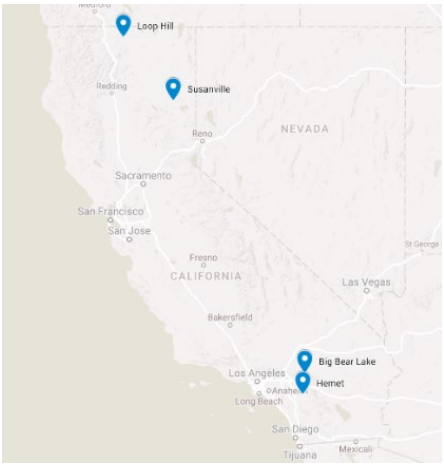
# Golden Squirrels – Part A

*Bergmann's rule* (named after German biologist Carl Bergmann) is an ecogeological rule that states that within a species, specimens will tend to be larger if they are from cooler climates or more extreme latitude. Bergmann's rule is most often applied to warm-blooded animals, but there has been some evidence of the rule in other species as well. A former Cal Poly Biology grad student (Nora Gerdes) wanted to investigate whether Bergmann's Rule applies to the golden mantled squirrel in California. She measured the body lengths (mm) of 18 squirrels from four California locations.




| Location | Avg temperature | Latitude |
|---|---|---|
| Hemet, CA | $64.7^0$ F | 33.7475° N, 116.9720° W (33.7475) |
| Big Bear Lake, CA | $47.6^o$ F | 34.2441° N |
| Susanville, CA | $50.25^0$ F | 40.4167° N |
| Loop Hill, CA (Yreka) | $51.25^o$F | 41.70° N |

**STEP 1: Ask a research question.**

**1.** State the research question along the lines of an alternative hypothesis. Also conjecture other possible sources of variation in squirrel lengths.

**STEP 2: Design a study and collect data.**

**2.** Identify the observational units (how many are there?), the response variable, and the explanatory variable. Classify each variable as quantitative or categorical.

**3.** Was this an experiment or an observational study? How are you deciding?

**4.** Describe any inclusion criteria used in this study.

**5.** Complete a possible Sources of Variation diagram for this study.

| Observed Variation in: | Sources of explained variation | Sources of unexplained variation |
|---|---|---|
| *Inclusion criteria* | | |

**STEP 3: Explore the data.**

Open the **Comparing Groups** applet. Type squirrels.txt in the Data window and press Use Data to preview the data and then press Use Data again to load the data and to produce numerical and graphical summaries of the squirrel lengths.

**6.** Summarize the distribution of squirrel lengths. What is the overall mean length and standard deviation of length? (Be sure to include the units of measurement.) What is the *SSTotal* for these data?

In the applet, check the box for **Show Groups**.

**7.** Does location appear to explain variation in squirrel lengths? How are you deciding? (Include relevant output to support your statements.)

**8.** Use the group means to write out a "separate locations" statistical model. Give the prediction equation and the standard error of the residuals (aka residual standard error).

*Predicted length* =

*SE residuals for separate locations model* =

**Practical Significance**

Recall that in evaluating practical significance you should consider the context of the research study as well as a numerical measure of group differences. When we have only two groups, we can compare the difference in means to the residual standard error: $|\bar{y}_1 - \bar{y}_2|/(residual\ SE)$. If the difference in means is larger than one residual SE, we tend to think of the difference as practically significant.

When there are more than two groups, we can use $R^2$ to summarize how different the groups are in terms of how much variation in the response variable we are able to explain.

**9.** Check the **Show $R^2$** box in the applet. Interpret $R^2$ for these data.

We can also consider whether any pairs of means are further apart than one residual standard error.

**10.** Are any pairs of group means farther apart than one residual standard error?

**11.** Do you consider these differences in the group means to be practically significant? How are you deciding? (*Hint*: Consider your answer to the previous two questions, as well as context here.)

**STEP 4: Draw inferences beyond the data. (Statistical Significance)**

Practical significance is only part of the story. We also want to consider whether the observed differences in mean lengths among the four locations are plausibly due to random chance alone or instead provide strong evidence of a true underlying association between location and length.

