Sections 8.1 and 8.2: Confidence Intervals for Population Means with a known Standard Deviation

1 Point Estimates

In the last section we discussed using a sample to make an inference about a population value. We infer that the average in a properly collected random sample is the population average as well. Do we really expect the sample average to be the exact population average? No. We expect it to be close to the population average. How close? Furthermore, how confident are we that an outlier didn't sneak into the sample by random chance and dramatically skew the results?

Definition 1 *A **point estimate** is a statistic from a sample that is used to estimate a parameter for a population. For example we use $\bar{x}$ (median, s, Q_3, etc.) from a sample to estimate μ (median, σ, Q_3, etc.) for a population.*

Example 1 *A random sample of 140 KSU students finds that 37 of those polled live on campus. Compute the point estimate for the proportion of KSU students who live on campus.*
The point estimate for the proportion is $\frac{37}{140} = 0.264\,29$.

2 Estimating a population mean from a sample

Definition 2 *A **confidence interval** for a parameter is an interval of numbers within which we expect the true value of the population parameter to be contained. The endpoints of the interval are computed based on sample information.*

Let's say we want to estimate the true average GPA on campus. We collect a sample and find that the sample average is 3.0. Thus, we reasonably believe that the true average is close to 3.0 and use it as our point estimate. A confidence interval can be constructed around the sample average to indicate how close we expect the sample average and population average to be. Some confidence intervals for the true population average are:

$$(2.95, 3.05)$$
$$(2.9, 3.1)$$
$$(2.0, 4.0)$$
$$(2.9999, 3.0001)$$

A struggle between precision and certainty exists in every confidence interval. To be more confident, we wind up being less precise. To be more precise, we wind up being less confident. Because of this, every confidence interval is a

balance between certainty and precision. The tension between certainty and precision is always there. Fortunately, there are certain standards to use so that we can all be both sufficiently certain and sufficiently precise to make useful statements. Most all confidence intervals are computed using one of four levels of confidence: 90%, 95%, 98% or 99%. The level of confidence indicates the probability that the true population average is contained in the confidence interval.

Due to the Central Limit Theorem, we know that the sampling distribution of the mean is approximately normal. Thus, by the 68-95-99.7% rule, we know that about 68% of all samples will yield a sample average that falls within 1 standard error of the true population average. So, about 95% of all samples will yield a sample average that falls within 2 standard errors of the true population average. And finally, about 99.7% of all samples will yield a sample average that falls within 3 standard errors of the true population average.

How do we compute the endpoints of our confidence interval? For large ($n \geq 30$), random samples, we compute the lower and upper bounds using

$$\bar{x} \pm \text{ margin of error.}$$

The margin of error depends on the level of confidence and standard error. For every level of confidence, the error rate α, is computed as 100%-level of confidence. So, for a 95% confidence level, $\alpha = 5\% = .05$. How do we interpret α? If there is a 95% chance the our confidence interval contains the true population mean then there is a 5% chance that our confidence interval does not contain the true population mean. Every confidence level is associated with a critical value $z_{\frac{\alpha}{2}}$.

level of confidence	90%	95%	98%	99%
α	$10\% = 0.1$	$5\% = 0.05$	$2\% = 0.02$	$1\% = 0.01$
$z_{\frac{\alpha}{2}}$	1.645	1.96	2.33	2.575

Example 2 *How is $z_{\frac{\alpha}{2}}$ computed? In the standard normal curve what two values contain the central 90% of the data? That would be P_5 and P_{95}.*

Example 3 *Find $z_{\frac{\alpha}{2}}$ for an 80% confidence interval. Since $\alpha = .2$, we want $z_{.10}$. The endpoints that contains the central 80% of the standard normal curve are P_{10} and P_{90}. So $z_{.10} = 1.28$.*

Exercise 1 *Find $z_{\frac{\alpha}{2}}$ for an 70% confidence interval.*

The margin of error for a confidence interval is the product of the critical value and standard error (where the standard deviation of the population is known), $z^* \times \sigma_{\bar{x}}$. The final form for our confidence interval is

$$\bar{x} \pm \text{ margin of error } = \bar{x} \pm z_{\frac{\alpha}{2}} \times \sigma_{\bar{x}} = \bar{x} \pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}.$$

Remark 1 *A confidence interval for the mean with known population standard deviation is computed on the TI-83/84 using the* ***ZInterval*** *command from statistical tests.*

Example 4 *A sample of 100 observations is collected and yields* $\bar{x} = 75$ *and from a population with known* $\sigma = 8$*. Find a 95% confidence interval for the true population average. The endpoints are computed from* $75 \pm 1.96\left(\frac{8}{\sqrt{100}}\right)$*. The lower bound is* $75 - 1.96\left(\frac{8}{\sqrt{100}}\right) = 73.432$ *and the upper bound is* $75 + 1.96\left(\frac{8}{\sqrt{100}}\right) = 76.568$*. So, we are 95% confident that the true population average falls somewhere in the interval* $(73.432, 76.568)$*.*

