Module D: Introductory Inference

Slide Deck D4:

Hypothesis Testing in R

The section in which we see how to perform hypothesis testing using the R Statistical Environment. We should pay attention to the process behind selecting the correct test, how to perform that test, and how to convey to your reader what you did and what it means.

Today's Objectives

By the end of this slidedeck, you should

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

Yes, these are the same goals as in the last slide deck. The difference is that we will use R to perform the tests.

Make sure you pay attention to the statistical process we follow.

Overview

Example

WacDnalds claims that the weight of a quarter-pounder hamburger patty (before cooking) is 4 ounces. In symbols, this is

$$H_R : \mu = 4$$

To test this, a dietary scientist, Kagome Higurashi, collects data by weighing a random sample of $n = 10$ patties. Here are the weights she collected:

3	2	1	2	2	4	2	3	5	3

Does this provide sufficient evidence that WacDnalds is incorrect in their claim?

Bootstrapping

Again, here are the data she collected:

3	2	1	2	2	4	2	3	5	3

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.

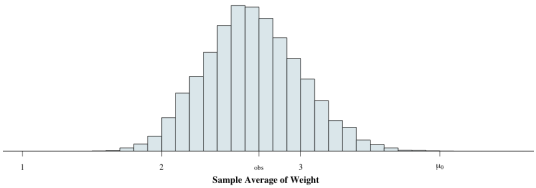
Bootstrapping

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

```
# The Data
weight = c(3, 2, 1, 2, 2, 4, 2, 3, 5, 3)

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

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

# Analysis Calculations
2 * mean( ts > 4 )                 ## p-value
quantile( ts, c(0.025,0.975) )     ## confidence interval
```


Bootstrapping

Here is the distribution of sample means:



Bootstrapping

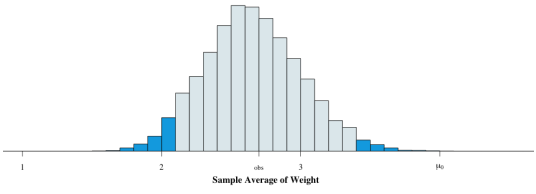
So, with this distribution of sample means, what can we conclude about the null hypothesis?

- **First**, we can directly test the null hypothesis. Since the p-value is approximately 0.0002, we can reject the null hypothesis and conclude that the average patty weight is not 4oz.
- **Second**, we can get a 95% confidence interval from 2.1 to 3.4 ounces.
  - Since the hypothesized value of $\mu = 4$ is not in the interval (between 2.1 and 3.4), we can conclude that the data do not support the null hypothesis at the 95% confidence level.

We conclude that the alternative hypothesis is correct, that $\mu \neq 4$. In fact, based on the confidence interval, we can also conclude $\mu <$ 4oz, that the average weight of a quarter-pounder at WacDnalds **is less than** the advertised 4 ounces.

Bootstrapping

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



Non-Parametric

Again, here are the data:

3	2	1	2	2	4	2	3	5	3

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 $H = 0.1725$.

```
hildebrand.rule(weight)
```

Thus, it is reasonable to treat the data as if they came from a symmetric process.

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

$$H_A : \mu \neq 4$$

```
wilcox.test(weight, mu=4, conf.int=TRUE)
```

Non-Parametric

Here is the R-output from this code:

```
Wilcoxon signed rank test with continuity correction

data:  weight

V = 2.5, p-value = 0.01868
alternative hypothesis: true location is not equal to 4
95 percent confidence interval:
 1.999960 3.500001

sample estimates:
(pseudo)median
      2.499996
```

Be able to interpret every part of this output.

Non-Parametric

Conclusion:

We are asked to determine if the average weight of hamburger patties is 4oz, before cooking. Because the Hildebrand ratio is $H = 0.1725$, we are able to use the Wilcoxon procedure to test our hypothesis.

The p-value of 0.0187 is less than our usual alpha-value of $\alpha = 0.05$. Thus, we reject the null hypothesis. There is significant evidence that the average weight of quarter-pounders at WacDnalds is less than a quarter of a pound. In fact, a 95% confidence interval for the average weight of a quarter-pounder burger is from 2.0 to 3.5 ounces.

