Comparison 2 – Solutions

COR1-GB.1305 – Statistics and Data Analysis

Case Study: New York City Taxi Tips

1. 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

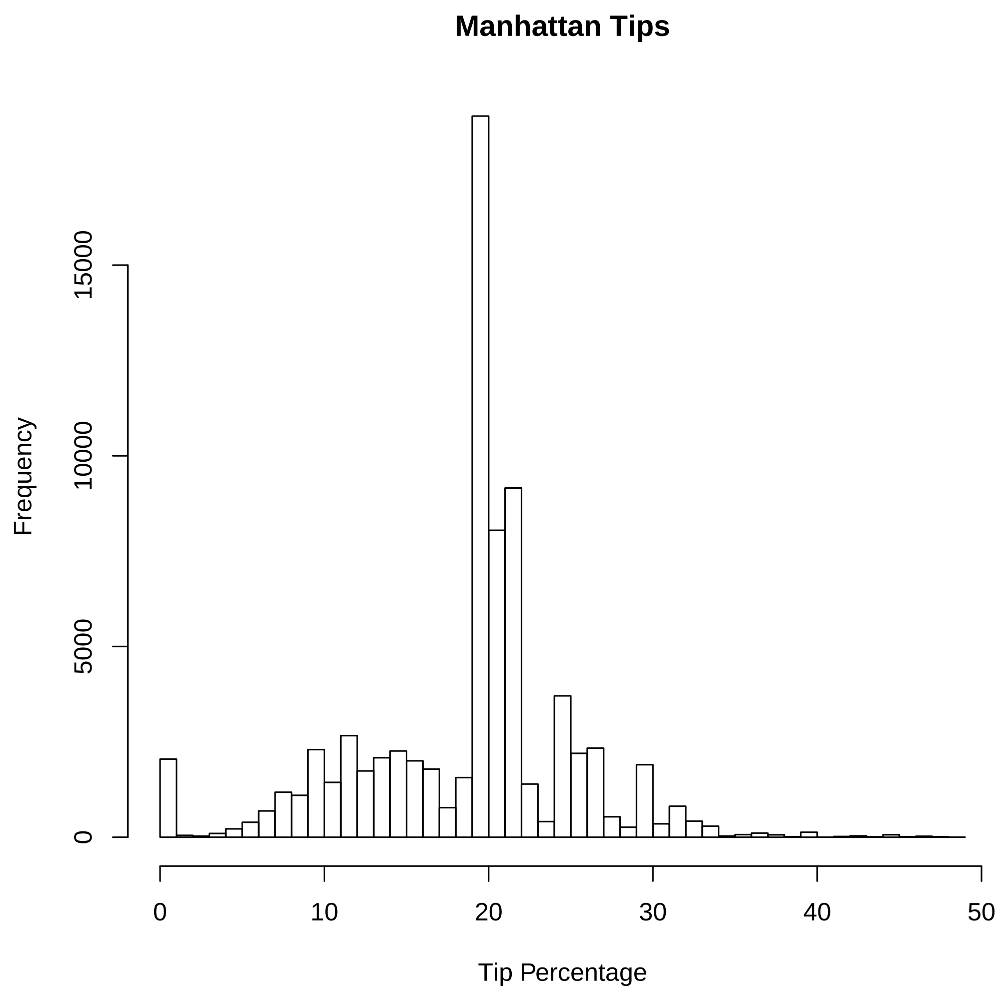
$$\begin{aligned}
(\bar{x}_1 - \bar{x}_2) \pm 2.576\text{se}(\bar{x}_1 - \bar{x}_2) &= (-1.40) \pm (2.576)(0.33) \\
&= -1.40 \pm 0.85 \\
&= (-2.25, -0.55).
\end{aligned}$$

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

2. (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.

Paired and unpaired samples

3. In the following situations, are the samples paired or unpaired?

   (a) You want to compare the performances of two restaurants. You measure the weekly profits of both restaurants for 10 consecutive weeks.

   Solution: Paired.

   (b) You want to compare expected starting salaries between males and females using the class survey data.

   Solution: Unpaired.

   (c) Your company can use one of two possible advertisements. You show one ad to one group of people, and ask them to rate the likelihood of buying your product after seeing the ad. You show the second ad to a second group of people, and ask them the same question.

   Solution: Unpaired.

   (d) Your company can use one of two possible advertisements. You show both ads to a group of people, and ask them to rate their opinions of both ads.

   Solution: Paired.