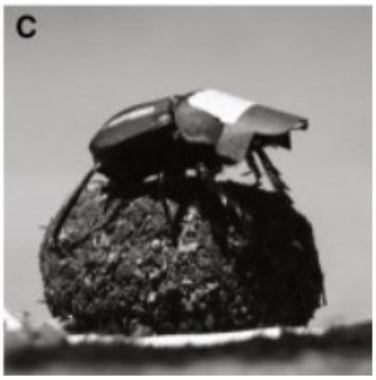
Starry Navigation – Part A

Learning Goals:

- Build simple statistical models to formally capture and summarize important sources of variation in a variable of interest.
- Explain how the variation in the response variable is partitioned into variation explained by the model and unexplained variation.
- Report the percentage of variation that is explained by the model.
- Explain what effect size measures.
- Explain how statistical significance is different from practical significance.

Background: 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 does not 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?



Researchers Dacke, Baird, Byrne, Scholtz, and Warrant ("Dung Beetles Use the Milky Way for Orientation," *Current Biology*, 23, 2013) report on several experiments they conducted 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 (clear) cap. (Why?) On a moonless, starry night beetles wearing the clear 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.

Figure 1: Possible Sources of Variation diagram for Starry Nights study

<table>
<tr><td>Observed Variation in:
Time to reach edge (sec)</td><td rowspan="2">Sources of explained variation
• Type of cap</td><td rowspan="2">Sources of unexplained variation
• Age of beetle
• Gender of beetle
• Unknown</td></tr>
<tr><td>Inclusion criteria
• Beetle species

Design
• Size of platform</td></tr>
</table>

4. Based on the averages provided (124.5 seconds with black cap and 40.1 seconds with clear 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 clear 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

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*).

Go to the **Multiple Variables** applet, use the Select data pull-down menu to load the Dung Beetles data and press **Use Data**. Drag the time variable into the **Response variable** box. Check the **Show descriptive** box.

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:

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

$$= \sqrt{\frac{37441.6}{17}} = \sqrt{\frac{SSTotal}{17}} = 46.93\ seconds$$

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) × (*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. Does the type of cap appear to explain variation in the times? Describe the aspects of the graph/summary statistics you are using to decide.

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.

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. In other words, it is the "pooled" (across the groups) "within group variation." Another phrasing for this is the variation in the rolling times 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*, is often used in statistics 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). (The calculations vary a bit 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).

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

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.

One important consideration in evaluating the treatment effects is whether they can be considered ***practically significant.***

Definition: *Practical significance* refers to whether the treatment effects and group differences are large enough to be of value for that particular context/research area. 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 value larger than one or two? [Often values larger than one or two are considered noteworthy, that the size of the difference between the treatment means is more than the typical variable in the response values...]

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.

Starry Navigation – Part B

Recall that in Part A of the Starry Navigation study about dung beetles navigating during the night 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 to have randomly assigned 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 in rolling speed 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."

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).

Additionally, 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 of 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 by 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 Do / Do Not (circle one) change depending on 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.

Calculate the mean time for each group and calculate the difference in means (clear cap – black cap). Record the two means, and also the difference in means.

6. 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.

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

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

Launch the **Comparing Groups** applet, and use the Select data pull-down menu to load the Dung Beetles data. Toggle the (Response, Explanatory) variable button to match the format of the data and press **Use Data**. Check the box to **Show Groups.** Also, from the **Select statistic** pull-down menu select "Difference in Means." (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.

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

9. Pess **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 998 (for 1,000 total), 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 re-random assignments, we can repeat the process a large number of times to approximate the ***null distribution*** of the statistic.

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

11. Recall that in their study, the researchers observed $\bar{y}_{clear} - \bar{y}_{black}$ to be -83.77. Did your shuffles ever produce a difference in means as small or smaller (more negative) 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 and consider it implausible, if the likelihood of the actual study's observed statistic is too small to plausibly occur by chance (random assignment) 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.)

12. 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.

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

14. 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)? To approximate the p-value for this statistic, enter the observed R^2 statistic value (as a decimal, not a percentage) in the **Count Samples** box and use the pull-down menu to specify the "more extreme" direction.

15. What is the new p-value? Has it changed much by using the R^2 statistic instead of the Difference in means 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. 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. Accounting for the standard error or in this case the shuffle-to-shuffle variation in the statistic is referred to as ***standardizing*** the statistic.

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

16. 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 used the difference in means as the statistic? (*Hint*: It's not all that close! Why do you think that is?)

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

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

18. What value is reported for the observed *t*-statistic?

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

One advantage to the *t*-statistic is it puts results from different studies 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***.

20. Examine the data to see whether in the context of this dung beetles study 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?

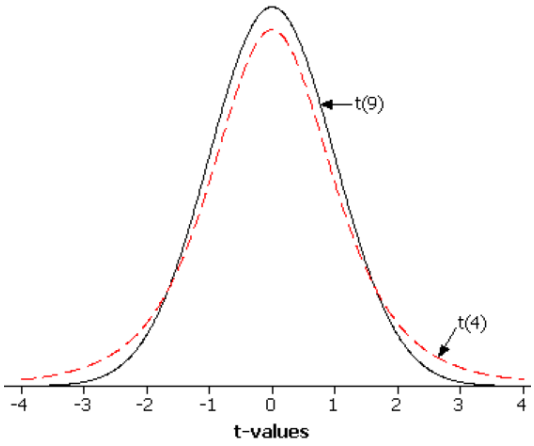
In the applet, check the box to **Overlay *t* distribution** on the null distribution of *t*-statistics.

23. Does the *t* probability distribution appear to predict the simulation results reasonably well (is the *t* probability distribution 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 2.) For a pooled *t*-test, this will equal the degree of freedom for the *SSError* calculation, total sample size minus two.

Figure 2: Example *t*-distributions with 4 and 9 degrees of freedom



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)x(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 $2\times(9.19)$? (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 subtraction 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.