Remark 2 *R Studio does not have a simple one step commend for computing a confidence interval from sample parameters. R Studio does have a simple command to construct a confidence interval from a file of data.*

```
> # R program to find the confidence interval
>
> # Input the mean of the sample data
> mean_value <- 75
>
> # Input the size of the sample
> n <- 100
>
> # Input the standard deviation
> standard_deviation <- 8
>
> # Find the standard error
> standard_error <- standard_deviation / sqrt(n)
> alpha = 0.05
> degrees_of_freedom = n - 1
> t_score = qt(p=alpha/2, df=degrees_of_freedom,lower.tail=F)
> margin_error <- t_score * standard_error
>
> # Calculating lower bound and upper bound
> lower_bound <- mean_value - margin_error
> upper_bound <- mean_value + margin_error
>
> # Print the confidence interval
> print(c(lower_bound,upper_bound))
[1] 73.41263 76.58737
```

Example 5 *A sample of 75 KSU students shows that students watch an average of 20 hours of Braves baseball a month during the baseball season with a known population standard deviation of 6.5 hours. Construct a 90% confidence interval for the true average number of hours per month that KSU students watch Braves baseball. The endpoints are computed from* $20 \pm 1.645\left(\frac{6.5}{\sqrt{75}}\right)$. *The lower bound is* $20 - 1.645\left(\frac{6.5}{\sqrt{75}}\right) = 18.765$ *and the upper bound is* $20 + 1.645\left(\frac{6.5}{\sqrt{75}}\right) = 21.235$.

```
> # R program to find the confidence interval
>
> # Input the mean of the sample data
> mean_value <- 20
>
> # Input the size of the sample
> n <- 75
>
> # Input the standard deviation
> standard_deviation <- 6.5
>
> # Find the standard error
> standard_error <- standard_deviation / sqrt(n)
> alpha = 0.10
> degrees_of_freedom = n - 1
> t_score = qt(p=alpha/2, df=degrees_of_freedom,lower.tail=F)
> margin_error <- t_score * standard_error
>
> # Calculating lower bound and upper bound
> lower_bound <- mean_value - margin_error
> upper_bound <- mean_value + margin_error
>
> # Print the confidence interval
> print(c(lower_bound,upper_bound))
[1] 18.74979 21.25021
>
```

Exercise 2 *Seventy KSU students individually solve the same Sudoku puzzle. The average solving time is 15.3 minutes with a known population standard deviation of 8 minutes. Construct a 95% confidence interval for the true population average time to solve this Sudoku puzzle. Interpret your confidence interval.*

Exercise 3 *Alex collects a sample of 175 KSU students from the Library. This sample yields the 95% confidence interval (31,37) for the true average number of hours all KSU students study per week. Do you have faith in Alex's estimate? Explain.*

3 Finding a necessary sample size

Example 6 *Let's revisit the 95% confidence interval constructed from a sample of 100 observations which yielded* $\bar{x} = 75$ *and* $\sigma = 8$*. The ME is* $1.96\left(\frac{8}{\sqrt{100}}\right) = 1.568$*. Perhaps we need the margin of error to be less than 1.5 but don't want to lower the level of confidence. We can decrease the ME by increasing the sample size. How large of a sampled to collect in order to lower the ME to 1.5?*

$$\begin{aligned} 1.96\left(\frac{8}{\sqrt{n}}\right) &\leq 1.5 \\ 1.96 * 8 &\leq 1.5\sqrt{n} \\ \frac{1.96 * 8}{1.5} &\leq \sqrt{n} \\ \left(\frac{1.96 * 8}{1.5}\right)^2 &\leq n \\ 109.27 &\leq n \end{aligned}$$

So we need to increase the sample size to at least $n = 110$ observations.

In general, the sample size needed to bound the ME by m with a fixed level of confidence is

$$n \geq \left(\frac{z_{\frac{\alpha}{2}}\sigma}{m}\right)^2.$$

Exercise 4 *At 95% confidence with* $\sigma = 15$*, how large of a sample is needed to bound the ME below by 4?*

$$n \geq \left(\frac{z_{\frac{\alpha}{2}}\sigma}{m}\right)^2 = \left(\frac{1.96 * 15}{4}\right)^2 = 54.023$$

4 Exercises

1. Kokoska 3rd edition Section 8.1: 8.12, 8.13a, b, c,
2. Kokoska 3rd edition Section 8.2: 8.21, 8.25, 8.27, 8.29, 8.31, 8.32, 8.33, 8.35, 8.37, 8.42