Module D: Introductory Inference

Slide Deck D1:

Theory of Confidence Intervals

The section in which we are introduced to the 'confidence interval,' sets of values (intervals) used to estimate population ***parameters****. As we move forward, we shall see the importance and limitation of this estimation.*

Today's Objectives

By the end of this slidedeck, you should

- compare and contrast three types of intervals
  - observation intervals
  - prediction intervals
  - confidence intervals
- state the theory behind calculating each
- understand what each represents

Observation Intervals

Definition (Observation Interval)

An **observation interval** is a set of values that contain the observed values a given proportion of the time.

That is: An observation interval describes your data.

Prediction Intervals

Definition (Prediction Interval)

A **prediction interval** is a set of values that contain a future observation a given proportion of the time.

That is: A prediction interval predicts a future observation.

Confidence Intervals

Definition (Confidence Interval)

A **confidence interval** is a set of values that theoretically contain the population parameter a given proportion of the time when the experiment is performed many times.

That is: A confidence interval estimates a parameter of the population (usually the mean, the proportion/success probability, or the variance: μ, p, or σ^2).

A Quick Example

Example

One of the variables in the `crime` data file is the percent of children enrolled in school in 2000 (`enroll00`). Calculate the three intervals for this variable. Use a 95% level for each.

Observation Interval:

```
interval(enroll00, type="observation", level=0.95)
```

This tells us that the 95% observation interval is from 85.575 to 97.700. In other words, about 95% of **these observations** are between 85.575 and 97.700, inclusive.

A Quick Example

Example

One of the variables in the `crime` data file is the percent of children enrolled in school in 2000 (`enroll00`). Calculate the three intervals for this variable. Use a 95% level for each.

Prediction Interval:

```
interval(enroll00, type="prediction", level=0.95)
```

This tells us that the 95% prediction interval is from 82.559 to 100.936. In other words, about 95% of the time the next state (**future observation**) will have an enrollment between 82.559 and 100.936, inclusive.

A Quick Example

Example

One of the variables in the `crime` data file is the percent of children enrolled in school in 2000 (`enroll00`). Calculate the three intervals for this variable. Use a 95% level for each.

Confidence Interval:

```
interval(enroll00, type="confidence", level=0.95)
```

This tells us that the 95% observation interval is from 90.473 to 93.021. In other words, we are 95% confident that the **population mean** is between 90.473 and 93.021, inclusive.

An Important Note

Note

The confidence (and prediction) intervals have requirements that must be met before was can calculate them. We will see those important requirements in the future. The above examples merely serve to illustrate the three types of intervals and what they mean.

Let us now focus on confidence intervals for the rest of this lecture, because it is the most used of the three, and the sciences require understanding them.

Definitions

Definition (Confidence Interval)

A **confidence interval** is a set of values that theoretically contain the population parameter a given proportion of the time when the experiment is performed many times.

That is: If (lcl, ucl) is a $(1-\alpha)\times 100\%$ confidence interval for the parameter, then $(1-\alpha)\times 100\%$ of the intervals — under repeated experiments — contain the parameter.

Example: If $(4,\ 9)$ is a 95% confidence interval for the mean, then 95% of the intervals — under repeated experiments — contain the mean.

Definitions

Note that the confidence interval

- gives some information about the population parameter
- is a set of *reasonable* values for that parameter
- is a function of the data (and thus a random variable)
- is a result of a probability distribution calculation

Note that the confidence interval does *not*

- provide a probability statement on the parameter

That is, one **cannot state** something like
"The *probability* that the parameter is between 4 and 9 is 95%.

Definitions

Definition (Confidence Level)

The **confidence level** is the theoretical frequency (or proportion) of possible confidence intervals that contain the true value of their corresponding parameter.

- The value of the confidence level is $c = 1 - \alpha$.
- By default, we will use $\alpha = 0.05$ in this course.
- The value α will be re-introduced when we cover hypothesis testing. It will represent the theoretical Type I error rate. It is selected by the researcher to reflect tradition in the discipline and the particularities of the current research.