**12.** Write out in words the null and alternative hypotheses for a test of significance. (*Hint*: You may do so simply in terms of "association" between the response and explanatory variables or in terms of population means–being sure to define any symbols that you use.)

**Applying the 3S Strategy**

**1. Choice of Statistic**

First, we need one number that summarizes collectively how different the groups are. The difference in means and *t*-statistic only work when we have just two groups to compare.

**13.** Suggest a statistic (a formula) that we could use to summarize the differences among these four groups.

**2. Simulation**

There are several different statistics we could use, the key is getting some sense *overall* of how different the squirrel lengths are among the four locations using *only one number*. This study

was not a randomized experiment, but we can still shuffle the observed lengths to the four locations many, many times and determine how often we would randomly get a value for this statistic as or more extreme than that found in the actual samples. So, after each shuffle, as before, we will need to calculate the value of this statistic, and build a null distribution for the statistic. As we've already computed $R^2$, let's start with $R^2$ as our statistic.

In the **Comparing Groups** applet, check the **Show Shuffle Options** box. With multiple groups, the applet start with $R^2$ as the statistic. Select the **Plot** radio button and press **Shuffle Responses** to get a sense of the randomness being modeled. Then change the **Number of Shuffles** to a large number, like 999, to create a null distribution of $R^2$ statistics.

**14.** Describe the behavior of the null distribution of the $R^2$ statistic. Is it roughly symmetric? Does this shape make sense? Explain.

### 3. Strength of Evidence

Use the applet to estimate the p-value. (*Hint*: What types of $R^2$ values do you consider "more extreme" than the observed value from the actual study?) Include a screen capture of your results.

**15.** Explain how you determined your p-value. Does this p-value provide strong evidence against the null hypothesis of no association between location and length? Explain.

In the applet, select all of the observations in the Sample data window (but not the column headers) and copy to your clipboard. Then scroll to the end of the data and paste in 3 copies of the data, so that you have a total of 4 copies of the data in the Sample data window. Press **Use Data**. Examine the summary statistics.

**16.** How have summary statistics (means and standard deviations and sample sizes) changed? How has the $R^2$ value changed? (Include supporting output to answer these questions.)

**17.** Now reshuffle these response values at least 1,000 times. How does the p-value (our strength of evidence against the null hypothesis) change? Explain why this change makes sense intuitively.

When we have more data, with the same differences in means and the same variability, we might find the additional "consistency" in the group differences more convincing, less likely to be due to "random chance" alone. The simulation reflects this, giving us a smaller p-value, but it would be nice if our statistic did so as well. The $R^2$ statistic was the same whether we had 18 total observations or 72 total observations. Because the $R^2$ doesn't take sample size into

account, we may prefer to look at a *standardized statistic* (something analogous to the *t*-statistic) that reflects both the sample sizes and the left-over or unexplained variation.

### Other Choices of Statistics

The ***F-statistic*** is one such statistic. Named after famous statistician R. A. Fisher, the *F*-statistic compares the explained to unexplained variance, adjusting for the sample size and number of groups, using the degrees of freedom for both the *SSError* and *SSModel*.

**Definition:** The ***F-statistic*** is

$$F = \left[\frac{R^2}{1-R^2}\right] \times \left[\frac{n - \#\ of\ groups}{\#\ of\ groups - 1}\right] = \left[\frac{SSModel/_{df\ for\ Model}}{SSError/_{df\ for\ Error}}\right]$$

where *n* is the total number of observations in the data set.

In the applet, use the Statistic pull-down menu to obtain the *F* statistic for the "four copies" data set.

**18.** Create the null distribution for the *F* statistic and determine the approximate p-value. (*Hint*: What else do you need to change in the applet?)

Now delete the extra copies of the data or reload in the data and examine the *F*-statistic for the *original data set*.

**19.** Verify that the value shown in the applet for the *F* statistic is $\left[\frac{R^2}{1-R^2}\right] \times \left[\frac{n-\#\ of\ groups}{\#\ of\ groups-1}\right]$.

**20.** How does the *F*-statistic for the original data compare to the *F*-statistic for the four copies data set? As you expected? The p-value?

