Chapter 11: Inferring Population Means

1 Sample Means of Random Samples

As you learned in previous chapter, we estimate population parameters by collecting a random sample from that population. We use the collected data to calculate a statistic, and this statistic is used to estimate the parameter. Whether we are using the statistic $\hat{p}$ to estimate the parameter p or are using $\bar{x}$ to estimate μ, if we want to know how close our estimate is to the truth, we need to know how far away that statistic is, typically, from the parameter.

Just as we did in previous chapter with $\hat{p}$, we now examine three characteristics of the behavior of the sample mean: its accuracy, its precision, and its probability distribution. By understanding these characteristics, we'll be able to measure how well our estimate performs and thus make better decisions. As a reminder, table below shows some commonly used statistics and the parameters they estimate.

Statistics	**Symbols**
Sample size	n
Sample mean	$\overline{x}$
Sample standard deviation	s
Sample variance	s^2
Sample proportion	$\widehat{p}$

Parameters	**Symbols**
Population mean	μ
Population standard deviation	σ
Population variance	σ^2
Population proportion	p

REMEMBER!!! It is extremely important to figure out the difference between statistics and parameters.

2 The Central Limit Theorem for Sample Means

The Central Limit Theorem (CLT) assures us that no matter what the shape of the population distribution, if a sample is selected such that the following conditions are met, then the distribution of sample means follows an approximately Normal distribution. The mean of this distribution is the same as the population mean. The standard deviation (also called the standard error) of this distribution is the population standard deviation divided by the square root of the sample size. As a rule of thumb, sample sizes of 25 or more may be considered "large."

When determining whether you can apply the Central Limit Theorem to analyze data, there are three conditions to consider:

Condition 1: Random Sample and Independence.

Condition 2: Large Sample. Either the population distribution is Normal or the sample size is large. General rule of thumb is sample size is greater than 25.
Condition 3: Big Population. The population must be at least 10 times larger than the sample size.

3 The $t-$distribution

The hypothesis tests and confidence intervals that we will use for estimating and testing the mean are based on a statistic called the $t-$statistic:

$$t = \frac{\bar{x}-\mu}{SE}$$

$$SE = \frac{s}{\sqrt{n}}$$

In real life, we almost never know the value of σ, the population standard deviation. Instead, we replace it with an estimate: the sample standard deviation, s.

The $z-$statistic follows an approximately Normal distribution if the sample size is large enough.

However, we rarely get to use a $z-$statistic and so must instead use the $t-$statistic. The $t-$statistic does not follow the Normal distribution. One reason for this is that the denominator changes with every sample. For this reason, the t-statistic is more variable than the $z-$statistic (whose denominator is the same in each sample of the same size). Instead, if the three conditions for using the Central Limit Theorem hold, the $t-$statistic follows a distribution called—surprise!—the $t-$**distribution.**

The $t-$distribution shares many characteristics with the Normal distribution. Both are symmetric, are unimodal, and might be described as “bell-shaped.” However, the $t-$distribution has thicker tails. This means that in a $t-$distribution, it is more likely that we will see extreme values (values far from the mean) than it is in a standard Normal distribution.

The t-distribution’s shape depends on only one parameter, called the **degrees of freedom (df).** The number of degrees of freedom is (usually) an integer: 1, 2, 3, and so on. If the df is small, then the $t-$distribution has very thick tails. As the degrees of freedom get larger, the tails get thinner. Ultimately, when the df is infinitely large, the $t-$distribution is exactly the same as the Normal distribution.

4 Constructing a confidence interval

We already studied in previous chapter how to construct the confidence intervals for one sample population proportions. To construct a confidence interval for one-sample population means, the steps are identical except here you have to use $t-$statistic. The steps to construct an one-sample $t-$interval are as follows:

- Identify sample mean $\overline{x}$, sample standard deviation s, and sample size n.

- Check if all the conditions of CLT are met.

- Find the estimated standard error calculated as $SE_{est} = \frac{s}{\sqrt{n}}$

- Find the margin of error. Margin of error (m) $= t^{*}SE_{est}$

- The t^* comes from the **t-table**. You need to look for $n-1$ degrees of freedom and the confidence level to find t^*. We will use technology to calculate the value of t^* and I will provide the value of t^* on an exam.

