Comparison – Solutions

COR1-GB.1305 – Statistics and Data Analysis

More p-values

1. Suppose we perform a hypothesis test and we observe a p-value of $p = .02$. True or false: There is a 2% chance that the null hypothesis is true.

> **Solution:** False. The p-value is the probability of getting a test statistic at least as extreme as what was observed. Heuristically, we can think of this as
>
> $$\mathrm{P}(\text{Data} \mid H_0 \text{ is true}) = 2\%.$$
>
> The statement in the problem is
>
> $$\mathrm{P}(H_0 \text{ is true} \mid \text{Data}) = 2\%.$$
>
> Clearly, this is not the same.

2. Suppose we perform a hypothesis test and we observe a p-value of $p = .02$. True or false: If we reject the null hypothesis, then there is a 2% chance of making a type I error.

> **Solution:** False. We can only make a type I error when the null hypothesis is true. Thus, the statement in question 2 is *exactly the same* as the statement in question 1.

3. Suppose we perform a hypothesis test and we observe a T test statistic $t = -2.02$, corresponding to a p-value of $p = .02$. True or false: If we were to repeat the experiment and the null hypothesis were actually true, then there would be a 2% chance of observing a test statistic at least as extreme as $t = -2.02$.

> **Solution:** True. The p-value is the probability of getting a test statistic as least as extreme as the observed value if the null hypothesis were true. Note: for a one-sided less-than alternative, extreme means "less than or equal to."

Confidence Intervals for Comparing Means

4. 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

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

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

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

(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

5. 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 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.

   (b) For the 27 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. 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:

   $$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{se}(\bar{x}_1 - \bar{x}_2)} \\ &= \frac{-0.5}{0.3} \\ &= -1.7 \\ p &\approx P(|Z| \geq 1.7) \\ &\approx 0.08913. \end{aligned}$$

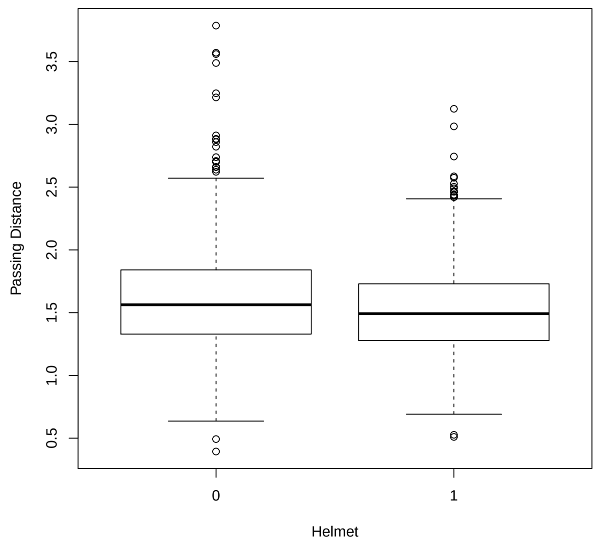
   It would be very typical to see a difference in sample means as large as observed.

(c) What is the relationship between the confidence intervals in Question 4 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

6. 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:

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

(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 5.733×10^{-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.