Note

Again, the purpose of the conclusion is more than just conveying your numerical results. A good conclusion provides the procedures used, a brief defense of that procedure, and an interpretation of the results. Use numbers in the conclusion.

Parametric

Recall that bootstrapping only requires that the data are representative of the population (which is what we always assume about our smaple). 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 in R:

```
weight = c(3, 2, 1, 2, 2, 4, 2, 3, 5, 3)
shapiroTest(weight)
```

The results of this code are

```
Shapiro-Wilk normality test

data:  y
W = 0.91645, p-value = 0.3283
```

The null hypothesis for this test is that the data are from a Normal distribution. Thus, since the p-value is greater than 0.05, we conclude that it is reasonable to conclude the data came from a Normal process.

Parametric

Since it is reasonable to conclude the data are from a Normal process, we can perform the t-test. Here is the code to perform it in R for our hypothesis:

$$H_A : \mu \neq 4$$

```
t.test(weight, mu=4)
```


Parametric

Here is the R-output from this code:

```
One Sample t-test

data:  weight

t = -3.5455, df = 9, p-value = 0.00626
alternative hypothesis: true mean is not equal to 4

95 percent confidence interval:
 1.870542 3.529458

sample estimates:
mean of x
      2.7
```

Parametric

Conclusion:

We are asked to determine if the average weight of hamburger patties is 4oz, before cooking. Because the Shapiro-Wilk test suggests that the data are from a Normal distribution (p-value = 0.3283), we are able to use the one-sample t-test on our hypothesis.

The p-value of 0.0063 is less than our usual alpha-value of $\alpha = 0.05$. Thus, we reject the null hypothesis. There is significant evidence that the average weight of quarter-pounders at WacDnalds is less than a quarter of a pound. In fact, a 95% confidence interval for the average weight of a quarter-pounder burger is from 1.87 to 3.53 ounces.

Note

Again, the purpose of the conclusion is more than just conveying your numerical results. A good conclusion provides the procedures used, a brief defense of that procedure, and an interpretation of the results. Use numbers in the conclusion.

Analysis Procedure Summary

Notice the procedure: 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 1a: Road to Wolfenbüttel

Example

It is well known that drinking alcohol increases reaction time. When driving a car, this effect can be fatal.

To determine if this "common wisdom" is correct, I gathered a sample of STAT 200 students. I first measured the reaction time of each person sober. I then measured their reaction after drinking sufficient quantities of Jager Meister to have a BAC of 0.05%.

Here are the change in reaction times (in milliseconds):

143	182	192	205	115	362	172	192	203	61

Example 1a: Road to Wolfenbüttel

The first step is to determine the appropriate test. I would like to use the most-powerful test, but those tests have a lot of assumptions.

The following load the data and test Normality:

```
rxJ = c(143, 182, 192, 205, 115, 362, 172, 192, 203, 61)

shapiroTest(rxJ)
```

According to the Shapiro-Wilk test (p-value = 0.1144), there is no reason to believe that the data are not Normally distributed. As such, we can use the t-test.

Example 1a: Road to Wolfenbüttel

The code to perform the appropriate t-test is

```
t.test(rxJ, mu=0, alternative="greater")
```

The results are

```
One Sample t-test

data:  rxJ
t = 7.437, df = 9, p-value = 1.973e-05
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 137.6669      Inf
sample estimates:
mean of x
    182.7
```


Example 1a: Road to Wolfenbüttel

Conclusion:

We are asked to determine if the average difference in reaction times between sober and drunk (BAC = 0.05%) is positive. Because the Shapiro-Wilk test suggests that the data are from a Normal distribution (p-value = 0.1144), we are able to use the one-sample t-test on our hypothesis.

The p-value of less than 0.0001 is less than our usual alpha-value of $\alpha = 0.05$. Thus, we reject the null hypothesis. There is significant evidence that the average difference in reaction times is positive. In fact, we are 95% confident that the average increase in reaction time is at least 137.67 milliseconds.

Example 1b: Road to Cupertino

Example

While it is well known that drinking alcohol increases reaction time, it is *also* well known that talking on a telephone does the same.

To determine which affects reaction time most, I use the same people in the first example. After they get sober, I measure their reaction time while talking on their cell phones.

The change in reaction times between no-phone and yes-phone (ms) is provided below.

