0.32) ≈ (0.617, 1.963). We are 95% confident that using the scent will increase the average consumer favorability rating between 0.617 and 1.963 points, among a population similar to that of the business majors who participated in the study.

Keep in mind that this interval is about the difference in the underlying treatment *means*. Also notice how we are able to use "action terms" here (e.g., "will increase") because of the random assignment in the study design. So our study conclusions could read something like the following.

STEP 5: Formulate conclusions.

We have significant evidence (*t*-statistic p-value with 46 df ≈ .002) that there is an underlying effect from exposing college business majors to scents while evaluating a store. Among a population similar to that of the business majors who participated in the study, those given a scent will increase the average consumer favorability rating between 0.617 and 1.963 points compared to those not given a scent. The scent model explained 26.3% of the variation in consumer ratings, probably meaningful enough for a store manager to care, as this corresponds to about a 1 point increase on a seven-point scale.

STEP 6: Look back and ahead.

Some further investigations could include a wider range of scents (in this study they only used "inoffensive" scents). Perhaps a pet store or a store near a paper mill might worry about negative impacts on their customers' shopping behaviors. Perhaps this "avoidance" response would be even larger than the effects found in this study. The researchers also noted that the generalizability of the study was limited because it was conducted in a simulated store, and so a follow-up study could also examine a wider variety of store types.

Exploration 1.3: Starry Navigation (cont.)

Recall the Starry Navigation study from Exploration 1.2 (***DungBeetles***). We found that the type of cap used (whether they could see the night sky or not, 9 beetles in each group) explained 84.3% of the variation in times for a beetle to roll the ball to the edge, and the difference in times of 83.77 seconds was large compared to the standard error of the residuals (19.77 seconds). This seems like a large and meaningful difference between the two groups, as also shown by the lack of overlap between the two sample distributions. We seem to have strong evidence that the difference between the two groups is larger than just natural variation in beetle times.
But—is it possible that there really is no treatment effect from the type of cap, and the random assignment process alone was responsible for the large difference between the two treatment groups? In other words, what if the cap didn't make a difference and each beetle's time would be exactly the same no matter which cap they had been using; could we have been so unlucky that the 9 fastest beetles happened to end up in the "no cap" group?
So we have two competing explanations here for the observed difference in the groups:

- There is an effect on dung beetles' rolling speed when they are not able to see the night sky
- There is no difference between whether or not dung beetles can see the night sky, and the only reason we saw a difference between the two groups in our study is "random chance."

You may recall from your first statistics course that the first statement, our research conjecture, is often set up as the *alternative hypothesis* and the second statement is often set up as the *null hypothesis.*

Key Idea: In assessing statistical significance, we typically define null and alternative hypotheses.

- The *null hypothesis* (H_0) is the "by chance alone" explanation for the observed results,
- The *alternative hypothesis* (H_a) typically corresponds to the research conjecture (e.g., the imposed treatments explain variation in the response variable).

You also may recall that a *parameter* summarizes the population or process, but the *statistic* is what we calculate from the observed sample data. Suppose we define our parameter to be $\mu_{\text{black cap}} - \mu_{\text{clear cap}}$ where $\mu_{\text{clear cap}}$ is the mean time to reach the edge for this population of beetles if they can see the sky, and $\mu_{\text{black cap}}$ is the mean time to reach the edge for this population of beetles if the view of the sky is blocked.

1. Restate the null hypothesis and the alternative hypothesis in terms of these μ values.

Notes: If you are looking for evidence of a difference in the average times, you state a *two-sided* (not equal to) alternative hypothesis. If you are looking for evidence that the beetles are faster when they can see the night sky, you specify a *one-sided* alternative. Also note that the null hypothesis says our "single mean" model is adequate, whereas the alternative hypothesis includes our "separate means" model.

2. For the data we provided you for this study, what was the observed value of the statistic corresponding to this parameter?

One way to decide between these two competing explanations (hypotheses), is what we call the ***3S Strategy***.

3S Strategy for Measuring Strength of Evidence:

1. **Statistic:** Compute a statistic from the observed sample data which measures the comparison of interest (e.g., difference in group means).
2. **Simulate:** Identify a "by-chance-alone" explanation for the data (the null hypothesis). Then use a computer to repeatedly simulate values of the statistic, mirroring the randomness of the study design, that could have happened if the chance explanation is true.
3. **Strength of evidence:** If the observed statistic is unlikely to have occurred when the chance explanation is true, then we say we have "strong evidence" against the reasonableness of chance alone as an explanation for the study results.

