Comparison 2 – Solutions

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?

   > **Solution:** There are two populations: all first-year female Stern MBA students, and all first-year male Stern MBA students.

   (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.

   > **Solution:**
   > Let sample 1 be the female GMAT scores: $n_1 = 14$, $\bar{x}_1 = 721$, $s_1 = 27$. Let sample 2 be the male GMAT scores: $n_2 = 28$, $\bar{x}_2 = 720$, $s_2 = 39$. We have
   >
   > $$\begin{aligned} \bar{x}_1 - \bar{x}_2 &= 721 - 720 \\ &= 1, \\ \mathrm{se}(\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= \sqrt{\frac{(27)^2}{14} + \frac{(39)^2}{28}} \\ &= 10. \end{aligned}$$
   >
   > An approximate 95% confidence interval for the difference between Stern MBA1 female and male average GMAT scores is
   >
   > $$\begin{aligned} (\bar{x}_1 - \bar{x}_2) \pm 2\mathrm{se}(\bar{x}_1 - \bar{x}_2) &= 1 \pm (2)(10) \\ &= 1 \pm 20 \\ &= (-19, 21). \end{aligned}$$
   >
   > (Note: a more precise confidence interval would use 1.96se instead of 2se.)

   (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.

   > **Solution:**
   > Let sample 1 be the female interest levels: $n_1 = 17$, $\bar{x}_1 = 5.8$, $s_1 = 1.8$. Let sample 2

be the male interest levels: $n_2 = 30$, $\bar{x}_2 = 6.3$, $s_2 = 2.1$. We have

$$\begin{aligned}
\bar{x}_1 - \bar{x}_2 &= 5.8 - 6.3 \\
&= -0.5, \\
\operatorname{se}(\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
&= \sqrt{\frac{(1.8)^2}{17} + \frac{(2.1)^2}{30}} \\
&= 0.3.
\end{aligned}$$

An approximate 95% confidence interval for the difference between Stern MBA1 female and male interest levels is

$$\begin{aligned}
(\bar{x}_1 - \bar{x}_2) \pm 2\operatorname{se}(\bar{x}_1 - \bar{x}_2) &= (-0.5) \pm (2)(0.3) \\
&= -0.5 \pm 0.6 \\
&= (-1.1, 0.1).
\end{aligned}$$

(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?

Solution: We need that the observed samples are simple random samples from the population. (We need the samples to be unbiased.) It is impossible to check this assumption, but it seems reasonable.

Since the sample sizes are small, we need for the populations to be normal. We could check this by looking at histograms of the samples.

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?

> **Solution:** To answer this, we first compute a test statistic:
>
> $$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{se}(\bar{x}_1 - \bar{x}_2)} \\ &= \frac{1}{10} \\ &= 0.1. \end{aligned}$$
>
> If the population means were equal, then the test statistic would be approximately normally distributed; the chance of seeing a difference in sample means as large as observed would be
>
> $$\begin{aligned} p &\approx P(|Z| \geq 0.1) \\ &\approx 0.9203. \end{aligned}$$
>
> That is, it would be very typical to see such a difference.

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?

Solution: Population 1: the tip percentages of all credit-card paying taxi trips starting and ending in Manhattan.

Population 2: the tip percentages of all credit-card paying taxi trips starting and ending in Brooklyn.

(b) What are the null and alternative hypotheses?

Solution:

$$H_0 : \mu_1 = \mu_2 \quad \text{(same mean tip amount for both boroughs)}$$
$$H_a : \mu_1 \neq \mu_2$$

Here μ_1 is the average tip amount for all trips starting and ending in Manhattan μ_2 is the average tip amount for all trips starting and ending in Brooklyn.

(c) What are the samples?

Solution: Sample 1: the $n_1 = 76060$ sampled tip percentages from Manhattan, with mean $\bar{x}_1 = 19.21$ and standard deviation $s_1 = 9.23$.

Sample 2: the $n_2 = 1197$ sampled tip percentages from Manhattan, with mean $\bar{x}_2 = 20.61$ and standard deviation $s_1 = 11.48$.

(d) What is the test statistic?

Solution: We first compute

$$\begin{aligned}\bar{x}_1 - \bar{x}_2 &= 19.21 - 20.61\\ &= -1.40,\\ \text{se}(\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\\ &= \sqrt{\frac{(9.23)^2}{76060} + \frac{(11.48)^2}{1197}}\\ &= 0.33.\end{aligned}$$

The test statistic is

$$\begin{aligned}t &= \frac{\bar{x}_1 - \bar{x}_2}{\text{se}(\bar{x}_1 - \bar{x}_2)}\\ &= \frac{-1.40}{0.33}\\ &= -4.24.\end{aligned}$$

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

Solution: The p-value is approximately given by

$$\begin{aligned}p &\approx P(|Z| > 4.24)\\ &\approx 0.00006334.\end{aligned}$$

This is very strong evidence to reject H_0. That is, there is strong evidence in a difference in average tip rates between Brooklyn and Manhattan for all New York City taxi trips paid for with credit cards.

