CHAPTER 8

AP STATISTICS

STUDY GUIDE INCLUDES:

① READING ON KEY CONCEPTS (pp 196-204)
② PRACTICE TEST
MC (pp 205)
Short Answer (pp 208)
③ PRACTICE TEST ANSWERS (pp 209)

ESTIMATION USING A SINGLE SAMPLE

This review section will cover methods for estimating a single population parameter. Both point estimates (single numbers) and interval estimates (intervals of plausible values for our population parameter) will be considered.

OBJECTIVES

- Estimate a population mean or proportion using a point estimate.
- Construct and interpret a confidence interval estimate for a population proportion.
- Construct and interpret a confidence interval estimate for a population mean.
- Calculate the sample size needed in order to achieve a given bound on error.

POINT ESTIMATION

(*Introduction to Statistics & Data Analysis* 3rd ed. pages 476–480/4th ed. pages 530–534)

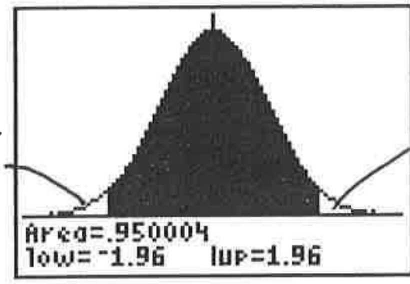
When we want to estimate some population characteristic, we will often use sample data to compute a single value, such as $\bar{x}$, the sample mean. This value is called a point estimate of μ, the population mean. Given a different sample, the sample mean would most likely be different—it is only an estimate of the value of μ. When estimating a population proportion, the usual point estimate is the sample proportion, $\hat{p}$. As an example, suppose a teacher wanted to estimate the proportion of student phone calls that are made to family members. Each student in a random sample of 100 students was asked whether the most recent call he or she made was to a family member.

In the case of proportions, $\hat{p}$, the sample proportion, is the sample statistic. If the sample size is large (check $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$), the sampling distribution of $\hat{p}$ is approximately normal. In this case, the critical value is found using the z distribution and depends on the desired confidence level. For a 95% confidence level, the critical value is found by finding the z value that corresponds to a central area o 0.95 under the z curve as shown below. For other confidence levels, adjust the central area.

To find $Z^* = 1.96$

invNorm(.025, 0, 1) = −1.96

.025



.025

For a 95% confidence level, the z critical value is 1.96. It follows that a 95% confidence interval for a population proportion is:

$$(\hat{p}) \pm (1.96)\begin{pmatrix}\text{standard deviation} \\ \text{of } \hat{p}\end{pmatrix}$$

Finally, as seen in the last review section, the estimated standard deviation of the sampling distribution of $\hat{p}$ is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, resulting in

$$\hat{p} \pm (1.96)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Returning to the example where 100 students were asked if the last phone call made was to a family member, a 95% confidence interval for the proportion of calls made to family members is

$$(0.29) \pm (1.96)\left(\sqrt{\frac{0.29(1-0.29)}{100}}\right) = (0.201, 0.379)$$

Note that the sample was a simple random sample and that the sample size is large enough $(n\hat{p} = 100(0.29) \geq 10, n(1-\hat{p}) = 100(0.71) \geq 10)$. These are the two conditions that must be checked in order for this confidence interval for a population proportion to be appropriate.

The 95% confidence interval is (0.201, 0.379). Interpreting the interval we would say that we are confident, in this case 95% confident, that the true proportion of students' phone calls that are to family members is between 0.201 and 0.379. Notice, instead of just having the point estimate of 0.29, we now have a range of reasonable values based on the sample data. This acknowledges the uncertainty in our original point estimate.

The sample statistic use is $\bar{x}$, the sample mean. When (1) the sample is a random sample and (2) either the sample size is large ($n \geq 30$) or the population distribution is approximately normal (these are the conditions that you need to check), the $\bar{x}$ sampling distribution is approximately normal with a standard deviation of $\frac{\sigma}{\sqrt{n}}$. There are two cases to consider depending on whether or not σ, the population standard deviation, is known.

