Comparison 2
COR1-GB.1305 – Statistics and Data Analysis

Confidence Intervals for Comparing Means

1. Recall the class survey. Seventeen female and thirty male students filled out the survey, reporting (among other variables) their GMAT scores and interest levels in the course. We will use this data to compare females and males.

   (a) What are the relevant populations?

   (b) For the 14 female respondents who reported their GMAT scores, the mean was 721 and the standard deviation was 27. For the 28 male respondents, the mean was 720 and the standard deviation was 39. Find a 95% confidence interval for the difference in population means.

   (c) For the 17 female respondents who reported their interest levels in the course (1–10), the mean was 5.8 and the standard deviation was 1.8. For the 30 male respondents, the mean was 6.3 and the standard deviation was 2.1. Find a 95% confidence interval for the difference in population means.

   (d) For the confidence intervals you constructed in parts (b) and (c) to be valid, what assumptions need to be satisfied? How could you check these assumptions?

Hypothesis Tests for Comparing Means

2. Consider again the class survey data. We will use the data to evaluate whether or not there is a significant difference between the female and the male population means.

   For the 14 female respondents who reported their GMAT scores, the mean was 721 and the standard deviation was 27. For the 28 male respondents, the mean was 720 and the standard deviation was 39. If the population means were equal what would be the chance of seeing a difference in sample means as large as observed?

Case Study: New York City Taxi Tips

3. To taxi riders tip differently in Brooklyn and Manhattan? We took a random sample of seventy-eight thousand New York City taxi trips from 2013 to find out. Of these, 76050 started and ended in Manhattan; 1197 started and ended in Brooklyn. All trips paid with credit card (not cash). For the Manhattan trips, the mean and standard deviation of the tip percentages were 19.21 and 9.23. For the Brooklyn trips, the mean and standard deviation of the tip percentages were 20.61 and 11.48.

   (a) What are the relevant populations?

   (b) What are the null and alternative hypotheses?

   (c) What are the samples?

(d) What is the test statistic?

(e) Approximately what is the p-value? What is the result of the test?

(f) Find a 99% confidence interval for the difference in average tip rates between Manhattan and Brooklyn.

(g) What assumptions do you need for the hypothesis test and the confidence interval to be valid?

Paired Comparisons

4. Suppose That you want to compare the mean daily rates of return for two stocks. For $n = 30$ consecutive trading days, you record the daily returns of the two stocks.

   (a) Consider two samples. The first sample is $X_{1,1}, X_{1,2}, \ldots, X_{1,30}$, the consecutive returns of the first stock. The second sample is $X_{2,1}, X_{2,2}, \ldots, X_{2,30}$, the consecutive returns of the second stock. Are these samples independent of each other? Why or why not?

   (b) How could we test whether or not one stock performs better than the other on average?

5. (Adapted from Stine and Foster, 4M Example 17.4)

   Two pharmaceutical companies (call them A and B) are about to merge. Senior management plans to eliminate one companies sales force. Which one should they eliminate?

   To decide this we will take sales data from similar products in 20 comparable geographical districts. For each district, we have the average dollar sales per representative per day in that district. Because each district has its own mix of population, cities, and cultures, it makes the most sense to directly compare the sales forces in each district. We will use the difference obtained by subtracting sales for Division B from sales of Division A in each district.

   (a) What is the population?

   (b) What is the sample?

   (c) Find a 95% confidence interval for expected difference in sales (per representative per day) between Division A and Division B after adjusting for district. Use the following information: number of districts is 20; average difference (A−B, in dollars) is −13.5; standard deviation difference is 26.7474.

   (d) Is there evidence that one division is better than the other?