Exploration 8.1 – 8.2: Steps per Day and All-Cause Mortality

Comparing Multiple Proportions

LEARNING GOALS

- Understand how multiple comparisons can increase the probability of a Type I error.
- Compute the Mean Group Diff statistic from a data set when comparing multiple proportions.
- 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 Proportions applet to carry out an analysis using the Mean Group Diff statistic to compare multiple proportions.
- Explain why the simulated null distribution of the Mean Group Diff statistic looks different from other simulated null distributions.
- Conduct a follow up analysis after using the Mean Group Diff statistic.
- Find the value of the chi-square test statistic using the Multiple Proportions applet, recognize that larger values of the statistic mean more evidence against the null hypothesis and why the distribution of the chi-square statistic is non-negative and follows a right-skewed distribution.
- Conduct a chi-square test of significance using the Multiple Proportions applet, including appropriate follow-up tests.
- Identify whether a chi-square test meets appropriate validity conditions.

Issues in Multiple Testing

You might think that to compare multiple proportions (or multiple means), one simply conducts all the possible two-sample comparisons between the groups using ”two-sample” methods. For example, with four groups, if our groups are labeled A, B, C, and D, we could test all six comparisons: A vs. B, A vs. C, A vs. D, B vs. C, B vs. D, and C vs. D. Suppose each of these tests is tested at the 5% significance level (i.e., rejecting the null hypothesis if the p-value is smaller than 0.05). Now suppose the null hypotheses in these tests are all true (all four of the group parameters are actually the same). This means that for each test there is a 5% chance that we will make a mistake and reject the null hypothesis even though each statistic actually did happen by chance alone. This type of mistake is defined as a ***Type I error***, or false alarm. We control the probability of making a Type I error through the level of significance. The problem is, these Type I errors “accumulate” when we do more and more tests on the same data. (For an analogy, suppose that every time you ski down a certain run, there’s a 5% chance you will fall. If you ski down the run 15 times, what’s the chance you will fall at least once?) If we conduct six tests each at the 5% significance level, our overall Type I error rate (the probability of rejecting at least one of the six null hypotheses that are all actually true) jumps to more than 26%.

An alternative approach to testing all possible pairs against each other is to use one test that compares all four parameters at once. If we fail to reject the null hypothesis, say at the 5% level, we are “done,” in that we will conclude that we don’t have evidence that any of the long-run parameters differ. By constructing a statistic that compares all sample proportions or sample means at once, we can perform just one test and thus we can keep the probability of a Type I error as small as we want. But, to do this, we need to find a single statistic that measures the differences in all our proportions (or means) at once so we can run one overall test. This is the approach we explore in this exploration. We’ll also see the

These materials were developed by the STUB Network and supported by the National Science Foundation under Grant NSF-DBI 1730668.

commonly used theory-based approach for comparing multiple proportions: the chi-square test. Finally, when evidence exists to reject this null hypothesis, we'll also explore a follow-up analysis that allows us to begin to construct a more detailed picture of the nature of the relationship.

STEP 1: State the research question.

Seems everyone is counting their daily steps. Regular physical activity is important in maintaining and improving physical health. The step volume for an individual is one way to measure total daily activity. Researchers conducted a prospective study in middle-aged Black and White adults to see whether the amount of steps taken was associated with mortality.

STEP 2: Design a study and collect data. The baseline ages of the 2110 participants ranged from 18-30 years and the subjects in the study were followed for 11 years. Along with age and race, sex, BMI, years of education, and smoking status were also recorded. The main focus of the study, however, looked at the association between step volume and mortality. Researchers measured step volume as the sum of the raw step counts as reported by an accelerometer for each valid day, then calculating the mean across days. Participants were grouped into three groups according to their step volume: low: <7,000 steps/day, moderate: 7,000 to <10,000 steps/day, high: ≥ 10,000 steps/day.

1. What are the variables? Are the variables categorical or quantitative? If quantitative, what are the measurement units. If categorical, how many categories do they have? Also, identify the roles (explanatory or response) of the variables.