Bootstrapping

Previously, we introduced bootstrapping as a way of estimating — and obtaining — a confidence interval for a population parameter. Let's look at the code:

```
dt = read.csv("http://rfs.kvasaheim.com/data/geography.csv")
attach(dt)

st = numeric()
for(i in 1:1e4) {
  x = sample(Score, replace=TRUE)
  st[i] = mean(x)
}

quantile(st, c(0.025,0.975))
```


Bootstrapping

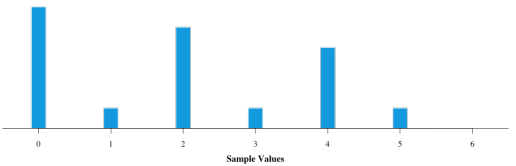
The output from this code is

```
     2.5%     97.5%
1.633333 2.566667
```

Thus, we are 95% confident that the population mean, μ, average understanding of geography, is between 1.63 and 2.57 (out of 6).

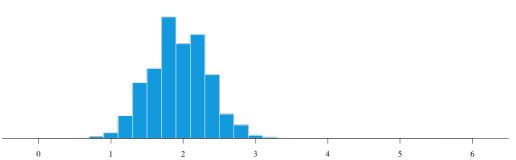
Bootstrapping

One thing I want to point out here is the difference between the distribution of the sample **values** and the distribution of the sample means:



Bootstrapping

One thing I want to point out here is the difference between the distribution of the sample values and the distribution of the sample **means**:



The Z Theory

This is fine if all we have is the data. However, we know that sample means tend toward Normality as the sample size increases (CLT). Thus, we may be able to create confidence intervals *without* bootstrapping and *with* using **additional information**.

Recall that for large n,

$$\overline{X} \sim \mathcal{N}\left(\mu,\ \frac{\sigma^2}{n}\right)$$

Applying the z-transform (i.e., standardizing formula, z-score formula) to $\overline{X}$ gives

$$\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,\ 1)$$

The Z Theory

$$\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0, 1)$$

This means the expression on the left is between -1.96 and 1.96 approximately 95% of the time.

Why? What are the 0.025 and 0.975 quantiles of the standard Normal distribution? Yeppers: -1.96 and 1.96.

Thus, a theoretical 95% interval for $\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$ is between -1.96 and 1.96, by definition.

The Z Theory

Since a theoretical 95% interval for $\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$ is between -1.96 and 1.96, we can solve for μ to get a 95% confidence interval for μ, the population parameter:
One bound of a 95% confidence interval for μ is

$$-1.96 = \frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$$

$$-1.96\sqrt{\sigma^2/n} = \overline{X}-\mu$$

$$-\overline{X}-1.96\sqrt{\sigma^2/n} = -\mu$$

$$\overline{X}+1.96\sqrt{\sigma^2/n} = \mu$$

The Z Theory

I leave it as an exercise for you to prove that the other bound is

$$\mu = \overline{X} - 1.96\sqrt{\sigma^2/n}$$

Thus, the endpoints of a confidence interval for μ are

$$\overline{X} \pm Z_{\alpha/2}\sqrt{\sigma^2/n}$$

Here, $Z_{\alpha/2}$ is the absolute value of the $\alpha/2$ quantile for the Z distribution. If $\alpha = 5\%$ (as usual), then $Z_{\alpha/2} = 1.96$.

The Z Theory

This confidence interval for μ works if

- You are estimating μ
- You are able to calculate $\bar{x}$ and n
- You know σ^2
- The data are generated from a Normal process (or n is sufficiently large)

The Z Theory

Recall the definition of the population variance. To calculate it, you need to know μ. However, since we are estimating μ in this procedure, we *rarely* know σ^2.

- Since we cannot use the z-procedure, we need to use a different procedure.

Important!!

This starts you on the journey to understand that statistical procedures have requirements to them. You need to check that the requirements are reasonable before you are able to use those procedures.