### Theoretical *F*-distribution

One advantage of the *F*-statistic is when certain ***validity conditions*** are met, it is well-approximated by a probability distribution, the ***F distribution*** (also named after R. A. Fisher).

**Validity Conditions:** To use the *F*-distribution to find the p-value for the *F*-statistic requires
(1) the samples are independent of each other,
(2) the standard deviations of the samples are similar (e.g., the largest is not more than twice the size of the smallest), and

(3) the distributions of the samples are approximately symmetric (implying the distribution of the residuals is approximately normal) or all group sizes are larger than 20 with no extreme skewness or outliers.

Notice these validity conditions are the same conditions we used for the (pooled) *t*-test.

Consider the original data set with 18 observations.

**21.** Do you consider condition (1) to be met for this study? Explain.

**22.** Do you consider condition (2) to be met for this study? Explain.

**23.** Do you consider condition (3) to be met for this study? (For now, examine the group dotplots and/or consider the group sizes.) Explain.

Under the null distribution that you created for the original data, check the box to **overlay *F* distribution** on your simulation results. Include a screen capture of your null distribution.

**24.** Does the theoretical *F*-distribution do a good job of approximating the shuffled null distribution even in the original study with such small sample sizes?

**25.** Based on your p-value (the simulated and theory-based p-values should be similar), what conclusions will you draw regarding the null hypothesis?

**STEP 5: Formulate conclusions.**

**26.** Based on your analysis so far, summarize the conclusions you would draw from this study. Be sure to address statistical significance, generalizability, and causation. Also be sure to put your comments into the context of this research study and Bergmann's rule.

**STEP 6: Look back and ahead.**

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

**More on Analysis of Variance**

The following illustrates some additional calculation details for the *F*-statistic.

**28.** What are the *SSModel* (or *SSlocation*) and *SSError* for this separate locations model? Arrange the information you have so far in the following table. (*Hint*: The degrees of freedom for Location will be the sum of (group size – 1) for each location.)

| Source of variation | df | SS |
|---|---|---|
| **Location** | | |
| **Error** | | |
| **Total** | | |

**29.** In the **Comparing Groups** applet, check the box for **Show ANOVA table**. Notice that this table keeps track of the sources of variation in squirrel lengths, degrees of freedom, sums of squares, and more.

a) Why are the degrees of freedom for Total 17?

b) Verify that *SSModel* is the weighted (by sample size) sum of the squared treatment effects (group mean – overall mean).

c) Verify that the "mean square" (*MS*) values equal the sum of squares values divided by the corresponding degrees of freedom.

d) Verify that the square root of the *MSError* is the standard error of the separate means model residuals.

**30.** Verify that the *F*-statistic is the ratio of *MStreatment* (the variance of the group means) and *MSError* (the unexplained variance of the squirrel lengths)

The above calculations show that the *F*-statistic can also be viewed as a ratio of variances, the *between group* variance in the group means, and the *within group* unexplained variance of the residuals. For this reason, the table keeping track of the degrees of freedom and sums of squares is often called an Analysis of Variances or ANOVA table. This theory-based approach is also often referred to as an *F*-test.

# Golden Squirrels – Part B

Recall that the body lengths of 18 golden mantled squirrels were measured from four locations in California. The locations were chosen so that the locations varied in average yearly temperature. Bergmann's Rule states that the members of a species are larger when they are from cooler climates (i.e., more extreme latitudes).

Our hypotheses of interest can be written using either of the formats shown below.

Option 1 – Hypotheses stated in terms of association

$H_o$: There is no underlying association between squirrel length and location in this population
$H_a$: There is an underlying association between squirrel length and location in this population

Option 2 – Hypotheses stated in terms of population means

$H_o$: $\mu_{\text{Hemet}} = \mu_{\text{Big Bear}} = \mu_{\text{Susanville}} = \mu_{\text{LoopHill}}$ (the single mean model is sufficient)
$H_a$: At least one $\mu$ differs from the others

where $\mu_{Hemet}$ is the mean length in the population of all golden mantled squirrels from Hemet; similarly for Big Bear, Susanville, and Loop Hill. Recall that the separate locations model explains about 60% of the observed variation in the lengths of these 18 squirrels ($R^2 = 0.60$) and the *F*-statistic was 7.059, indicating the between-location variation is about 7 times more than the unexplained variation left-over within the groups (i.e., after accounting for location). Using the *F*-distribution with 3 and 14 degrees of freedom gives the theory-based p-value of 0.004. This small p-value gives us strong evidence of a true association between the length of golden mantled squirrels and where they live. (In general, *F*-statistics larger than about 4 usually correspond to small p-values.)