88	233	240	207	188	194	184	198	193	208

Example 1b: Road to Cupertino

The first step is to determine the appropriate test. I would like to use the most-powerful test, but those tests have a lot of assumptions.

The following load the data and test Normality and symmetry:

```
rxP = c(88, 233, 240, 207, 188, 194, 184, 198, 193, 208)

shapiroTest(rxP)
hildebrand.rule(rxP)
```

The Shapiro-Wilk test (p-value = 0.0081) indicates that the data are *not* from a Normal distribution. However, the Hildebrand Rule suggests that the data are generated from a symmetric distribution ($H = -0.0653$). As such, we should use the Wilcoxon test to estimate the population mean (and test that it differs from zero).

Example 1b: Road to Cupertino

The code to perform the appropriate Wilcoxon test is

```
wilcox.test(rxP, mu=0, alternative="greater", conf.int=TRUE)
```

The results are

```
Wilcoxon signed rank test

data:  rxP
V = 55, p-value = 0.0009766
alternative hypothesis: true location is greater than 0
95 percent confidence interval:
184 Inf
sample estimates:
(pseudo)median
         198
```


Example 1b: Road to Cupertino

Conclusion:

We are asked to determine if the average difference in reaction times between using a phone and not is positive. The Shapiro-Wilk test suggests that the data are not from a Normal distribution (p-value = 0.0081), we are unable to use the one-sample t-test on our hypothesis. The Hildebrand ratio ($H = -0.0653$) suggests the data are from a symmetric distribution. As such, we should use the Wilcoxon procedure.

The p-value (0.0010) is less than our usual alpha-value of $\alpha = 0.05$. Thus, we reject the null hypothesis. There is significant evidence that the average difference in reaction times is positive. In fact, we are 95% confident that the average increase in reaction time of a phone-user is at least 184 milliseconds.

Example 1c: Road to the District

Example

It is known that alcohol *and* cell phone use increase reaction time. However, this raises the question of which is worse. To determine this, I use the previously-collected data and appropriately compared their means.

The change in reaction times between sober and drunk, and between no-phone and yes-phone (ms) is provided below.

Alcohol:	143	182	192	205	115	362	172	192	203	61
Phone:	88	233	240	207	188	194	184	198	193	208

Example 1c: Road to the District

We are to compare the means of the two populations. To achieve the best power, we should test symmetry and Normality of both samples.

The phone data suggests non-Normality. It is the most (limiting) of the two subsets. Because of this, we should perform the two-sample Mann-Whitney test:

```
rxJ = c(143, 182, 192, 205, 115, 362, 172, 192, 203, 61)
rxP = c(88, 233, 240, 207, 188, 194, 184, 198, 193, 208)

wilcox.test(rxJ,rxP, conf.int=TRUE)
```

Example 1c: Road to the District

The results from this analysis are

```
Wilcoxon rank sum test with continuity correction

data:  rxJ and rxP

W = 32, p-value = 0.1857
alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:
 -60.99997  10.99994

sample estimates:
difference in location
             -15.99997
```


Example 1c: Road to the District

Conclusion:

We are asked to determine which suppresses reaction times more: being legally drunk or using a cell phone. The Shapiro-Wilk test suggests that the phone data are not from a Normal distribution (p-value = 0.0081). Thus, we should not use the two-sample t-test. As such, we should use the Mann-Whitney procedure.

The p-value (0.1857) is greater than our usual alpha-value of $\alpha = 0.05$. Thus, we do not know if being drunk or talking on the telephone slows reaction time more. Note that this conclusion may change with additional data, but this is the conclusion based on what we currently know.

Example 2: Remember the Cowboys

Example

There is always a lot of discussion about the "home field advantage" in sports. If such an advantage existed, then we would see home teams tend to score more points than the visiting teams.

To check this, I will use the 2015 Big XII conference.

```
source("http://rfs.kvasaheim.com/stat200.R")

dt = read.csv("http://rfs.kvasaheim.com/data/big12football2015.csv")
attach(dt)

names(dt)
```


Example 2: Remember the Cowboys

The two variables we will compare are `ptsFor` and `ptsAgainst`.

We would like to use the most powerful test. However, those tests have requirements that must be met. Let us check the requirements for this data:

```
shapiroTest(ptsFor)
shapiroTest(ptsAginst)
```

The Shapiro-Wilk test indicates that the `ptsAgainst` variable is not from a Normal distribution (p-value = 0.0378). Thus, we will use the Mann-Whitney test.

