Module E: Advanced Inference

Slide Deck E1:

Handling One and Two Means

The section in which we see how to perform hypothesis tests concerning population means using the R Statistical Environment. This deck will focus not only the code used to perform the analysis, it will also emphasize the analysis process, including testing assumptions before being able to use a particular statistical procedure.

Today's Objectives

By the end of this slidedeck, you should

- understand the theory behind, and test hypotheses about:
  - a single population mean
  - the difference between two population means
- better understand the p-value and how to test hypotheses
- clearly specify how confidence intervals and p-values both give important information about the population parameter

Note that we are moving beyond the general theory of confidence intervals and hypothesis testing. We are looking at how to specifically perform the procedures. Make sure you pay attention to the statistical process we follow.

Overview

Example

Elliot Bainbridge, a former student of mine, did a project concerning international political economy (IPE). As a part of his research, he needed to draw conclusions about the average GDP per capita in countries around the world. Were his theory correct, the average would be 15,000 USD.

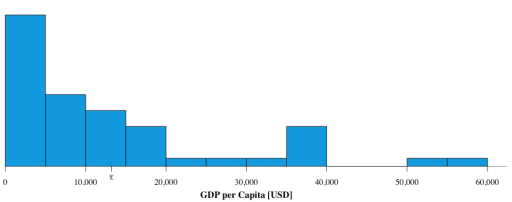
$H_R : \mu = 15{,}000$ *USD*

To test this, he collected a random sample of size $n = 50$. These data are in the `bainbridge` data file.

Does this provide sufficient evidence that Bainbridge is incorrect in his theory?

The Data

Here is the distribution of Elliot's sample:



Bootstrapping

Note that the hypothesis concerns the **population mean** and the likelihood of a *claimed* population mean. If the data are representative of the population, we could generate the distribution of sample means and see how likely the hypothesized mean is.

This is called "**bootstrapping the data**." It *only requires* the data are representative of the population. It makes no distributional assumptions about the process that generated the data.

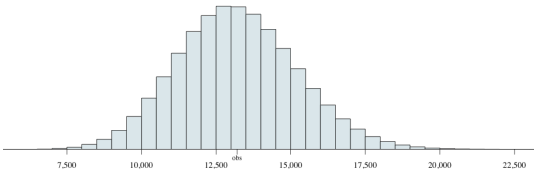
Bootstrapping

Here is the R-code to analyze the population mean using bootstrapping:

```
# The Data
dt = read.csv("http://rfs.kvasaheim.com/data/bainbridge.csv")

# Initialize variables
B = 10000                                ## Specify the number of iterations
ts = numeric()                           ## Tell R to set aside memory

# Simple bootstrapping of the mean
for(i in 1:B) {
  x = sample(dt$gdpcap, replace=TRUE)    ## Sample from data
  ts[i] = mean(x)                        ## Calculate mean
}

# Analysis Calculations
2 * mean( ts > 15000 )                   ## p-value
quantile( ts, c(0.025,0.975) )           ## confidence interval
```

Bootstrapping

Here is the distribution of **sample means**:



Bootstrapping

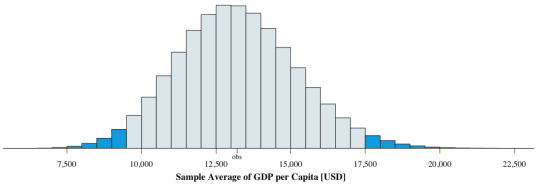
So, with this distribution of sample means, what can we conclude about the null hypothesis?

- **First**, we can directly test the null hypothesis.
  - Since the p-value is approximately 0.3508, we should not reject the null hypothesis. We cannot draw any conclusion about the population mean with respect to this particular claim. It could be greater than 15,000, less than 15,000, or equal to 15,000 USD.
- **Second**, we can get a 95% confidence interval from 9458 to 17,268 USD.
  - Since the hypothesized value of $\mu = 15{,}000$ is in the interval (between 9458 to 17,268 USD), we can conclude that the data do support the null hypothesis at the 95% confidence level.

Ultimately, it is reasonable to assume that the average GDP per capita is $15,000$ USD.

Bootstrapping

Here is the distribution of sample means with the confidence interval highlighted:



Non-Parametric

Bootstrapping only requires that the data are representative of the population. If we are willing to *also* assume that the population is symmetric, then we can obtain a test that is more-powerful — the Wilcoxon test (1945).

How can we tell if it is *reasonable* to state that the population is symmetric? Yeppers, the Hildebrand Rule will accomplish that for us, giving a ratio of $H = 0.3786$.

```
hildebrand.rule(dt$gdpcap)
```

Thus, it is *not* reasonable to treat the data as if they came from a symmetric process.

