**Comparison – Solutions**
COR1-GB.1305 – Statistics and Data Analysis

# Confidence Intervals

1. Recall the class survey. Thirteen female and twenty-four 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 Langone students, and all first-year male Langone students.

   (b) For the 7 female respondents who reported their GMAT scores, the mean was 671 and the standard deviation was 27. For the 15 male respondents, the mean was 682 and the standard deviation was 38. Find a 95% confidence interval for the difference in population means.

   > **Solution:**
   > Let sample 1 be the female GMAT scores: $n_1 = 7$, $\bar{x}_1 = 671$, $s_1 = 27$. Let sample 2 be the male GMAT scores: $n_2 = 15$, $\bar{x}_2 = 682$, $s_2 = 38$. We have
   >
   > $$\begin{aligned} \bar{x}_1 - \bar{x}_2 &= 671 - 682 \\ &= -11, \\ \text{se}(\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= \sqrt{\frac{(27)^2}{7} + \frac{(38)^2}{15}} \\ &= 14. \end{aligned}$$
   >
   > An approximate 95% confidence interval for the difference between Langone female and male average GMAT scores is
   >
   > $$\begin{aligned} (\bar{x}_1 - \bar{x}_2) \pm 2\text{se}(\bar{x}_1 - \bar{x}_2) &= (-11) \pm (2)(14) \\ &= -11 \pm 28 \\ &= (-39, 17). \end{aligned}$$
   >
   > (Note: a more precise confidence interval would use 1.96se instead of 2se.)

   (c) For the 13 female respondents who reported their interest levels in the course (1–10), the mean was 6.2 and the standard deviation was 2.5. For the 24 male respondents, the mean was 6.5 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: $n_1 = 13$, $\bar{x}_1 = 6.2$, $s_1 = 2.5$. Let sample 2

be the male GMAT scores: $n_2 = 24$, $\bar{x}_2 = 6.5$, $s_2 = 2.1$. We have

$$\begin{aligned}
\bar{x}_1 - \bar{x}_2 &= 6.2 - 6.5 \\
&= -0.3, \\
\text{se}(\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
&= \sqrt{\frac{(2.5)^2}{13} + \frac{(2.1)^2}{24}} \\
&= 0.8.
\end{aligned}$$

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

$$\begin{aligned}
(\bar{x}_1 - \bar{x}_2) \pm 2\text{se}(\bar{x}_1 - \bar{x}_2) &= (-0.3) \pm (2)(0.8) \\
&= -0.3 \pm 1.6 \\
&= (-1.9, 1.3).
\end{aligned}$$

(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

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.

   (a) For the 7 female respondents who reported their GMAT scores, the mean was 671 and the standard deviation was 27. For the 15 male respondents, the mean was 682 and the standard deviation was 38. 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:

   $$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{se}(\bar{x}_1 - \bar{x}_2)} \\ &= \frac{-11}{14} \\ &= -0.79 \end{aligned}$$

   If the population means were equal, then the test statistic would be normally distributed; the chance of seeing a difference in sample means as large as observed would be

   $$\begin{aligned} p &\approx P(|Z| \geq 0.79) \\ &\approx 0.4237. \end{aligned}$$

   That is, it would be very typical to see such a difference.

   (b) For the 13 female respondents who reported their interest levels in the course (1–10), the mean was 6.2 and the standard deviation was 2.5. For the 24 male respondents, the mean was 6.5 and the standard deviation was 2.1. If the population means were equal what would be the chance of seeing a difference in sample means as large as observed?

   **Solution:** This is similar to the previous problem:

   $$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{se}(\bar{x}_1 - \bar{x}_2)} \\ &= \frac{-0.3}{0.8} \\ &= -0.375 \\ p &\approx P(|Z| \geq 0.375) \\ &\approx 0.6892. \end{aligned}$$

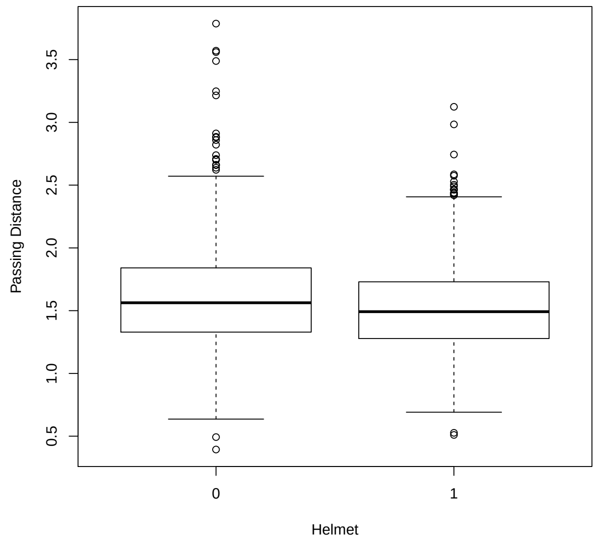
   It would be very typical to see a difference in sample means as large as observed.

(c) What is the relationship between the confidence intervals in Question 1 and your answers to parts (a) and (b)?

**Solution:** When the 95% confidence intervals contain 0, the $p$-values for testing for a difference are greater than 0.05 (the differences are not significant at level 5%).

## Case Study: Bicycle Passing Distance

3. Here are boxplots of the passing distances (in meters) for a bike rider with and without a helmet. Is there evidence that the passing distance differs when the rider has a helmet?




Here are the sample statistics for the passing distance without a helmet: $n_1 = 1206$, $\bar{x}_1 = 1.61$, $s_1 = 0.405$. Here are the sample Here are the sample statistics for the passing distance with a helmet: $n_2 = 1149$, $\bar{x}_2 = 1.52$, $s_2 = 0.354$.

Formulate the problem as a hypothesis test, using significance level 5%.

(a) What are the populations?

> **Solution:** Population 1: all passing distances while not wearing a helmet.
> Population 2: all passing distances while wearing a helmet.

(b) What are the null and alternative hypotheses?

> **Solution:**
>
> $$H_0 : \mu_1 = \mu_2 \quad \text{(same mean distance for both populations)}$$
> $$H_a : \mu_1 \neq \mu_2.$$

(c) What are the samples?

**Solution:** All recorded passing distances, without and with a helmet.

(d) What is the test statistic?

**Solution:**

$$\begin{aligned}
\bar{x}_1 - \bar{x}_2 &= 1.61 - 1.52 \\
&= 0.09 \\
\mathrm{se}(\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{(0.405)^2}{1206} + \frac{(0.354)^2}{1149}} \\
&= 0.016 \\
z &= \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{se}(\bar{x}_1 - \bar{x}_2)} \\
&= \frac{0.09}{0.016} \\
&= 5.6
\end{aligned}$$

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

**Solution:**

$$\begin{aligned}
p &\approx P(|Z| \geq 5.6) \\
&< 5.733 \times 10^{-7}.
\end{aligned}$$

If there were no difference in average passing distance with and without a helmet, then there would be less than a $5.733 \times 10^{-5}$ chance of seeing data like that observed. There is substantial evidence of a difference; we reject $H_0$.

(f) Find a 95% confidence interval for the difference in passing difference with and without a helmet.

**Solution:**

$$0.09 \pm 2(0.016)$$

With 95% confidence, the difference in population means is between 0.058 and 0.122 meters.