   Explanatory variable: Type:
   Response variable: Type:

2. Let's write out the hypotheses.
   a. Write the appropriate null and alternative hypotheses (in words) using the language of association between the explanatory and response variables.

   b. Express the null and the alternative hypothesis in words but now in terms of population proportions instead of the language of association.

   c. Finally, write the null and alternative hypotheses using symbols π_{low}, $\pi_{moderate}$, and π_{high} (define at least one of these symbols in context)

STEP 3: Explore the data.

3. The researchers found that 32 of 448 subjects in the low step volume group died during the 11-year study, 16 of 863 in the moderate step volume group died, and 24 of the 799 in the high step volume group died. Enter the data in the Multiple Proportions applet, click the check box by **Enter table** and select 2x3. Enter the labels low, moderate, and high for the three groups and fill in the six squares in the body of the table with success being that the participant died and failure being that the patient survived (note that is different than the total in the group). Though it may feel strange to consider "death" a "success," oftentimes the more rare outcome is the one we focus on for "success." Then click on **Use Table**. Comment on what the segmented bar graph or mosaic plot, and the conditional proportions reveal about whether the step volume appears to be associated with whether or not died during the study (Check the **Show table** box to see the two-way table and conditional proportions.)

STEP 4: Draw inferences beyond the data.

We see some differences in the sample proportions of individuals who have died across the three step volume groups, but are these differences large enough to be statistically significant? In other words, do these data provide strong evidence of an association between the step volume and mortality in the population?

Applying the 3S Strategy

To investigate, we will see how we can apply the 3S strategy to these data. There are actually several reasonable ways of summarizing the differences among the groups in one number. Once you have settled on a statistic, you apply the 3S strategy as before—simulate the distribution of the statistic under the null hypothesis and then see where the observed value of the statistic falls in that simulated null distribution.

1. Statistic: A reasonable statistic to calculate is the mean of the absolute values of differences for each pair of groups. We will call this statistic ***Mean Group Diff*** for mean of the group differences.

4. Let's construct the Mean Group Diff statistic for these data by going through the following steps.
   a. Using the conditional proportions of mortality shown in the applet calculate the differences in these proportions for each pair:

      Low minus moderate step volume:
      Low minus high step volume:
      Moderate minus high step volume:

   b. Calculate the mean of the absolute values of these differences. (In the **Multiple Proportions** applet, change the statistic to the Mean Group Diff and verify that your calculation matches.)

2. Simulation: We have seen how we can simulate a null hypothesis of no association by shuffling the response variable outcomes across the explanatory variable groups. This models the random assignment process used in experimental studies that assign treatments as part of the data collection process, and, assuming the null hypothesis is true, "breaks any potential association" between the response and explanatory variables. This re-randomizing process works in studies where treatments weren't randomly assigned as part of the study design as this still breaks any potential association between the explanatory and response variables. We could model this with cards representing the observational units and words on the cards representing the response variable.

5. Describe how you would model a simulation of the null hypothesis with cards now that there are three explanatory variable groups instead of two. Make sure to specify what you would write on these cards and how many of each type of card there would be. How many cards would you deal out to each group? What would you calculate after dealing them out?

6. Now let's use the **Multiple Proportions** applet to simulate a null distribution of these Mean Group Diff statistics.

a. Check **Show Shuffle Options** and leave **Number of Shuffles** at 1. Press **Shuffle Response** to perform one shuffle of the response variable values. You should see the shuffled response variable, the new two-way table for the simulated data, and the value of the simulated statistic (Most Recent Shuffled Mean Group Diff) which is also placed in blue on the graph on the right. Select the **Plot** radio button to see the shuffled segmented bar graph or mosaic plot. How does the distribution across the groups for the shuffled data compare to the original data? Is the simulated Mean Group Diff statistic value closer to zero than the observed value of the Mean Group Diff statistic?

b. Now enter 999 for the **Number of Shuffles** and press **Shuffle Responses**. This repeats the shuffling of the response variable 999 more times for a total of 1,000 repetitions. You should see a graph of a null distribution of the Mean Group Diff statistics. What is the shape of this distribution? Why is it *not* centered at zero?