- Find the confidence interval using: $\overline{x} \pm m$

We discussed in detail about the theory being constructing confidence intervals and hypothesis testing for population proportions. The same logic translates to population means, so I won't be theory heavy for population means.

5 Some examples:

i. A random sample of 35 two-year colleges in 2014–2015 had a mean tuition (for in-state students) of \$4173, with a standard deviation of \$2590. Find a 90% confidence interval for the mean in-state tuition of all two-year colleges in 2010–2011. Interpret the intervals.

a. We are told that the sample is random and that the sample size is larger than 25, so the necessary conditions for CLT hold.

b. $\bar{x} = 4173$, $s = 2590$ and $n = 35$

c. We find the estimated standard error:

$$SE_{\text{est}} = \frac{2590}{\sqrt{35}} = 437.7899$$

d. We find the appropriate values of $t*$ from $t-$table and compute margin of error:

$$t^* = 1.691 \text{ (for 90\% confidence level)}$$

$$m = t^* SE_{\text{est}} = 1.691 \times 437.7899 = 740.3027$$

e. For the 90% confidence interval,

$$\begin{aligned} \bar{x} \pm t^* SE_{\text{est}} \;\; & \text{becomes } 4137 \pm 740.3027 \\ \text{Lower limit: } & 4137 - 740.3027 = 3432.70 \\ \text{Upper limit: } & 4173 + 740.3027 = 4913.30 \end{aligned}$$

A 90% confidence interval for the mean tuition of all two-year colleges in the $2014-2015$ academic year is $(\$3433, \$4913)$.

ii. A random sample of 50 12th-grade students was asked how long it took to get to school. The sample mean was 16.2 minutes, and the sample standard deviation was 12.3 minutes. Find a 95% confidence interval for the population mean time it takes 12th-grade students to get to school.

a. We are told that the sample is random and that the sample size is larger than 25, so the necessary conditions for CLT hold.

b. $\bar{x} = 16.2$, $s = 12.3$ and $n = 50$

c. We find the estimated standard error:

$$SE_{\text{est}} = \frac{16.2}{\sqrt{50}} = 1.739$$

d. We find the appropriate values of $t*$ from $t-$table and compute margin of error:

$$t^* = 2.001 \text{ (for 95\% confidence level)}$$

$$m = t^* SE_{\text{est}} = 2.001 \times 1.739 = 3.479$$

e. For the 90% confidence interval,

$$\bar{x} \pm t^* SE_{\text{est}} \text{ becomes } 16.2 \pm 3.479$$
$$\text{Lower limit: } 16.2 - 3.479 = 12.7$$
$$\text{Upper limit: } 16.2 + 3.479 = 19.7$$

A 95% confidence interval for the mean commute time of all 12th grade students is $(12.7, 19.7)$.

iii. Try it yourself and verify the answer: Data on the speed (in mph) for random sample of 30 cars travelling on a highway was collected. The mean speed was 63.3 mph with a standard deviation of 5.23 mph. Find the 95% confidence interval for the mean speed of all cars travelling on the highway. $(t* = 2.045)$

We are 95% confidence that the mean speed of cars on the highway is between 61.35 and 65.25 mph.

6 One sample hypothesis testing for means

- Step 1: Figure out μ, $\overline{x}$, s, and n from the problem.
- Step 2: Write down the pair of hypotheses as:

Two-sided	One-sided (Left)	One-sided (Right)
$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$	$H_0 : \mu = \mu_0$ $H_a : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_a : \mu > \mu_0$

- Step 3: Find the corresponding one-sample $t-$statistics. $t = \frac{\overline{x}-\mu_0}{\frac{s}{\sqrt{n}}}$
- Step 4: Find the corresponding p-value of your $t-$test statistics.
- Step 5: Compare your $p-$value with significance level α
  If p-value $< \alpha$, you will reject H_o
  If p-value $> \alpha$, you will fail to reject H_o
- For one sample hypothesis testing for means, $p-$value will be provided to you.