But, to illustrate the process, let us pretend that the data are sufficiently symmatric ($|H| < 0.20$). In such a case, we should use the Wilcoxon procedure (it is more powerful than bootstrapping).

Non-Parametric

So, this data allows us to use the Wilcoxon test. Here is the code to perform it in R for our original alternative hypothesis of

$$H_A : \mu \neq 15{,}000 \ \ USD$$

```
wilcox.test(dt$gdpcap, mu=15000, conf.int=TRUE)
```


Non-Parametric

Here is the R-output from this code:

```
Wilcoxon signed rank test with continuity correction

data:  dt$gdpcap

V = 453, p-value = 0.07569
alternative hypothesis: true location is not equal to 15000

95 percent confidence interval:
  7150   15950

sample estimates:
(pseudo)median
         9950
```

Be able to interpret every part of this output.

Non-Parametric

Interpretation:

The p-value of 0.07569 is greater than our usual alpha-value of $\alpha = 0.05$. Thus, we should not reject the null hypothesis. We do not know if the average GDP per capita is greater than 15,000 USD, less than 15,000 USD, or equal to 15,000 USD. It *is*, however, reasonable to assume that the average GDP per capita is 15,000 USD at this level of certainty.

In fact, we are 95% confident that the average GDP per capita in the world is between 7150 and 15,950 USD.

Parametric

Recall that bootstrapping only requires that the data are representative of the population (which is what we always assume about our sample). The Wilcoxon test *also* requires that the data are from a symmetric population.

IF we are *also* willing to assume that the data come from a **Normal** distribution, then we can obtain a test that is even more-powerful than the Wilcoxon test — the Student's t-test.

How can we tell if it is *reasonable* to state that the population is Normal? There are a large number of "Normality tests" available. One of the best is the **Shapiro-Wilk test** (1965).

Parametric

Here is how to perform the Shapiro-Wilk test for this data in R:

```
shapiroTest(dt$gdpcap)
```

The results of this code are

```
Shapiro-Wilk normality test

data:  y
W = 0.80819, p-value = 1.361e-06
```

The null hypothesis for this test is that the data are from a Normal distribution. Thus, since the p-value is less than 0.05, we conclude that it is *not* reasonable to conclude the data came from a Normal process (p-value $= 1.36 \times 10^{-6} = 0.000\,001\,361 \ll \alpha = 0.05$).

Parametric

While it is *definitely* ***not*** reasonable to conclude the data are from a Normal process, we should not perform the t-test.

However, to illustrate the analysis process, let up pretend that the data *are* from a Normal process. Here is the code to perform that analysis in R:

```
t.test(dt$gdpcap, mu=15000)
```

Parametric

Here is the output from this code:

```
One Sample t-test

data:  dt$gdpcap
t = -0.88841, df = 49, p-value = 0.3787

alternative hypothesis: true mean is not equal to 15000
95 percent confidence interval:
9147.993 17264.007

sample estimates:
mean of x
  13206
```


Parametric

Interpretation:
The p-value of 0.07569 is greater than our usual alpha-value of $\alpha = 0.05$. Thus, we should not reject the null hypothesis. We do not know if the average GDP per capita is greater than 15,000 USD, less than 15,000 USD, or equal to 15,000 USD. It is reasonable to assume the GDP per capita is 15,000 USD, given this stated level of certainty.

In fact, we are 95% confident that the average GDP per capita in the world is between 9148 and 17,264 USD.

Power Summary

The following are the confidence intervals and p-values for the three methods above:

Method	Lower CL	Upper CL	p-value
Bootstrap	9458	17,268	0.3508
Wilcoxon	7150	15,950	0.0757
Student's t-test	9148	17,264	0.3787

Note that all three methods tell the same substantive story about the population mean. Also, note:

- The interval width for the parametric test is narrower than for the non-parametric test. This illustrates that the parametric test is more powerful than the non-parametric test.
- The interval width for the bootstrap is *not* narrower than that of the non-parametric test. This illustrates that power is a general tendency; it does not hold perfectly for every example.

Analysis Procedure Summary

Notice the analysis process: We use the test that has the most requirements met by the data. The more assumptions met, the more powerful the test.

Test	Assumption(s)	Power
Student's t-test	Representative and Symmetric *and* Normal	High
Wilcoxon test	Representative *and* Symmetric	Moderate
Bootstrap	Representative	Low

A powerful test is better able to reject a false null hypothesis. All things being equal, it has a lower Type II error rate.

- Because of this, it is a better test. One should use it *if* allowed.

Example 1: The DFW Rate

Example

Every legitimate college goes through an accreditation process in which an outside reviewer examines the claim of the college and compares them to the realities. As a part of that process, the college evaluates each department and faculty member across several measures. One of which is the DFW rate — the proportion of students who receive a D, an F, or who withdraws from a course.