3. Strength of evidence

7. Estimate the p-value by determining how often the observed value of the statistic, or something even larger, occurred in the null distribution. (*Hint*: Enter the observed value for the Mean Group Diff statistic from the research study in the **Count Samples** box and press **Count**.)

8. Based on the p-value, summarize the strength of evidence that the sample data provide against the null hypothesis.

Another choice of statistic: chi-square

The Mean Group Diff statistic is fairly easy to understand but is not widely used in part because it cannot be easily predicted theoretically. A more commonly used statistic is called a chi-square (χ^2) statistic. In this exploration, we will use the more general formula that works with data that has a categorical response with any number of categories. Using this more general formula, the chi-square statistic is obtained by squaring the differences in the observed and expected counts, dividing by the expected count (a form of standardizing), and then summing the six values. This can be thought of as

$$\chi^2 = \sum \frac{(Observed\ count - Expected\ count)^2}{Expected\ count}$$

where Σ asks you to sum across all the cells in the 2x3 table.

The *observed counts* are what was actually observed in the study. The *expected cell counts* are what you would expect for the count in a cell if the null hypothesis were true. The observed counts for the step volume data are shown in the following table. We will use the row and column totals to calculate the chi-square statistic.

	Low	Moderate	High	Total
Deceased	32	16	24	72
Alive	416	847	775	2038
Total	448	863	799	2110

The first thing we will do is find the expected counts under the assumption of no association. Remember that these expected counts will make it so the proportion of deaths is the same in each step volume group. To find out how many subjects we expect to be in the (Low, Deceased) cell of the table we first find the overall proportion of subjects who were deceased in the sample and multiply this by the total number of subjects who are in the low step volume group.

$$Expected\ count = \frac{72}{2110} \times 448 \approx 15.29$$

As you can see, if there was no association between the step volume group and mortality, we would expect to see a lot fewer subjects in the (Low, Deceased) cell. (As this is an "expected" count, you do not need to round it to an integer value.)

9. Now let's calculate the chi-square statistic.
   a. We've included the expected count for the (Low, Deceased) group in the table below. Now calculate the expected counts for the remaining five cells in the table, rounding to two decimal places for each.

	Low	Moderate	High	Total
Deceased	32/15.29	16/	24/	72
Alive	416/	847/	775/	2038
Total	448	863	799	2110

   b. To compare the discrepancy between the observed and expected counts for the (Low, Deceased) cell, we find the square of the difference in the observed and expected count and divide it by the expected count.

$$\frac{(Observed - Expected)^2}{Expected} = \frac{(32 - 15.29)^2}{15.29} \approx 18.27$$

   Now find these terms for the remaining five cells.

	Low	Moderate	High
Deceased	18.27		
Alive			

   c. Add up the six values (often called the components of the chi-square statistic) to calculate the chi-square statistic.

10. In the applet, change the **Statistic** in the pull-down menu to **χ^2**. What does the applet report for the observed value of the chi-square statistic? Confirm that this matches your answer to the previous questions. (*Note:* Value may be a bit different due to rounding.) Also click on the **Show χ^2 output** box and verify that the components of the chi-square statistic match those that you calculated.

11. The applet will also now display the null distribution for the χ^2 statistic rather than the Mean Group Diff statistic. Determine the p-value based on the simulated χ^2 statistics. (*Hint*: Change the value in the **Count Samples** box to the observed value of the χ^2 statistic.) How does the p-value based on the χ^2 statistic compare to the one based on the Mean Group Diff statistic? Are

they similar? Is your conclusion about strength of evidence provided by the data against the null hypothesis the same?

