Hypothesis Test and Comparison – Solutions
STAT-UB.0001 – Statistics for Business Control

Test on a Population Mean

1. (Adapted from Stine and Foster, 4M 16.2). Does stock in IBM return a different amount on average than T-Bills? We will attempt to answer this question by using a dataset of the 264 monthly returns from IBM between 1990 and 2011. Over this period, the mean of the monthly IBM returns was 1.26% and the standard deviation was 8.27%. We will take as given that the expected monthly returns from investing in T-Bills is 0.3%.

   (a) What is the sample? What are the sample mean and standard deviation?

   Solution: The $n = 264$ monthly IBM returns from 1990 to 2011. The sample mean and standard deviation (in %) are

   $$\bar{x} = 1.26$$
   $$s = 8.27$$

   (b) What is the relevant population? What are the interpretations of population mean and standard deviation?

   Solution: All monthly IBM returns (past, present, and future). The population mean, μ represents the expected return for a month in the future. The population standard deviation, σ, represents the standard deviation of the monthly returns for all months (past, present, and future).

   (c) What are the null and alternative hypotheses for testing whether or not IBM gives a different expected return from T-Bills (0.3%)?

   Solution:

   $$H_0 : \mu = 0.3$$
   $$H_a : \mu \neq 0.3$$

(d) Use an appropriate test statistic to summarize the evidence against the null hypothesis.

> **Solution:** If the null hypothesis were true ($\mu = 0.3$), then the sample mean would have been a normal random variable with mean $\mu_{\bar{X}} = 0.3$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. The test statistic
> $$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$
> would follow a t distribution with $n - 1 = 263$ degrees of freedom. The observed value of this statistic is
> $$t = \frac{1.26 - 0.3}{8.27/\sqrt{264}} = 1.886$$

(e) If the null hypothesis were true (there were no difference in expected monthly returns between IBM and T-Bills) what would be the chance of observing data at least as extreme as observed?

> **Solution:** If we approximate the distribution of the test statistic under H_0 as a standard normal random variable, then the chance of observing data at least as extreme as observed would be
> $$p = P(|Z| > 1.886) \approx 0.05743.$$

(f) Is there compelling evidence (at significance level 5%) of a difference in expected monthly returns between IBM and T-Bills?

> **Solution:** No, since $p \geq 0.05$, there is not compelling evidence.

(g) What assumptions do you need for the test to be valid? Are these assumptions plausible?

> **Solution:** Since $n \geq 30$, we do not need to assume that the population is normal. We need that the observed sample is a simple random sample from the population; this might not hold if the period under observation (1990–2011) is not typical for IBM.

Test Statistic and Observed Significance Level (p-value)

2. In each of the following examples, for the hypothesis test with

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

find the test statistic (t) and the p-value.

(a) $\mu_0 = 5$; $\bar{x} = 7$; $s = 10$; $n = 36$.

Solution:

$$\begin{aligned} t &= \frac{7-5}{10/\sqrt{36}} \\ &= 1.2 \\ p &\approx P(|Z| > 1.2) \\ &= 0.2301 \end{aligned}$$

(b) $\mu_0 = 90$; $\bar{x} = 50$; $s = 200$; $n = 64$.

Solution:

$$\begin{aligned} t &= \frac{50-90}{200/\sqrt{64}} \\ &= -1.6 \\ p &\approx P(|Z| > 1.6) \\ &= 0.1096 \end{aligned}$$

(c) $\mu_0 = 50$; $\bar{x} = 49.4$; $s = 2$; $n = 100$.

Solution:

$$\begin{aligned} t &= \frac{49.4-50}{2/\sqrt{100}} \\ &= -3 \\ p &\approx P(|Z| > 3) \\ &= 0.002700 \end{aligned}$$

3. For each example from problem 2:

(a) Indicate whether a level 5% test would reject H_0.

Solution: We reject H_0 if $p < 0.05$: (a) do not reject H_0; (b) do not reject H_0; (c) reject H_0.

(b) Indicate whether a level 1% test would reject H_0.

Solution: We reject H_0 if $p < 0.01$: (a) do not reject H_0; (b) do not reject H_0; (c) reject H_0.

4. Before Facebook's recent redesign, the mean number of ad clicks per day was 100K. In the 49 days after the redesign, the mean number of ad clicks per day was 105K and the standard deviation was 35K. Is there significant evidence that the redesign affected the expected number of ad clicks? Perform a test at the 5% level.

   (a) What is the sample? What is the population?

   > **Solution:** The sample is the number of ad clicks on the measured $n = 49$ days.
   > The population is the number of ad clicks on all days after the redesign.

   (b) What are the null and alternative hypotheses?

   > **Solution:** Let μ be the expected clicks per day after the redesign We use "thousands of clicks" as the units for all relevant quantities.
   > The null hypothesis is that the redesign had no effect on expected ad clicks. The alternative hypothesis is that μ changed after the redesign:
   >
   > $$\begin{aligned} H_0 &: \mu = 100 \\ H_a &: \mu \neq 100 \end{aligned}$$

   (c) What is the test statistic?

   > **Solution:** The test statistic is based on the sample, the observed clicks in the $n = 49$ days after the redesign. Let $\bar{x}$ denote the mean clicks per day in the sample; let s denote the sample standard deviation. Our test statistic is
   >
   > $$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{105 - 100}{35/\sqrt{49}} \\ &= 1. \end{aligned}$$

   (d) Approximately what is the p-value?