If σ is known, a confidence interval for the population mean is

$$\bar{x} \pm (z \text{ critical value})\frac{\sigma}{\sqrt{n}}$$

The z critical value depends on the desired confidence level and is found in the same way as it was for the confidence interval for a population proportion. Because it is rare that σ is known, the more commonly used interval is the one that follows.

If σ is not known, a confidence interval for the population mean is

$$\bar{x} \pm (t \text{ critical value})\frac{s}{\sqrt{n}}$$

If the sample size is not large, we need to assume that the population distribution is approximately normal. This is reasonable when a dotplot or a boxplot constructed from the sample data is roughly symmetric and there are no outliers.

The t critical value in the interval above is found using a t distribution rather than the standard normal distribution. The t distributions have a mound shape, similar to the z distribution, but t distributions are more spread out than the z distribution. A t distribution is characterized by a quantity known as degrees of freedom (df). As df increases, the t distributions become more and more like the z distribution. For finding the critical value for a confidence interval for a population mean, df – $n - 1$.

The t critical value needed to compute the confidence interval can be found using the t distribution table (also provided during the AP Statistics exam) or generated on your calculator. To use the table provided by the College Board® on the day of the exam, find the degrees of freedom in the left-most column. Remember, df = $n - 1$. Then follow this row over to the column that contains the critical values for the desired confidence level. The partial *t table below* shows t critical values for df 8, 9, and 10. This is from the actual table you will be provided with on exam day. Keep in mind, the tail probability is ½ of the area left *outside* the interval you are seeking.

EXAMPLE If we want the t critical value for a 95% confidence interval and the sample size is 10, we would go down the first column to locate df = $n - 1$ = 9 and follow that row over to the column for a 95% confidence level.

EXAMPLE A cable TV company is interested in estimating the average time that customers must wait on hold when they call the company's customer support line. Twenty calls are selected at random from the calls made to the customer support line and the hold time is recorded for each call. The sample mean hold time was 10.5 minutes and the sample standard deviation was 1.1 minutes. A dotplot of the 20 hold times was approximately symmetric and there were no outliers. Estimate the mean hold time for customers calling the support line using a 90% confidence interval.

Because the population standard deviation is not known, we will consider using the t confidence interval. First we must verify that the conditions are met. The problem states that the sample is a random sample. The sample size is not large ($20 < 30$), so we must be willing to assume that the hold time population distribution is approximately normal. The actual sample data is not given here (if it were, we would construct a dotplot or a boxplot), but we are told that the dotplot was approximately symmetric and that there were no outliers.

With a sample size of $n = 20$, df = 19. For a 90% confidence level, the t critical value is 1.729. This gives

$$\bar{x} \pm (t^* \text{ critical value})\left(\frac{s}{\sqrt{n}}\right)$$
$$= 10.5 \pm (1.729)\left(\frac{1.1}{\sqrt{20}}\right)$$
$$= 10.5 \pm (1.729)(0.246)$$
$$= 10.5 \pm 0.426$$
$$= (10.08, 10.93)$$

We can now say that, based on this sample, we are 90% confident that the true mean hold time is between 10.08 and 10.93 minutes. The 90% confidence level means that we used a method that will capture the true mean wait time about 90% of the time in repeated sampling.

AP Tip

Prior to calculating a confidence interval of any type, be sure you verify that the necessary conditions are met. In the case of a t-interval, we must verify (and state for the AP Readers) that

- The sample was a simple random sample.
- The sample size is large or the population distribution is approximately normal. To decide if it is reasonable to think that the population distribution is approximately normal, look at a dotplot or ~~a Boxplot~~. You want the plot to be reasonably symmetric with no outliers. (You could also look at a normal probability plot—if the plot looks linear, normality is plausible. NOT NEEDED)

BETTER TO LOOK AT A HISTOGRAM TO SEE SHAPE AND IDENTIFY OUTLIERS AND SKEWNESS
BUT remember for means
CLT says distribution is approximately normal if large enough ($n \geq 30$).

As a final note, consider what happens if instead of using a 90% confidence level, we change to a 95% confidence level. Using the same information from the example above, the only change in the

The method used to construct the interval will capture the true value about 90% of the time in repeated sampling.

