Explorations 9.1 and 9.2: Exercise and Brain Volume

Comparing Multiple Means: Simulation- and Theory-Based Methods

LEARNING GOALS

- Understand how multiple comparisons can increase the probability of a Type I error.
- Apply the Mean Group Diff statistic to a dataset, including the relationship with the statistic comparing two means.
- Understand that larger values of the Mean Group Diff statistic suggest stronger evidence against the null hypothesis.
- Use the 3S strategy with the Mean Group Diff statistic.
- Use the Multiple Means applet to carry out an analysis using the Mean Group Diff statistic to compare multiple means.
- Understand why the simulated null distribution of the Mean Group Diff statistic looks different from other simulated null distributions.
- Conduct a follow-up test after using the Mean Group Diff statistic.
- Find the value of the ANOVA *F*-statistic using the Multiple Means applet, recognize that larger values of the statistic indicate more evidence against the null hypothesis, and explain why the distribution of the *F*-statistic is skewed right.
- Identify whether an ANOVA (*F*) test meets appropriate validity conditions.
- Conduct an ANOVA using the Multiple Means applet, including appropriate follow-up tests.

STEP 1: State the research question.

Brain size typically shrinks as people age past adulthood, and such shrinkage may be linked to dementia. Therefore, any intervention that can protect against brain shrinkage could help to protect the elderly against dementia and Alzheimer's disease. Researchers in China recently investigated whether different kinds of exercise/activity might help to prevent brain shrinkage or perhaps even lead to an increase in brain volume (Mortimer et al., 2012).

STEP 2: Design a study and collect data.

The researchers randomly assigned elderly adult volunteers into four activity groups: tai chi, walking, social interaction, and no intervention. Except for the group with no intervention, each group met for about an hour three times a week for 40 weeks to participate in their assigned activity. The tai chi group was led by a tai chi master and an assistant, the walking group walked around a track, the social interaction group met at a community center and discussed topics that interested them, and the no-intervention group just received four phone calls during the study period. A total of 120 participants started the study, and 13 dropped out along the way, so 107 completed the study.

Each participant had an MRI to determine brain volume before the study began and again at its end. The researchers measured the percentage change in brain volume in each participant's brain during that time. If a person's brain volume increased, then this percentage change was positive; if brain volume decreased, then this percentage change was negative. The researchers thought that physical activity would help increase brain volume; hence they anticipated that the tai chi and walking groups would tend to show larger increases in brain volume during the study than the control group and the social interaction group.

To start the analysis, we will first consider whether there are any differences in the brain volume changes across these four groups. The null hypothesis can be written in terms of no association between the explanatory and response variables, whereas the alternative asserts that there is an association between the explanatory and response variables. More specifically, the null hypothesis asserts that all four long-run means are equal, and the alternative hypothesis states that at least one long-run mean differs from the others.

1. State the null and the alternative hypotheses, in symbols and in words, in the context of this study.

STEP 3: Explore the data.

These data can be found in the file **Brain**.

2. Paste the data into the **Multiple Means** applet and press **Use Data**. The applet will produce dotplots and descriptive statistics. Check the **Add boxplots** box to overlay those as well.
   a. Which activity group tended to have the largest increase in brain volume percentage change? Which tended to have the smallest increase (i.e., largest decrease)?

   b. Describe what the graphs and statistics reveal about whether the four activities appear to differ with regard to percentage change in brain volume for the participants in this study.

STEP 4: Draw inferences. (Using 3S strategy)

1. Statistic: To minimize the chance of making a Type I error (rejecting a true null hypothesis), rather than carrying out six different tests comparing each group to each other, we want a single statistic to help test the overall null hypothesis of no association. We will start with the mean of the absolute values of the pair-wise differences in means (called ***Mean Group Diff***). In this exploration there are four groups and six ways to pair them up. So our Mean Group Diff statistic can be calculated using the following:

$$\text{Mean Group Diff} = \frac{|Avg1-Avg2|+|Avg1-Avg3|+|Avg1-Avg4|+|Avg2-Avg3|+|Avg2-Avg4|+|Avg3-Avg4|}{6}$$

3. Explain why it's important to take the absolute values before averaging the six differences.

4. What is the value of the Mean Group Diff statistic shown in the applet?

2. Simulate

5. Suppose that you were to conduct a tactile simulation to generate many possible values of the Mean Group Diff statistic that could have happened if the null hypothesis of no association were true, that is, by random chance alone. For example, you could use index cards to represent each person in the study. Answer the following questions about understanding how to carry out the

tactile simulation:

a. How many cards will you need?

b. What would you write on the cards?

c. When you shuffle and deal, how many piles will you make?

d. How many cards will you place in each pile?

e. What should you be recording after the completion of each shuffle and deal?

f. What else do you need to do to find the p-value?

