Module D: Introductory Inference

Slide Deck D2:

Practice of Confidence Intervals

The section in which we learn how to calculate confidence intervals using the R statistical environment. This will allow one to estimate a population parameter. Such parameters include the mean, the success probability, and the variance.

Today's Objectives

By the end of this slidedeck, you should

1. state what a confidence interval concerns
2. calculate confidence intervals using R
3. understand the analysis procedure for
   - the value of one population mean
   - the value of one population proportion
   - the value of one population variance
4. understand the analysis procedure for
   - the difference between two population means (independent)
   - the difference between two population means (dependent)
   - the difference between two population proportions
   - the ratio of two population variances

Analysis Procedure Overview

All of the analysis procedures follow the same process (an abstraction):

1. Determine the parameter(s) to estimate
2. Determine the number of populations
3. Determine the most powerful test available
   - Determine the requirements of that procedure
   - Test the requirements
   - If no requirement is violated, use that procedure
   - If any requirement is violated, check the next procedure
4. Repeat Step 3 until you obtain a procedure whose requirements are met.

Note: There are many, many, many more procedures than what we cover in this course.

One-Parameter Procedures: μ

Parametric Procedure: Student's t-procedure

- Graphic: box-and-whiskers plot
  `boxplot(x)`
- Requires: Data generated from Normal distribution
  - Requirement test: Shapiro-Wilk test
  - `shapiroTest(x)`
- `R` function: `t.test(x)`

One-Parameter Procedures: μ (and $\tilde{\mu}$)

Non-parametric Procedure I: Wilcoxon procedure

- Graphic: box-and-whiskers plot
  `boxplot(x)`
- Requires: Data generated from symmetric distribution
  - Requirement test: Hildebrand Rule
  - `hildebrand.rule(x)`
- R function: `wilcox.test(x, conf.int=TRUE)`

One-Parameter Procedures: μ

Non-parametric Procedure II: non-parametric bootstrap

- Graphic: box-and-whiskers plot
  `boxplot(x)`
- Requires: Nothing
- R code:

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

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

One-Parameter Procedures: p

Parametric Procedure: Binomial procedure

- Graphic: Binomial plot
  `binom.plot(x, n)`
- Requires: Data generated from Binomial distribution
- R function: `binom.test(x, n)`

Note: This is *not* the procedure Hawkes covers. They use something close to what we refer to as the Wald test. The function to perform the Wald test in R is `wald.test`.

One-Parameter Procedures: p

Non-parametric Procedure: Non-Parametric Bootstrap procedure

- Graphic: box-and-whiskers plot
  `binom.plot(x, n)`
- Requires: Nothing
- R code:

```
st = numeric()
for(i in 1:1e4) {
  x = sample(y, replace=TRUE)
  st[i] = mean(x)
}
quantile(st, c(0.025,0.975))
```

This code assumes the data, `y`, consist of 0 and 1 values.

One-Parameter Procedures: σ^2

Parametric Procedure: Chi-Square Procedure

- Graphic: box-and-whiskers plot
  `boxplot(x)`
- Requires: Data generated from Normal process
  - Requirement test: Shapiro-Wilk test
  - `shapiroTest(x)`
- `R` function: `onevar.test(x)`

One-Parameter Procedures: σ^2

Non-parametric Procedure: Non-Parametric Bootstrap procedure

- Graphic: box-and-whiskers plot
  `boxplot(x)`
- Requires: Nothing
- `R` code:

```
st = numeric()
for(i in 1:1e4) {
  x = sample(y, replace=TRUE)
  st[i] = var(x)
}
quantile(st, c(0.025,0.975))
```

Two-Parameter Procedures: $\mu_1 - \mu_2$

Parametric Procedure: two-sample t-procedure

- Graphic: side-by-side box-and-whiskers plot
  `boxplot(x ~ g)`
  `boxplot(x1,x2)`
- Requires: Data generated from Normal distribution — in *each* population
  - Requirement test: Shapiro-Wilk test
  - `shapiroTest(x ~ g)`
- R function: `t.test(x ~ g)`
- R function: `t.test(x1,x2)`

Note: This is not the same test as Hawkes uses. As such, you should not use this to do your HLS homework.

Two-Parameter Procedures: $\mu_1 - \mu_2$

Non-Parametric Procedure: Mann-Whitney procedure

- Graphic: side-by-side box-and-whiskers plot
  `boxplot(x ~ g)`
  `boxplot(x1,x2)`
- Requires: Beyond the Scope of this Course (BSC)
- R function: `wilcox.test(x ~ g, conf.int=TRUE)`
- R function: `wilcox.test(x1,x2, conf.int=TRUE)`

Two-Parameter Procedures: μ_{1-2}

Parametric Procedure: two-sample dependent t-procedure

- This is just a one-sample procedure if you focus on the differences instead of the measurements.
- This should help to illustrate the need to understand the data and what you want to understand.

Note: The dependent samples means procedure is just a one-sample means procedure in which you are using the *differences* between the data values. As such, refer to the one-sample mean procedure.

Two-Parameter Procedures: $p_1 - p_2$

Parametric Procedure: Proportions Procedure

- Graphic: Binomial plot
  `binom.plot(x=c(x1,x2), n=c(n1,n2))`
- Requires: Expected number of successes is at least 5 in each group
- `R` function: `prop.test(x=c(x1,x2), n=c(n1,n2))`

Note: This is *not* the procedure Hawkes covers. They use something close to this, but this procedure makes adjustments for the fact that the Binomial distribution is discrete and the Normal distribution is not. As such, you should use the Wald test for this type of Hawkes problem, too. `wald.test`

Two-Parameter Procedures: σ_1^2/σ_2^2

Parametric Procedure: Fisher's F-test

- Graphic: Side-by-side box-and-whiskers plot
  `boxplot(x ~ g)`
  `boxplot(x1, x2)`
- Requires: Data generated from Normal distribution — in *each* population
  - Requirement test: Shapiro-Wilk test
  - `shapiroTest(x ~ g)`
- R function: `var.test(x ~ g)`
- R function: `var.test(x1, x2)`

Today's Objectives

Now that we have concluded this lecture, you should

1. state what a confidence interval concerns
2. calculate confidence intervals using R
3. understand the analysis procedure for
   - the value of one population mean
   - the value of one population proportion
   - the value of one population variance
4. understand the analysis procedure for
   - the difference between two population means (independent)
   - the difference between two population means (dependent)
   - the difference between two population proportions
   - the ratio of two population variances

In Other Words:

In today's slide deck, we covered procedures for estimating:

- μ
- $\mu_1 - \mu_2$
- μ_{1-2}
- p
- $p_1 - p_2$
- σ^2
- σ_1^2/σ_2^2

We also looked at testing the Normality and the symmetry requirements:

- Shapiro-Wilk test
- Hildebrand Rule

Today's R Functions

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

- `boxplot`
- `binom.plot`
- `shapiro.Test`
- `hildebrand.rule`

- `t.test`
- `wilcox.test`
- `binom.test`
- `wald.test`
- `onevar.test`

- `t.test`
- `wilcox.test`
- `prop.test`
- `wald.test`
- `var.test`

These functions are *in addition to* the bootstrapping methods covered.

Supplemental Activities

The following activities are currently available from the STAT 200 website to give you some practice in performing statistical analyses concerning estimating a parameter or estimating the relationship between two parameters

- SCA 6a
- SCA 6b
- SCA-11
- SCA-12
- SCA-13
- SCA-21
- SCA-22
- SCA-23

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

Note: Some of these SCAs also deal with hypothesis testing, which will be covered in the near future.

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