** **AP Tip** VERY IMPORTANT TO DO! CORRECTLY

To interpret a confidence interval, say something like

I am ____% confident that the true value of the population ________ (proportion or mean) ______ (supply context here!) is contained within the interval (_____, _____).

To interpret the confidence level, say something like

The method that was used to construct the interval will capture the true _________ (proportion or mean) about ______% of the time in repeated sampling.

SAMPLE PROBLEM 1 A real estate company in Colorado has noticed the housing market seems to be doing better in some markets than others. In order to help new home sellers set a price, they selected a random sample of 20 recent sales in Colorado and found the average selling price was \$142,000 and the standard deviation was \$4,000.

(a) What conditions would need to be met in order to generate a confidence interval?

(b) Assuming the conditions are met, calculate a 95% confidence interval for the mean price in Colorado.

(c) Interpret the confidence interval.

(d) Interpret the confidence level of 95%.

SOLUTION TO PROBLEM 1

(a) Since the standard deviation is from the sample, a *t*-interval would be used. The conditions for the *t* confidence interval are that the sample is a simple random sample and that the sample size is large or the population distribution is approximately normal. We would need to assume these conditions are met to construct this confidence interval.

(b) $142{,}000 \pm 1.729\left(\frac{4{,}000}{\sqrt{20}}\right)$, with df = 19.

(c) I am 95% confident that the true mean selling price of homes in Colorado is contained in the interval (\$141,128 – \$143,872).

(d) The method I have used to generate the interval will capture the true mean housing price in Colorado about 95% of the time in repeated sampling.

4. A researcher wants to find a confidence interval estimate of the mean time it takes to complete an order over the phone at a call center for a large retail company. He has selected a random sample of 18 orders and recorded the time to complete the order. The sample mean time was 4.27 minutes and the sample standard deviation was 0.78 minutes. Assuming the conditions needed for inference are met, which of the following is the appropriate 98% confidence interval?

(A) $4.27 \pm 2.326\left(\frac{0.78}{\sqrt{18}}\right)$

(B) $4.27 \pm 2.054\left(\frac{0.78}{\sqrt{17}}\right)$

(C) $4.27 \pm 2.552\left(\frac{0.78}{\sqrt{18}}\right)$

(D) $4.27 \pm 2.567\left(\frac{0.78}{\sqrt{18}}\right)$

(E) $4.27 \pm 2.567\left(\frac{0.78}{\sqrt{17}}\right)$

5. What size is needed to achieve a margin of error of 0.03 with 95% confidence when the population proportion of successes is 0.36?
(A) 984
(B) 983
(C) 256
(D) 255
(E) 250

6. Which of the following would not decrease the width of a confidence interval?
I. Increasing the sample size
II. Decreasing the degrees of freedom
III. Decreasing the confidence level
(A) I only
(B) II only
(C) III only
(D) I and II
(E) II and III

7. In a random sample of 64 children who received a new antibiotic for an infection, 52 of the children had positive results within 12 hours. Find a 95% confidence interval for the proportion of children with infections who will experience positive results within 12 hours when treated with this antibiotic.
(A) 0.81 ± 0.096
(B) 0.81 ± 0.080
(C) 0.81 ± 0.071
(D) 0.83 ± 0.079
(E) 0.83 ± 0.094

12. You have been asked to compute a 96% confidence interval for a population mean. If the population standard deviation is known to be 7 and the sample size is 40, what critical value would be used in computing the interval?
    (A) 1.751
    (B) 1.960
    (C) 2.054
    (D) 2.122
    (E) 2.125

13. A random sample of size 150 resulted in a sample proportion of 0.45. What is the approximate standard error of $\hat{p}$ for a sample of size 150?
    (A) 0.03
    (B) 0.24
    (C) 0.02
    (D) 0.04
    (E) 0.002

14. A large company wants to estimate the proportion of employees who would prefer a pay increase to an increase in retirement benefits using a 95% confidence interval. What sample size should be used in order to achieve a margin of error of 0.04?
    (A) 156 employees
    (B) 157 employees
    (C) 307 employees
    (D) 600 employees
    (E) 601 employees