In other words, we are going to assume the null hypothesis is true and *simulate* thousands outcomes for the study that could happen in that case. We will then be able to determine whether our observed result from the actual study (where we don't know whether the null hypothesis is true) is consistent with these simulated outcomes (where we do know the null hypothesis is true). We will do this my mimicking the randomness that was involved in the study protocol, in this case the random assignment of the beetles to the type of cap. So will we assume that which cap they are assigned to had no impact on their performance, they would have had the same time either way. But the statistic, in this case the difference in the treatment means, could change depending on how the random assignment had turned out.

3. Take enough index cards to represent each beetle. How many index cards do you need?

4. Write each beetle's time on a different card. This represents the beetle times not changing regardless of which treatment group they will be assigned.

5. Shuffle the cards and deal them out in two groups, matching the group sizes of the study.

6. Calculate the mean time for each group and calculate the difference in means (clear cap – black cap).

7. Is the re-randomized difference in means larger or smaller than the original difference in means for these data? Is this what you would expect? Explain why or why not.

8. Does this convince you that it's impossible for random assignment alone to have created the groups that we saw?

We need to repeat this process a large number of times to see what values are possible for the re-randomized differences in means.

Open the **Comparing Groups** applet and paste in the beetle data. Make sure the explanatory and response variables are ordered to match the button above the data and press **Use Data**. Check the box to **Show Groups.** (Because the first category pasted in is "clearcap," the applet reports $\bar{y}_{clear} - \bar{y}_{black}$ as the observed difference.) Check the **Show Shuffle Options** box. Select the **Plot** radio button and press **Shuffle Responses**. The applet mimics what you did with the card shuffling, randomly re-distributing the observed response values back to one of the two groups, 9 in each group.

9. What is the shuffled difference in means after this shuffle?

10. If you press **Shuffle Responses** again, do you get a different value for the shuffled difference in means?

Now change the **Number of Shuffles** to some large number, like 1000, and press **Shuffle Responses** again.

Definition: A ***randomization test*** assumes the null hypothesis to be true and examines all possible re-random assignments of the observed responses among the groups (the *null distribution of the statistic*), recalculating the statistic each time. Instead of doing all possible arrangements, we can repeat the process a large number of times to approximate the ***null distribution*** of the statistic.

11. Describe the shape, center, and variability of the null distribution of shuffled differences in means.

12. Did your shuffles ever produce a difference in means as small or smaller than -83.77? Is it possible we could find a difference in means that negative? Is it very probable?

Remember that this simulation mimics what would happen by random assignment alone ***if*** we assume the treatments have no effect. In other words, if the null hypothesis of no treatment differences is true. We will reject this null hypothesis if the likelihood of the actual study's observed statistic is too small to plausibly occur by chance alone when the null hypothesis is true.

Enter the -83.77 value in the **Count Samples** box and use the **Less Than** pull-down menu option to count how many of your simulated statistics are equal to or smaller than the observed statistic. (If you

specified a two-sided alternative above, then use **Beyond** in the pull-down menu to compute a *two-sided p-value* from both tails of the distribution.)

13. How often does shuffling create a difference in means of -83.77 or smaller?

Key Idea: The *p-value* for a randomized experiment is how often random assignment alone could have produced a statistic at least as extreme as the statistic found in the actual study. A small p-value (e.g., below 0.05) constitutes strong evidence against the null hypothesis of "random chance alone," with even smaller values (e.g., below 0.01 or 0.001) providing even stronger and stronger evidence against the null hypothesis. When the p-value is small, we say that the observed difference is *statistically significant*, meaning the observed value of the statistic is unlikely to have happened by random chance alone.

Note that the values we consider "at least as extreme" as the observed statistic in determining the p-value will depend on the direction of the alternative hypothesis and whether that hypothesis is one-sided or two-sided.

Other Choices of Statistics

Use the **Statistic** pull-down to change from the difference in means to the *R*-squared value.