Instead of doing the tactile simulation (which would be fun, of course, but time consuming) let us use the **Multiple Means** applet to run the simulation. Check the **Show Shuffle Options** box in the applet. Enter 1 for **Number of Shuffles** and press **Shuffle Responses**. Notice in the **Most Recent Shuffle** data window, the **Treatment** column has not changed, but the **BrainChange** column has mixed up and reordered all its values. If you select the **Plot** radio button, you will see the new dotplots.

6. Notice that the new dotplots and statistics are produced for the shuffled data. Press **Shuffle Responses** four more times (with the Plot button selected) to visualize the re-randomization of the observations to the four groups. Are any of these new shuffled Mean Group Diff values more extreme than the observed value of the Mean Group Diff statistic from the observed data?

7. Now enter 995 for the **Number of Shuffles**, producing a total of 1000 shuffles.
   a. Describe the shape of the null distribution of the Mean Group Diff statistic. Is it symmetric or skewed? If it is skewed, in which direction is it skewed?

   b. Is the null distribution centered at zero? Explain why your answer makes sense.

3. Strength of evidence

8. Now you will calculate a p-value in order to assess the strength of evidence that the experimental data provide against the null hypothesis that type of activity has no association with change in brain volume.
   a. To calculate the p-value, you will count how many of the simulated Mean Group Diff statistics are equal to _____ or (larger or smaller or beyond).

   b. Enter the observed value of the Mean Group Diff statistic and use the pull-down menu to select the appropriate direction. Report the proportion of samples that are at least as extreme as the observed statistic. This estimates the p-value.

c. You should find that the p-value is fairly close to a common cutoff value for assessing statistical significance. So, go ahead and produce 9000 more shuffles to produce a more accurate estimate of the p-value (there will be a pause). Report this p-value.

d. Interpret this p-value: It is the probability of obtaining a Mean Group Diff statistic of _____ or _____, assuming that _____.

9. **Significance**
   a. Based on this p-value, evaluate the strength of evidence: The experimental data provide _____ evidence against the null hypothesis that the type of activity has no association on percentage change in brain volume.

   b. Does this analysis allow you to determine *which* activities differ significantly from which others with regard to percentage change in brain volume? Explain why or why not.

Although the Mean Group Diff statistic is fairly easy to understand and calculate, it is not commonly used, in part because there is no theory-based model for the null distribution. Another downside is that the Mean Group Diff statistic is not standardized so it is not comparable across studies (e.g., a Mean Group Diff statistic of 1 in one study might be strong evidence, but in another study might not show convincing evidence).

Key Idea
Instead of comparing only the differences in means, a statistic could standardize those differences by comparing to the amount of within-group or natural variability in the response variable while also taking sample sizes into account.

A much more commonly used statistic for comparing multiple groups on a quantitative response, which is standardized and does have a theory-based distribution, is called an ***F*-statistic**. As with the Mean Group Diff statistic, the *F*-statistic equals zero only when the group means are all identical. Otherwise the *F*-statistic is positive, with larger values indicating larger differences across the group means.

The *F*-statistic is a ratio of "between-group" and "within-group" variability. Thus,

$$F = \frac{between\ group\ variability}{within\ group\ variability}$$

The numerator is a measure of how much the group means differ from each other, and the denominator is a measure of how much variation there is within the groups (related to the SDs within the groups).

10. Select the **Statistic** pull-down menu to select ***F*-statistic**, and the null distribution changes from the display of the simulated Mean Group Diff statistics to the *F*-statistics for each random re-assignment.
    a. Report the observed value of the *F*-statistic, which appears on the left side of the applet.

    b. Use the simulation results and the observed value of the *F*-statistic to estimate the p-value.

c. Is this p-value similar to the one based on the Mean Group Diff statistic? Is your conclusion about strength of evidence the same with both statistics?

11. The *F*-statistic has a known theoretical distribution when the null hypothesis is true, called (surprisingly enough) an ***F-distribution***. We can use that to find a theory-based p-value. Check the box to **Overlay *F* distribution**. Would you say there is good agreement between this theoretical prediction and your simulated null distribution? How does the theory-based p-value compare to the simulation-based p-value using the *F*-statistic?

Digging Deeper: R^2 and the *F*-Statistic

Let's dig in a little deeper to both understand what the *F*-statistic means as well as another statistic we can calculate from the ANOVA (ANalysis Of VAriance) Table. To get this started, check the **Show ANOVA Table** box on the left to display the ANOVA table. Notice that one column is labeled SS. The SS stands for sum of squares and is simply a measure of variability found by squaring (and then adding up) the difference between numbers and their mean.