   > **Solution:**
   >
   > $$\begin{aligned} p &\approx P(|Z| > 1) \\ &= 0.3137. \end{aligned}$$

   (e) What assumptions are you making?

   > **Solution:** We need to assume that the observed sample is a simple random sample from the population.

Admittedly, the assumption does *not* hold, since there is strong selection bias in the sample: we are sampling days right after the website redesign with higher probability than days far into the future. For example, we have no chance of sampling a day three years into the future.

Since the assumption does not hold, we are in a bit of an awkward position. The hypothesis test may not be valid. One way in which it may not be valid is the following: it usually takes people a few weeks to adjust to a website redesign, so the ad click behavior in our sample of 49 days may not be representative of all future ad click behavior.

(f) What is α? What is the result of the test?

Solution: Despite the caveats mentioned in part (d), we will proceed with the test. For a level-5% test, $\alpha = 0.05$. Since $p \geq 0.05$, we do not reject H_0. If there were no difference in expected ad clicks before and after the redesign, there would be a 31.37% chance of seeing data like we observed. There is no evidence of a difference.

Types of Errors

5. In a hypothesis test, our decision will either be "reject H_0" or "do not reject H_0". Under what situations will each of these decisions be in error?

> **Solution:** Type I error: H_0 is true, but we reject it.
> Type II error: H_0 is false, and we fail to reject it.

6. We reject H_0 when the p-value is below α.

   (a) If H_0 is true, what is the probability of making a Type I error?

   > **Solution:** We reject H_0 when the p-value is less than α. This happens when $|T| > t_{\alpha/2,n-1}$. So, if the null hypothesis is true, then the probability of making a Type I error is
   > $$P(|T| > t_{\alpha/2,n-1}) = \alpha.$$

   (b) If H_0 is false, what is the probability of *not* making a Type II error?

   > **Solution:** We cannot give a direct answer to this question, because it depends on the true value of μ. In general, the probability of not making a Type II error is called the "power" of the test; it is given the symbol β or $\beta(\mu)$. If μ is close to μ_0, then β will be small (close to α); if μ is far from μ_0, then β will be large (close to 1).

More p-values

7. 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.

8. 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 8 is *exactly the same* as the statement in question 7.

9. 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

10. 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 female Stern MBA students, and all 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

$$\begin{aligned} \bar{x}_1 - \bar{x}_2 &= 721 - 720 \\ &= 1, \\ \mathrm{sd}(\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= \sqrt{\frac{(27)^2}{14} + \frac{(39)^2}{28}} \\ &= 10. \end{aligned}$$

An approximate 95% confidence interval for the difference between Stern MBA female and male average GMAT scores is

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) \pm 2\mathrm{sd}(\bar{x}_1 - \bar{x}_2) &= 1 \pm (2)(10) \\ &= 1 \pm 20 \\ &= (-19, 21). \end{aligned}$$

(Note: a more precise confidence interval would use 1.96sd instead of 2sd.)

(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

$$\begin{aligned}
\bar{x}_1 - \bar{x}_2 &= 5.8 - 6.3 \\
&= -0.5, \\
\mathrm{sd}(\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
&= \sqrt{\frac{(1.8)^2}{17} + \frac{(2.1)^2}{30}} \\
&= 0.58.
\end{aligned}$$

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

$$\begin{aligned}
(\bar{x}_1 - \bar{x}_2) \pm 2\mathrm{sd}(\bar{x}_1 - \bar{x}_2) &= (-0.5) \pm (2)(0.58) \\
&= -0.5 \pm 1.16 \\
&= (-1.66, 0.66).
\end{aligned}$$

(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

11. 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:

    $$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{sd}(\bar{x}_1 - \bar{x}_2)} \\ &= \frac{1}{10} \\ &= 0.1. \end{aligned}$$

    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:

    $$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{sd}(\bar{x}_1 - \bar{x}_2)} \\ &= \frac{-0.5}{0.58} \\ &= -0.86 \\ p &\approx P(|Z| \geq 0.86) \\ &\approx 0.3898. \end{aligned}$$

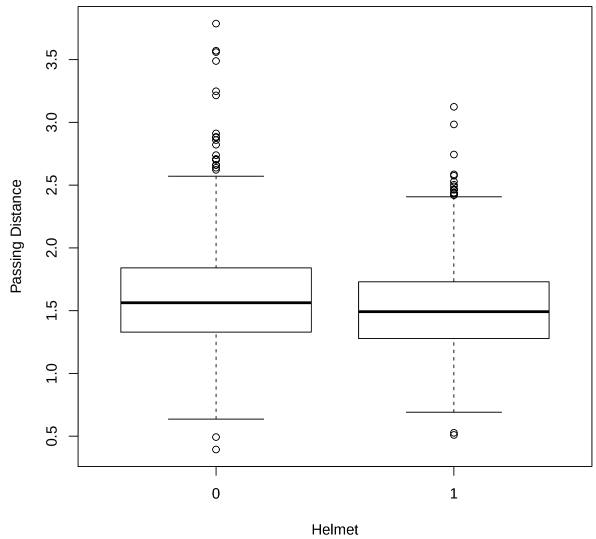
    It would be very typical to see a difference in sample means as large as observed.

(c) What is the relationship between the confidence intervals in Question 10 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

12. 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:

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

(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.