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

Solution: For a 99% confidence interval, $\alpha = .01$ and $z_{\alpha/2} = 2.576$ (use the df $= \infty$ box from the t table). The 99% confidence interval for the difference (Manhattan − Brooklyn) is

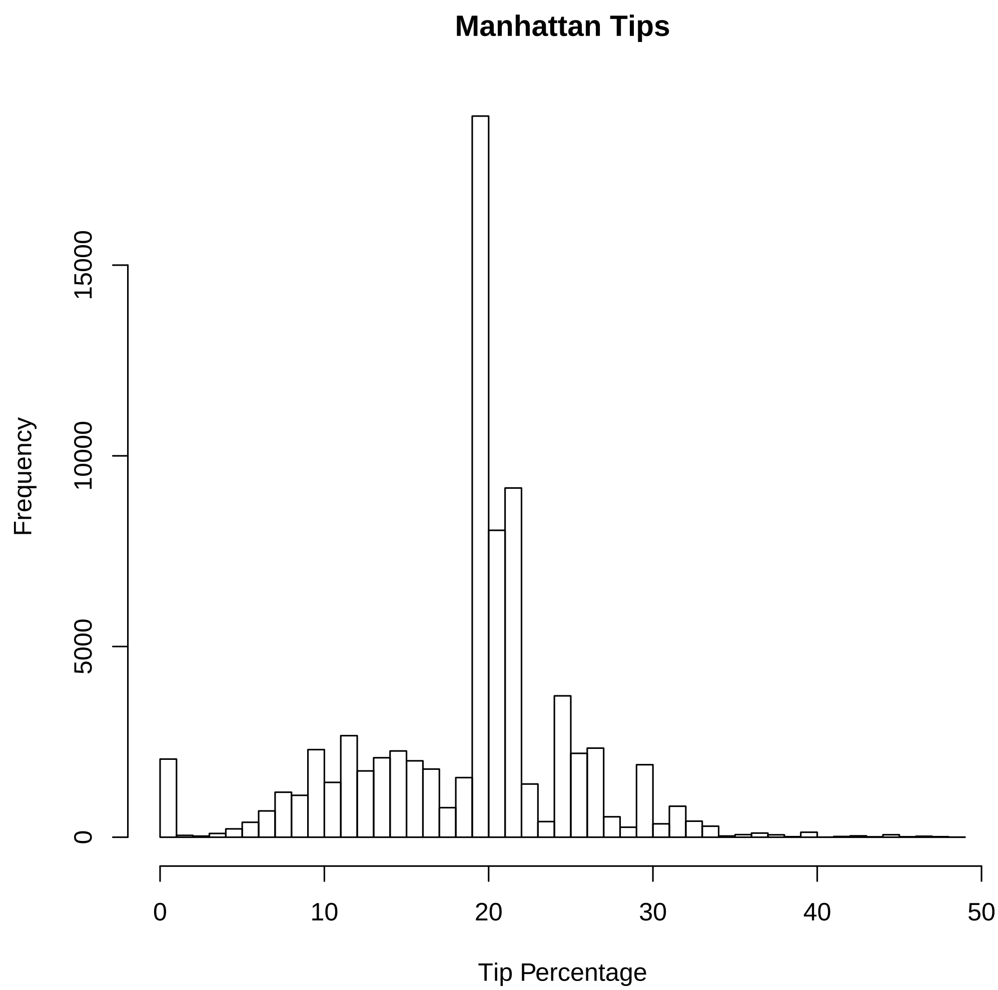
$$\begin{aligned}(\bar{x}_1 - \bar{x}_2) \pm 2.576\text{se}(\bar{x}_1 - \bar{x}_2) &= (-1.40) \pm (2.576)(0.33)\\ &= -1.40 \pm 0.85\\ &= (-2.25, -0.55).\end{aligned}$$

We are 99% confident that on average, riders in Brooklyn tip between 2.25% and 0.55% more than riders in Manhattan.

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

Solution: We need that the observed samples are simple random samples from the population (no bias in the samples). We do not need that the populations are normal, because the sample sizes are so large ($n_1 \geq 30$ and $n_2 \geq 30$).

(Here is a histogram of the Manhattan tip percentages, which are clearly non-normal.)



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?

   Solution: Dependent. There could be daily shocks (market events) affecting both stocks; these shocks would make the performances of the two stocks dependent.

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

   Solution: Look at the differences $D_i = X_{1,i} - X_{2,i}$, and test whether or not $\mu_D = 0$ using a test on a population mean (the usual z or t test).

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?

   Solution: The differences in sales per representative per day between the two companies within all comparable districts (not just the 20 districts in the sample).

   (b) What is the sample?

   Solution: The observed differences between the 20 districts in 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.

Solution: For an approximate 95% confidence interval, we use

$$\begin{aligned}\bar{x} \pm 2\frac{s}{\sqrt{n}} &= (-13.5) \pm 2\frac{26.7474}{\sqrt{20}} \\ &= -13.5 \pm 12.0 \\ &= (-25.5, -1.5).\end{aligned}$$

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

Solution: Yes, there is evidence of a difference. We are 95% confident that sales for Division B are higher.