14. How does the null distribution change (shape, center, variability)?

15. To approximate the p-value for this statistic, we need to use the observed value (as a decimal) in the Count Samples box. What is the observed value of the R^2 statistic for these data? What values do you consider "more extreme" (even strong evidence against the null hypothesis)? Enter the observed value in the Count Samples box and use the pull-down menu to specify the "as extreme as" direction (as a decimal, not a percentage).

16. What is the new p-value? Has it changed much by changing this statistic?

Typically when the R^2 value is large, the p-value will be small. But the p-value also considers the sample sizes involved in the study. If the sample sizes are quite large, then even a modest R^2 value could still be statistically significant. In general, it is good practice to comment on both statistical and practical significance.

Another possible statistic that you may remember from your first course is a "*t*-statistic." The formula below is called a ***pooled t-statistic*** because it assumes the standard deviation of the response outcomes is the same for both treatments and uses one value to estimate that standard deviation. You maybe have also seen the "unpooled" version which does not assume the population standard deviations are the same and so uses a different estimate for the

standard error of the statistic (see Example 1.3). The point is we are now not just dividing by the unexplained variation in the data but are also taking the sample sizes into account. Including the sample sizes in the denominator, which approximates the shuffle to shuffle variation in the statistic, is referred to as ***standardizing*** the statistic.

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{(SE\ of\ residuals) \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

17. Calculate the denominator of this statistic for the dung beetle data. How does it compare to the standard deviation of the null distribution when you use the difference in means as the statistic? (*Hint*: It's not all that close! Why do you think that is?)

18. Calculate the *t*-statistic for these data.

19. If the sample sizes had been 90 and 90 but the difference in means the same, would the *t*-statistic be larger or smaller?

In the applet, use the **Statistic** pull-down menu to select the *t*-statistic.

20. What value is reported for the observed *t*-statistic? How does it compare to your prediction in the previous question?

21. Find the corresponding p-value (explain your steps).

One advantage to the *t*-statistic is it puts our results all on the same scale. We can compare *t*-statistics from different studies directly against each other. Typically *t*-values larger than 2 (or smaller than -2) are considered extreme. So once we see a *t*-statistic below -9, we already know we are going to rule out the random assignment process as a plausible explanation for the differences in mean times between the treatment groups.

Another advantage to the *t*-statistic is it is often well-approximated by a probability model, the ***t-distribution*** ("discovered" by W. S. Gosset and published under the name "Student" in 1908).

Validity Conditions: For the pooled *t*-statistic, when comparing two population means, if
(1) the samples are independent of each other,
(2) the sample standard deviations are roughly equal (e.g., the larger SD is not more than twice the size of the smaller), and

(3) the sample distributions are roughly symmetric or both sample sizes are at least 20 without strong skewness or outliers in the distributions,
then we can approximate the null distribution of the *t*-statistic with a ***t-distribution*** with (*total sample size* – 2) ***degrees of freedom***.

22. Examine the data to see whether this is a case where the *t*-distribution is likely to be a good approximation of the null distribution:

a) Did the study protocol involve random assignment to two treatment groups? If so, then we will consider condition (1) to be met.

b) Is the larger standard deviation less than twice the size of the smaller standard deviation? If so, then we will consider condition (2) to be met.

c) Does either treatment group show severe skewness or extreme outliers? If not, then we will consider condition (3) to be met. [Note: An even better graph to examine here is a distribution of the residuals. If that distribution is approximately normal, we will consider this condition met.]

d) Do you consider all three conditions met for this study?

In the applet, check the box to **Overlay *t* distribution** on the null distribution of *t*-statistics.

23. Does the *t* probability distribution appear to predict the simulation results reasonably well (is it a good approximation of the null distribution)?

24. What degrees of freedom (df) is reported by the applet for this "theory-based" test? Where have you seen this value before?

Key Idea: There is actually a family of *t*-distributions, indexed by a "degrees of freedom" value. (See Figure 1.3.5.) For a pooled *t*-test, this will equal the degree of freedom for the *SSError* calculation, total sample size minus two.

Technical notes: You should find that, visually, the simulation and theory-based *t*-statistic distributions show good agreement. This is because the validity conditions are met, even though our prediction of the standard deviation of the null distribution of difference in means was much too small. This underestimation of the null distribution standard deviation can happen when the treatment effects are large, as found in this study. The theory-based *t*-statistic assumes the data are coming from separate populations with the same mean, but the *within group* variation is estimated by "averaging" the within group variation. This average (after adjusting for the group differences) will be much smaller than when we pool all the observations together in the randomization test. The *t*-statistic corrects for this in a way—when the numerator is large the

denominator will tend to be smaller—and things tend to balance out like a *t*-distribution would predict.