Theory-based approach

The primary advantage of using the χ^2statistic is that its null distribution can be predicted with a theoretical distribution. In fact, a theory-based approach could be used without conducting a simulation in the first place.

12. Below the graph of the simulated chi-square statistics, check the box to **Overlay Chi-square distribution**. Does the theoretical distribution match the distribution of simulated statistics reasonably well? How does the theory-based p-value compare to your simulation-based p-value using the chi-square statistic?

Validity conditions

Like all theory-based tests, this one also comes with the validity condition of having large sample sizes. We will (very conservatively) consider the sample size large enough if the sample data include at least 10 observations in each cell of the two-way table.

13. Go back to the applet and make sure the **Show Table** box is checked.
    a. How many cells are in this table? In other words, how many counts need to be checked that they are at least 10?

    b. Is the validity condition for a theory-based chi-square test satisfied for these data? Justify your answer.

Follow-up analysis

When a chi-square test produces a significant result, we conclude that at least one probability or population proportion differs from at least one of the others. A sensible next step is to try to identify *which* sample proportions differ significantly from which others. We can do this by producing confidence intervals for pairwise differences in population proportions.

14. Below the p-value output, check the box to **Compute 95% CI(s) for difference in proportions**.
    a. How many intervals are produced?

    b. Which sample proportions are significantly different from each other? How are you deciding?

    c. For one of the intervals you just identified, write a one-sentence interpretation of the interval, being very clear what is supposed to be captured inside the interval and which of the two step volumes has a larger probability of death within 11 years.

Generalization and Causation

15. Comment on how the sample was obtained and to which population your conclusion can be generalized.

16. Comment on how the study was designed and whether a cause-and-effect conclusion is warranted from this study. (Be sure to also consider the statistical significance of the results.)

Look back and ahead

17. Summarize your conclusion for the researchers. Do you have any concerns about the study design, any comments on sample size? Are there other limitations that you feel need to be addressed? Are there specific improvements you would suggest to make this study better?

Extension: What if we don't have a binary response?

The response for this study truly is binary; however, there are possible confounding variables that were also measured in the study that are of interest. Smoking status was one of those possible confounding variables. If step volume had been randomly assigned, we would expect there to be no association between smoking status and step volume. Because this was an observational study and no random assignment was used it is important to show that possible confounding variables are not associated with the explanatory variable, step volume. Use the following table of data gathered on smoking status to determine whether an association exists between step volume and smoking status.

	Low	Moderate	High
Never	273	541	489
Former	83	191	161
Current	90	124	142

To enter the data in the **Multiple Proportions** applet, click the check box by **Enter table** and select 3x3. Enter the labels low, moderate, and high for the three groups and never, former, current for the three outcomes. Fill in the nine squares in the body of the table with the counts of subjects in each cell/square. Then click on **Use Table**. You should also check the **Show table** box to see the two-way table and conditional proportions.

18. Create a segmented bar graph or mosaic plot to summarize this table. Discuss what the graph reveals (in context).

19. State your null and alternative hypotheses in terms of associations between step volume and smoking status. Note that this is a test of significance where we are hoping not to find evidence against the null hypothesis, rather we would like the null to be plausible.

20. Carry out the chi-square test. Report the test statistic and p-value. What is the strength of evidence of a genuine association between the variables in the population? Is this what we were hoping to find? Explain.

JAMA Network Open. 2021;4(9):e2124516. doi:10.1001/jamanetworkopen.2021.24516

Original Investigation | Nutrition, Obesity, and Exercise

Steps per Day and All-Cause Mortality in Middle-aged Adults in the Coronary Artery Risk Development in Young Adults Study

Amanda E. Paluch, PhD; Kelley Pettee Gabriel, PhD; Janet E. Fulton, PhD; Cora E. Lewis, MD; Pamela J. Schreiner, PhD; Barbara Sternfeld, PhD;
Stephen Sidney, MD; Juned Siddique, PhD; Kara M. Whitaker, PhD; Mercedes R. Carnethon, PhD