But, have we really answered the research question? Not yet! So far, we've only found evidence of an association between length and location. To determine whether Bergmann's Rule applies, we need to understand the nature of the association between length and location. We need to address questions such as:

- Do squirrels from colder locations tend to be longer, on average?
- Which population mean or means is/are different than the others? How much do they differ?
- Does the average length differ in every one of these four populations? Maybe just one of them?

In other words, we need to understand how the population mean lengths of the different locations *compare to each other*.

### Post-hoc Analyses

Once we find a significant association, the natural follow-up question is the nature of that association. The process of following up a statistically significant *F*-test is called ***post-hoc-analysis***.

**Definition:** The process of assessing how the means of the treatment groups relate to one another in a follow-up analysis to a significant *F*-test is called a ***post-hoc analysis***.

We will now see how to conduct a post-hoc analysis to evaluate whether Bergmann's Rule applies.

**Pairwise Comparisons of the Treatment Groups**

Arguably the most common type of post-hoc analysis involves comparing each group mean to each other group mean by conducting ***pairwise comparisons***.

**Definition:** ***Pairwise comparisons*** are used to compare each group to every other group. Often pairwise comparisons take place as part of a post-hoc analysis.

**1.** Why do you think it is considered OK to conduct pairwise comparisons as part of a post-hoc (*follow-up*) analysis to a *significant F*-test, but not *before* the *F*-test?

You may recall that the reason to conduct an overall test of significance when testing multiple groups was to control the Type I error rate. A key idea is to try to control the experiment-wise Type I error rate.

**Definition:** The ***experiment-wise Type I error rate*** is the chance of making at least one Type I error when conducting numerous tests of statistical significance.

To better control this rate, we will only conduct post-hoc analyses after a significant *F*-test.

**Key Idea:** We can protect against an inflated experiment-wise Type I error rate by only conducting post-hoc analyses using pairwise comparisons after obtaining a statistically significant *F*-statistic

Once again, enter the ***squirrels*** data into the **Comparing Groups** applet. Select **Show Groups**, then check the box for **95% CI(s) for difference in means**.

**2.** Do any of these confidence intervals (CI) contain 0? If so, which one(s).

**3.** What does it tell you about the population means when the CI for the difference in means contains 0?

**4.** What does it tell you about the population means if the 95% CI has two negative endpoints?

The results of these pairwise comparisons can be summarized in a ***letters "plot"*** or ***letters table***.

**Definition:** A ***letters "plot" or letters table*** is a table which indicates which groups are and are not statistically significantly different than each other when conducting pairwise comparisons.

**5.** Fill in the table below to make a letters plot of the means. When two groups have the same letter, it indicates that the group means are *not* statistically significantly different. For example, if Susanville and Big Bear are not significantly different they would be assigned the same letter (e.g., "b"). Typically, letter plots use the letters, a, b, c, d, ….

| Location (Avg Temp) | Mean Length (in mm) | Letters (Groups with the same letter are not significantly different) |
|---|---|---|
| Loophill (51.25 °F) | 280.75 | |
| Susanville (50.25 °F) | 262.20 | |
| Big Bear (47.6 °F) | 260.75 | |
| Hemet (64.7 °F) | 252.0 | |

**6.** Write a brief summary of your findings from the pairwise confidence intervals. Be sure to address whether/how these intervals support the validity of Bergmann's Rule to the golden mantled squirrel in California.
Summary:

**Confidence Intervals on Other Parameters**

The confidence intervals for the difference in population means allows us to compare the mean lengths of squirrels from two locations. But, what if we also wanted to estimate the average length of the population of squirrels from say Big Bear Lake?