Unfortunately, there is no "perfect" DFW rate. With that said, there are definitely some values that raise questions: a rate "too" anything suggests there is a mismatch between the expectations on and the performance of the students.

With this in mind, I would like to estimate the DFW rate of my classes.

Example 1: The DFW Rate

While I been at Knox for several years, here are the data (DFW rates from ten randomly-selected class sections) I was able to sample:

$$15.2, 16.3, 22.5, 0.0, 33.3, 14.9, 21.4, 0.0, 0.0, 24.2$$

According to this data, I am 95% confident that my *average* DFW rate is between 6.5% and 23.1%.

This is the code I used for this analysis:

```
DFWrate = c(15.2, 16.3, 22.5, 0.0, 33.3, 14.9, 21.4, 0.0, 0.0, 24.2)
shapiroTest(DFWrate)
t.test(DFWrate)
```

Example 2: The STAT 200 DFW Rate

Example

To continue the previous example, it may be that certain classes have a higher (or lower) DFW rate than others. For instance, a faculty member may expect lower-division sections to average a C, but upper-division should average an A-.

With this in mind, I would like to estimate the DFW rate of my STAT 200 class to determine if it is in accord with expectations.

Example 2: The STAT 200 DFW Rate

While I been at Knox for several years, here are the data (DFW rates from eight randomly-selected class sections of STAT 200) I was able to sample:

$$15.2, 16.3, 14.9, 21.4, 24.2, 18.8, 21.3, 17.6$$

According to this data, I am 95% confident that my *average* DFW rate for STAT 200 is between 15.9% and 21.5%.

This is the code I used for this analysis:

```
DFWrate = c(15.2, 16.3, 14.9, 21.4, 24.2, 18.8, 21.3, 17.6)
shapiroTest(DFWrate)
t.test(DFWrate)
```

Example 3: The STAT 225 DFW Rate

Example

To continue the previous examples, it may be that certain classes have a higher (or lower) DFW rate than others. For instance, a faculty member may expect lower-division sections to average a C, but upper-division should average an A-.

With this in mind, I would like to estimate the DFW rate of my STAT 225 class to determine if it is in accord with (my) expectations of a 200-level course.

Example 3: The STAT 225 DFW Rate

While I been at Knox for several years, here are the data (DFW rates from six randomly-selected class sections of STAT 225) I was able to sample:

$$0.0, 0.0, 0.0, 25.0, 0.0, 16.7$$

According to the bootstrap method for this data, I am 95% confident that my *average* DFW rate for STAT 225 is between 0.0% and 15.3%.

Example 4: The DFW Rate of Prof F vs. Prof X

Example

To continue the previous few examples, I would like to compare my DFW rate with that of another STAT program member. To do this, I sampled from mine and their results over the past few years.

Example 4: The DFW Rate of Prof F vs. Prof X

According to the Knox Registrar's office, the DFW rates in STAT 200 for the past few years for Professor F are

$$15.2, 16.3, 14.9, 21.4, 24.2, 18.8, 21.3, 17.6$$

for Professor X,

$$11.2, 15.2, 9.1, 4.2$$

According to the Mann-Whitney procedure, I am 95% confident that Professor F has a higher *average* DFW rate for STAT 200 than Professor X between 2.0% and 15.6%.

Today's Objectives

Now that we have concluded this lecture, you should be able to

1. identify the research, null, and alternative hypotheses
2. calculate the p-value for a given alternative hypothesis
3. properly interpret the p-value

Today's R Functions

Here is what we used the following R functions:

- `shapiroTest(x)` performs the Shapiro-Wilk test for Normality
- `hildebrand.rule(x)` performs the Hildebrand Rule for skewness

- `t.test(x, mu)` performs the one-sample t-test
- `wilcox.test(x, mu, conf.int=TRUE)` performs the Wilcoxon test

- `t.test(x, y)` performs the two-sample t-test
- `wilcox.test(x, y, conf.int=TRUE)` performs the Mann-Whitney test

Supplemental Activities

The following activities are currently available from the STAT 200 website to give you some practice in performing hypothesis tests concerning population means.

- SCA 9a
- SCA 9b
- SCA 11
- SCA 21

Source: `https://www.kvasaheim.com/courses/stat200/sca/`

In addition to the SCAs, there are **Laboratory Activity E** (confidence intervals) and **Laboratory Activity F** (hypothesis testing).

Source: `https://www.kvasaheim.com/courses/stat200/labs/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of the material covered in this slidedeck:

- Hawkes Learning: Chapters 10 and 11
- Intro to Modern Statistics: Chapters 19–21
- `R` for Starters: Chapters 5 and 6
- Wikipedia: Confidence Intervals
  Hypothesis Testing

Please do not forget to use the `allProcedures` document that lists all of the statistical procedures we will use in `R`.