15. Based on a random sample of 1,000 adult Americans, a consumer group states that 72% of adult Americans believe corporations are not concerned about public safety. They also reported a margin of error of 2 percentage points with 90% confidence. What does this mean?
    (A) If the poll were conducted again, the probability that 72% believe corporations are not concerned about safety is 0.90.
    (B) The probability that the proportion of adult Americans who believe corporations aren't concerned with public safety is between 70% and 74% is 0.98.
    (C) Between 88% and 92% of adult Americans believe corporations are not concerned with public safety.
    (D) About 90% of all random samples of 1000 adult Americans will result in a sample percentage that is within 2 percentage points of the actual population proportion.
    (E) Ninety out of every 100 samples of 1,000 adult Americans will have between 70% and 74% who believe that corporations are not concerned with public safety.

FREE-RESPONSE PROBLEMS (PRACTICE SHORT ANSWER)

1. Many of the trees in a national forest suffer from a virus that attacks the bark of the tree. Trees with this virus should be removed in order to minimize the risk to nearby trees. To estimate the proportion of trees that have this virus, a random sample of

3. **D.** A large sample size is required. For proportions, this means that we need to have $np \geq 10$ and $n(1-p) \geq 10$ (*Introduction to Statistics & Data Analysis* 3rd ed. pages 482–492/4th ed. pages 536–545).

4. **D.** Since the standard deviation is from the sample, this is a *t*-interval and will require a *t* critical value be based on 17 df (*Introduction to Statistics & Data Analysis* 3rd ed. pages 494–505/4th ed. pages 549–559).

5. **A.** The calculation would be done in the following manner

$$0.03 = 1.96\left(\sqrt{\frac{0.36(0.64)}{n}}\right);\ 0.0009 \approx 3.842\frac{0.2304}{n};\ n = 3.842(256);$$

$n = 984$

(*Introduction to Statistics & Data Analysis* 3rd ed. pages 482–492/4th ed. pages 536–545).

6. **B.** AS smaller df corresponds to a smaller sample size and a larger *t* critical value, making the interval wider (*Introduction to Statistics & Data Analysis* 3rd ed. pages 494–505/4th ed. pages 549–559).

7. **A.** $\frac{52}{64} \pm 1.96\sqrt{\frac{0.81(0.19)}{64}}$.

(*Introduction to Statistics & Data Analysis* 3rd ed. pages 482–492/4th ed. pages 536–545).

8. **C.** The confidence interval for 90% is 11.2 ± 0.863 (*Introduction to Statistics & Data Analysis* 3rd ed. pages 494–505/4th ed. pages 545–559).

9. **D.** The population standard deviation is not used in the computation of a *t*-interval. However, since the sample size is so small, both I and III will be needed for the interval to be appropriate (*Introduction to Statistics & Data Analysis* 3rd ed. pages 494–505/4th ed. pages 545–559).

10. **D.** WE have used a method that produces an interval that contains the true mean value about 98% of the time in repeated sampling (*Introduction to Statistics & Data Analysis* 3rd ed. pages 508–513/4th ed. pages 563–567).

11. **B.** Since the sample size is found in the denominator, any increase in this value will make the width of the confidence interval smaller. However doubling the sample size does not result in an interval that is half as wide (*Introduction to Statistics & Data Analysis* 3rd ed. pages 494–505/4th ed. pages 545–559).

12. **C.** A z critical value would be used. C is the *correct* z critical value for a confidence level of 96% (*Introduction to Statistics & Data Analysis* 3rd ed. pages 494–505/4th ed. pages 545–559).

The boxplot is approximately symmetric with no outliers and the normal probability plot is realtively straight, so it is not unreasonable to think that the population distribution is approximately normal.

The confidence interval is

$$72.4 \pm 1.701 \frac{10.77}{\sqrt{28}} = 72.4 \pm 4.1 = (68.3, 76.6)$$

(b) Based on the interval, we are confident that the mean time to complete the multiple-choice section is between 68.3 minutes and 76.6 minutes. Even though we expect about half of all students to take longer than the mean time (because we think that the population distribution of times is approximately normal), 90 minutes is quite a bit above the upper end of the confidence interval. So, 90 minutes seems like a reasonable amount of time (*Introduction to Statistics & Data Analysis* 3rd ed. pages 494–505/4th ed. pages 549–559).