**Definition**: A ***t*-interval for a population mean** is:

$$\bar{y}_i \pm (t^*_{df}) \frac{residual\ SE}{\sqrt{n_i}}$$

where, $n_i$ is the sample size of the treatment group, and the *t** multiplier is approximately 2 for 95% confidence intervals.

**7.** Earlier (in part A) when we did the *F*-test and when we computed the pairwise intervals above, we assumed that the standard deviations were approximately equal within the groups. The best estimate of this value is called the pooled estimate of the standard deviation, comparing the group standard deviations together. This calculation exactly corresponds to our residual SE. Using the Comparing Groups applet, what is the value of

the pooled SD for these data? Intuitively, explain why its value makes sense given the values of the SDs for the four different groups.

**8.** Use the group means, the residual SE (previous question), the sample size of each group, and the *t*-multiplier to find determine 95% confidence intervals for the mean length of all squirrels within each location. Write an interpretation of one of these intervals in the context of this study. *Important note: Just use a t-multiplier of 2 for each interval to yield an approximate interval. A more precise interval could be obtained by finding a different value of t depending on the error df.*

| **Location (Avg Temp)** | **Confidence interval for $\mu_j$** |
|---|---|
| Loophill (51.25 $^{o}$F) | |
| Susanville (50.25 $^{o}$F) | |
| Big Bear (47.6 $^{o}$F) | |
| Hemet (64.7 $^{o}$F) | |

**9.** How would increasing the sample size of the location groups change these intervals?

**10.** How would increasing the confidence level to 99% change these intervals?

**11.** Will any of the confidence intervals computed above allow you to predict the length of an individual new squirrel at a particular location? Why or why not?

**Prediction Intervals**

Up until now, our focus has been on confidence intervals for population means (or differences in means). These intervals provide estimates of ranges of plausible values for the unknown population mean value. These confidence intervals for the mean do not allow us to make predictions about the lengths of individual squirrels (e.g., how long would we estimate a squirrel to be if we randomly sampled one more squirrel from a particular location).

**Definition:** A ***prediction interval*** gives an interval of values within which we predict the response of a new individual observation (e.g., person) to occur with some degree of confidence. For example, a 95% prediction interval means we are 95% confident that the responses for 95% of individuals in the population will be captured in the interval.

**Definition:** A ***t*-prediction interval for a new individual** from the population is

$$\bar{y}_i \pm t^*_{df} \times (SE\ of\ residuals) \times \sqrt{1 + \frac{1}{n_i}}$$

where $t^*$ is roughly 2 for a 95% prediction interval.

**12.** Use the formula above to compute an approximate 95% *t*-prediction interval for a new squirrel at each location. Once again use the pooled standard deviation (residual standard error) and a *t** value of 2 in your computation.

| Location (Avg Temp) | Prediction interval for one squirrel length |
|---|---|
| Loophill (51.25 ºF) | |
| Susanville (50.25 ºF) | |
| Big Bear (47.6 ºF) | |
| Hemet (64.7 ºF) | |

**13.** Write an interpretation of one of the prediction intervals in the previous question. Comment on how your interpretation of this interval differs from the interpretation of the confidence intervals in #8.

**14.** Will increasing the sample size within the location groups have a large impact on the width of these prediction intervals? Explain why or why not.

**15.** In general, for a group of interest, which is wider, a 95% prediction interval or a 95% confidence interval? Explain.

As we've seen before, these theory-based intervals have certain conditions that must be met in order to be valid.

**Validity condition for confidence intervals and prediction intervals**

- The data distributions should be reasonably bell-shaped and symmetric, especially if the sample sizes are small. This condition is particularly important for prediction intervals. (With confidence intervals, the distribution of the sample mean should become more normal when the sample size increases, but the distribution of the responses themselves does not change shape as we increase sample size.)
- For comparing pairs of population means simultaneously, the standard deviations should be approximately equal among all the populations. This is because these confidence intervals all use the same residual standard error from the separate means model.

**16.** Based on the dotplots of the squirrel data, do you think that the validity conditions are met? Why?