For example, the sum of squares total, 130.39, is found by adding up all the squared differences between each brain change number and the overall mean of 0.138. The sum of squares error, 119.56, does the same thing except the differences are between each brain change number and its group mean. This number then represents variability that is left over (or not explained) after accounting for the activity group.

The sum of square error (variability not explained by the treatment) divided by the sum of square total (total variability) is the proportion of total variability in the brain change numbers that is **not** explained by the treatment. This means that one minus this number is the proportion of total variability that **is** explained by the treatment. We call this number R^2.

Definition
The proportion of the variation in the response variable which is explained (accounted for) by the treatments (groups) is called $\boldsymbol{R^2}$ = 1 - $\frac{Sum\ of\ Squares\ Error}{Sum\ of\ Squares\ Total}$. This number is often reported as a percentage (e.g., 25%) instead of a proportion (e.g., 0.25)

12. Using the ANOVA table, find the value of R^2 for this study and describe what it means.

13. What percentage of the variability in brain volume change is not explained by the activity group?

The sum of squares treatment (10.83 in the ANOVA table) is a measure of variability between the group means and when added to the sum of squares error equals the sum of squares total. So because the sum of squares error is the variability not explained by the treatment, the sum of squares treatment is the variability that is explained by the treatment. This means that R^2 can also be calculated by:

$$R^2 = \frac{Sum\ of\ Squares\ Treatment}{Sum\ of\ Squares\ Total}.$$

The values for the sum of squares numbers are not only dependent on the variability of the data (or variability of the means) but are also dependent on the sample sizes and the number or groups. Therefore, R^2 is not a standardized statistic. To create the *F*-statistic, which is standardized, we need to take the sample size and number of groups into account. This is done through the degrees of freedom numbers that are also listed in the table. The degrees of freedom for the treatment is the number of groups – 1 and the degrees of freedom for the error is total sample size – number of groups. These sum of squares are divided by their corresponding degrees of freedom to determine the numbers in the MS column (which stands for mean square). Finally, the *F*-statistic is the ratio of the mean square treatment (a number that describes the variability between the groups) and the mean square error (a number that describes the variability within the groups).

14. Confirm that the *F*-statistic is the ratio of the ***mean square for treatment*** to the ***mean square for error***.

Follow-up analysis. The ANOVA *F*-test is an "overall" test. A significant test result tells us that we have strong evidence that at least one population mean is different from the others but does not tell us which one(s) differ from which other one(s). To explore these differences, we need to apply a follow-up procedure to ANOVA. There are many different follow-up tests that can be done. We will use pairwise confidence intervals for the difference in two means.

Key Idea

If an ANOVA test finds that there is at least one long-run or population mean that is different, there are many different follow-up tests designed to help pinpoint where difference(s) occur. One option is to look at all the pairwise theory-based confidence intervals for the difference in long-run or population means.

15. When you have the *F*-statistic selected and the *F*-distribution overlaid (or are showing the ANOVA table) you can check the box to **Compute 95% CI(s) for difference in means**. Which of the confidence intervals calculated indicate a significant difference between groups? How are you deciding?

16. Based on your analysis, would you conclude that there is one activity that works significantly better than the other three? If so, which one?

Validity conditions

As with other theory-based approaches of conducting tests of significance, certain validity conditions must be met in order to conduct an ANOVA *F*-test.

Validity Conditions for the ANOVA *F*-Test

The *F*-distribution is a good approximation to the null distribution of the *F*-statistic as long as:

Either the sample size is at least 20 for all the groups without strong skewness or outliers in the response variable, or if the sample sizes are less than 20, then the distribution of the response variable is approximately symmetric in all the samples (examine the dotplots for skewness or outliers).

The standard deviations of the samples are approximately equal to each other (Largest standard deviation is not more than twice the value of the smallest standard deviation.)

17. Do the sample data for the brain change study appear to satisfy the validity conditions for conducting an ANOVA *F*-test? How are you deciding?

STEP 5: Formulate conclusions.

18. Are you comfortable with concluding from this study that the activity type causes a difference in brain volume change? Justify your answer.

19. To what population are you comfortable generalizing the results of this study? Justify your answer.

STEP 6: Look back and ahead.

20. **Looking back:** Did anything about the design and conclusions of this study concern you? Issues you may want to critique include:
    - The match between the research question and the study design
    - How the experimental units were selected
    - How the treatments were assigned to the experimental units
    - How the measurements were recorded
    - The number of experimental units in the study
    - Whether what we observed is of practical value

21. **Looking ahead**: What should the researchers' next steps be to fix the limitations or build on this knowledge?