7 Some examples

1. In 2010, the mean years of experience among a nursing staff was 14.3 years. A nurse manager took a survey of a random sample of 35 nurses at the hospital and found a sample mean of 18.37 years with a standard deviation of 11.12 years. Do we have evidence that the mean years of experience among the nursing staff at the hospital has increased? Use a significance level of 0.05.

   Firstly, population mean (μ) = 14.3, sample mean $\bar{x}$ = 18.37, sample standard deviation (s) = 11.12, and sample size (n) = 35

   1. Hypotheses: $\mathrm{H_0} : \mu = 14.3 \quad \mathrm{H_a} : \mu > 14.3$

   2. Conditions of CLT are satisfied as our sample is random, independent sample; sample size is large $(35 > 25)$

   3. Compute $t-$statistic

$$SE_{EST} = \frac{s}{\sqrt{n}} = \frac{11.12}{\sqrt{35}} = 1.88$$
$$t = \frac{\bar{x}-\mu}{SE_{EST}} = \frac{18.37 - 14.3}{1.88} = 2.16$$

   4. Use technology to find the p-value that corresponds with $t = 2.16$ and $df = 34$.

   5. We get $p-$value of 0.019. Since $p-$value $>$ significance level i.e. $0.0109 < 0.05$, we will reject H_0

   6. Evidence suggests that mean experience among the nursing staff has increased.

2. In the 2010 season, average home attendance for NHL hockey games was $17,072$. Suppose a sports statistician took a random sample of 30 home hockey games during the 2014 season and found a sample mean of $18,104$ with a standard deviation of 1203.5. Can we conclude that mean attendance at NHL games has changed since the 2010 season?

   Firstly, population mean (μ) = 17072, sample mean $\bar{x} = 18104$, sample standard deviation (s) = 1203.5, and sample size (n) = 30

   1. Hypothesize: $\mathrm{H}_0 : \mu = 17072 \quad \mathrm{H}_\mathrm{a} : \mu \neq 17072$

   2. Conditions of CLT are satisfied as our sample is random independent sample, sample size is large (i.e. > 25), assume population is large.

   3. Compute $t-$statistic

$$SE_{EST} = \frac{s}{\sqrt{n}} = \frac{1203.5}{\sqrt{30}} = 219.72$$
$$t = \frac{\bar{x} - \mu}{SE_{EST}} = \frac{18104 - 17072}{219.72} = 4.696$$

   4. Use technology to find the p-value that corresponds with $t = 4.696$ and $df = 29$.

   5. We get $p-$value of 0.0001. Since $p-$value $<$ significance level i.e. $0.0001 < 0.05$, we will reject H_0

   6. Evidence suggests that, yes, mean attendance at hockey games has changed since 2010.

3. A new report claims that the mean daily time spent online in Egypt is 8.03 hours. Researchers believe that the mean daily time spent is actually much larger and pulls a random sample of 72 adults in Egypt. The sample had a mean daily time spent online of 8.14 hours with standard deviation of 0.67 hours. Complete one-sample $t-$test for population means at a significant level of 0.05.

   Firstly, population mean (μ) = 8.03, sample mean $\bar{x} = 8.14$, sample standard deviation (s) = 0.67, and sample size (n) = 72

   1. Hypothesize: $\mathrm{H}_0 : \mu = 8.03 \quad \mathrm{H}_\mathrm{a} : \mu > 8.03$

   2. Conditions of CLT are satisfied as our sample is random independent sample, sample size is large (i.e. > 25). We have large population.

   3. Compute $t-$statistic

$$SE_{EST} = \frac{s}{\sqrt{n}} = \frac{0.67}{\sqrt{72}} = 0.07896$$
$$t = \frac{\bar{x} - \mu}{SE_{EST}} = \frac{8.14 - 8.03}{0.07896} = 1.393$$

   4. Use technology to find the p-value that corresponds with $t = 1.393$ and $df = 71$.

5. We get $p-$value of 0.084. Since $p-$value $>$ significance level i.e. $0.084 > 0.05$, we fail to reject H_0

6. Evidence suggests that we fail to reject the null hypothesis at the 5% significance level.