The t Theory

So, if we do not know the population variance (as usual), what can we do? Thanks to William Sealy Gosset and Karl Pearson (1908), it *can* be shown that the following is true:

$$\frac{\overline{X} - \mu}{\sqrt{s^2/n}} \sim t_\nu$$

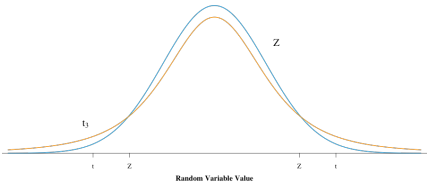
where $\nu = n - 1$ (in this one-sample case). From this, we have the following confidence interval

$$\overline{X} \pm T_{\alpha/2,\nu} \sqrt{s^2/n}$$

Here, $T_{\alpha/2,\nu}$ is the $\alpha/2$ quantile for the t distribution with $\nu = n - 1$ degrees of freedom. For instance, if $\alpha = 5\%$ and $n = 4$, then $T_{\alpha/2,\nu} = t_{\alpha/2,n-1} = 3.182$.

The t Theory

To illustrate the difference between the Z and the t distributions, here is the t_3 distribution — as compared to the Z distribution:



The t Theory

This confidence interval for μ works if

- You are estimating μ
- You are able to calculate $\bar{x}$, n, and s
- The data are generated from a Normal process (or n is "large enough")

Important!!

This continues your journey understanding that statistical procedures have *requirements*. In creating these procedures, statisticians made assumptions. In using these procedures, the assumptions became requirements.

You need to check that the requirements are reasonable before you are ethically able to use those procedures.

Other Procedures

Note the process we followed to obtain the endpoints of the confidence intervals:

1. Create a test statistic
2. Determine the distribution of that statistic
3. Determine the quantiles of that distribution
4. Solve the test statistic for the parameter of interest

Behind the scenes:
This process can/will be repeated for other parameters to create appropriate analysis procedures.

In fact, this procedure will be used many times in MATH 322: Mathematical Statistics, II. Apparently, this is one of the easiest courses in the department. Because of this, it is also one of the most satisfying for students... a crowning achievement!

Other Procedures

Other useful procedures you will need to know for one-population procedures

μ:	$\overline{X}$ from Normal distribution	t-procedure
	$\overline{X}$ from symmetric and continuous distribution	Wilcoxon procedure
	otherwise	non-parametric bootstrap
$\widetilde{\mu}$:	$\overline{X}$ from continuous distribution	Wilcoxon procedure
p:	X from Binomial	Binomial test
σ^2:	$\overline{X}$ from Normal distribution	Chi-Square test
	otherwise	non-parametric bootstrap

For the R functions, please see `allProcedures.pdf` handout.

Today's Objectives

In today's slide deck, we covered

- compare and contrast three types of intervals
  - observation intervals
  - prediction intervals
  - confidence intervals
- state the theory behind calculating each
- understand what each represents

Today's R Functions

In this slide deck, we covered three R functions. This is in addition to one we have already experienced and ones we will experience:

- `interval(x, type, level="observation")`
- `interval(x, type, level="prediction")`
- `interval(x, type, level="confidence")`

The `interval` function provides intervals based on observed data.

The first version has no requirements. The other two require Normality in the population. In the future, we will explore how to test Normality of the population and how to calculate confidence intervals more generally.

Supplemental Activities

The following are supplements for the topics covered today.

- SCA 8a
- SCA 8b

Note that you can access all Statistical Computing Activities here:
`https://www.kvasaheim.com/courses/stat200/sca/`

In addition to these SCAs, there is also **Laboratory Activity E**.
`https://www.kvasaheim.com/courses/stat200/labs/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding confidence intervals and what they actually tell us about the population:

- Hawkes Learning: Chapters 8 (and 9)
- Intro to Modern Statistics: Chapters 11–15
- R for Starters: Chapters 5 (and 6)

- Wikipedia: Central Limit Theorem
  Confidence Intervals