Another huge advantage to the *t*-distribution is we can use it to predict how far the statistic is likely to fall from the parameter we are trying to determine. In other words, we can use the *t*-distribution to calculate confidence intervals.

Estimating the Size of the Difference

In this study, the parameter we are trying to estimate is the underlying difference in the treatment means ($\mu_{\text{black}} - \mu_{\text{clear}}$). The statistic is the observed difference in group means ($\bar{y}_1 - \bar{y}_2$). A confidence interval will start with this estimate, plus and minus a *margin of error*, an indication of the precision of the estimate. Confidence intervals typically have the form: *statistic* $\pm$ *(multiplier)(standard error of statistic)* where the multiplier comes from a probability distribution. When the above validity conditions are met, we will use the *t* distribution to find the multiplier corresponding to our *confidence level*, an indication of the reliability of the procedure.

Definition: A two-sample (pooled) *t*-confidence interval for the difference in two population means, assuming the population standard deviations are equal is

$$\bar{y}_1 - \bar{y}_2 \pm t^* \times (std\ error\ of\ residuals)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the t^* *critical value* comes from the *t* distribution with (*total sample size* – 2) *degrees of freedom*. For a 95% confidence interval, the multiplier will be roughly 2.

In the applet, check the **95% CI(s) for difference in means** box (on the far left) to find the pooled *t*-interval.

25. What is the margin of error (i.e., half-width) of this interval? How does it compare to 2×(9.19)? (For 95% confidence, the multiplier t^* will be roughly 2.)

26. Write a one-sentence interpretation of this interval. (*Hint*: Pay attention to the order of subtract of the group means in the applet.)

STEP 5: Formulate conclusions.

27. Write a summary of your conclusions from this study, including discussion of significance, estimation (from the confidence interval), generalizability, and causation. Does this tell you whether the view of the stars makes a difference in whether a beetle can keep their ball away from others?

STEP 6: Look back and ahead.

28. Suggest at least one way you would improve this study if you were to carry it, or a follow-up study, out yourself.

Section 1.3 Summary

Step 4 of the statistical investigation process is to draw inferences beyond the data. Oftentimes this means stating a null hypothesis (chance explanation) and an alternative hypothesis (typically the research conjecture). The 3S strategy (obtain a Statistic that summarizes the evidence in the data, Simulate values of the statistic when the null hypothesis is true, and evaluate the Strength of evidence by comparing the observed statistic to the simulated values) helps us remember the reasoning behind a test of significance. Asking whether the observed statistic is unlikely to have occurred when the null hypothesis is true, allows us to evaluate the strength of evidence against the null hypothesis. The strength of evidence against the null is typically measured by computing a p-value and seeing how small it is (with p-values less than 0.05 providing strong evidence against the null hypothesis).

In this section, we illustrated two methods for exploring the null distribution of the statistic of interest and finding a p-value: simulation (randomization test) and probability theory (theory-based approach). The theory-based approach should predict the simulation results when certain validity conditions are met. But p-values are only one way to quantify evidence against the null hypothesis. Standardized statistics also do by dividing the statistic of interest (e.g., a difference in two group means) by an estimate of the chance variation in the statistic. A standardized statistic for comparing two group means is called the *t*-statistic. The distribution of the *t*-statistic assuming the null hypothesis is true follows a bell-shaped, symmetric shape, centered at zero, with values larger than 2 or smaller than -2 unlikely to occur by chance alone (corresponding to a p-value of <0.05; strong evidence against the null hypothesis).

Standardized statistics and probability theory also give us methods for estimating the size of the parameter. The two-sample *t*-interval has a common form: confidence interval, *statistic* $\pm$ *(multiplier)* $\times$ *(standard error of the statistic)* and provides a range of plausible values for the unknown difference between two population means or two long-run treatment means. With 95% confidence *t*-intervals, the multiplier is approximately 2. Confidence intervals, along with subject matter knowledge provide another way to assess *practical significance*. Where *statistical significance* helps us decide whether or not we are convinced there is a difference, and confidence intervals estimate the magnitude of that difference.