Example 2: Remember the Cowboys

The test:

```
wilcox.test(ptsFor,ptsAgainst, alternative="greater", conf.int=TRUE)
```

The results:

```
        Wilcoxon rank sum test with continuity correction

data:  ptsFor and ptsAgainst
W = 4573, p-value = 0.01905
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 0.99998     Inf
sample estimates:
difference in location
             6.000033
```


Example 2: Remember the Cowboys

Conclusion:

We are asked to determine whether there is evidence of a "home field advantage" using the 2015 Big XII football season. We will compare the average points scored by the home team to that scored by the away team. The Shapiro-Wilk test suggests that the point scored by the visiting team are not from a Normal distribution (p-value = 0.0378). Thus, we should not use the two-sample t-test. We should use the Mann-Whitney procedure.

The p-value (0.0191) is less than our usual alpha-value of $\alpha = 0.05$. Thus, we reject the null hypothesis. There is significant evidence that the average points scored by the home team is greater than those scored by the visiting team (a 95% confidence interval of at least 1 point). Thus, these data provide evidence of a small home-field advantage.

Example 3: Risky Business

Example

One of the more important parameters to estimate in financial mathematics is **risk**. Typically, "risk" is defined as the variance of the stock prices. Higher variance in prices is evidence of higher risk.

I would like to determine whether International Business Machines (IBM) or Apple (AAPL) has a higher risk.

To test this, we will need to import the prices for the two stocks, adjust their prices to ensure they are on a common scale, and finally compare their variances over time.

The first step is aided by the `quantmod` package that can be added to R. The third step just requires us to divide each by their averages. The last step is just a two-sample variance procedure.

Example 3: Risky Business

This is how to install and load the package:

```
### Preamble
install.packages("quantmod")    ## Only once per computer
library(quantmod)               ## Each time you use it
```

It is a new package, so working your way around it is slow at first. However, the following lines will download the IBM and Apple data from Yahoo Finance from Jan 3, 2007, until Aug 12, 2022, and store them in two variables.

```
getSymbols("IBM",src="yahoo")
stock1 = as.numeric(IBM$IBM.Close)

getSymbols("AAPL",src="yahoo")
stock2 = as.numeric(AAPL$AAPL.Close)
```

Example 3: Risky Business

Next, we have to adjust the prices so that they have the same average. This is done because risk is defined in terms of a set portfolio value of $1. The adjustment also allows us to interpret the changes in values as percent changes.

```
stock1 = stock1/mean(stock1)
stock2 = stock2/mean(stock2)
```

Now, we just compare the two variances.

```
var.test(stock1, stock2)
```


Example 3: Risky Business

The output:

```
F test to compare two variances

data:  stock1 and stock2
F = 0.034952, num df = 3930, denom df = 3930, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.03283305 0.03720756
sample estimates:
ratio of variances
        0.03495193
```

Example 3: Risky Business

Conclusion:

We are asked to determine which of two stocks, IBM and Apple (AAPL), is more risky. To do this, we will use the F test.

According to this test, there is a significant difference in variances between the two stocks (p-value $\ll 0.0001$), with Apple being the riskier stock. A 95% confidence interval for the ratio of Apple to IBM variance is from 26.87 to 30.46.

Example 4: A Tree Grows in Galesburg

Example

According to SAMHSA's 2004–2005 National Surveys on Drug Use and Health, males had higher rates than females for all measures of alcohol drinking in the past month. In fact, they found males were twice as likely as females to have met the criteria for alcohol dependence or abuse in the past year (10.5% vs. 5.1%).

I would like to determine if students at Knox College are typical in both proportions by gender and differences by gender.

To do this, I survey 50 students who identify as male and 25 who identify as female. For each student surveyed, I performed a typical "Alcohol Abuse Survey" (AAS). In my sample, 24 males and six females tested positive on the AAS.

Example 4: A Tree Grows in Galesburg

This problem compares two population proportions. Thus, we have two options: bootstrapping and the proportions test.

Note

While the sample size is "sufficient" to use the proportions test, let us use both to get a better feel for the data and what it is telling us.

Example 4: A Tree Grows in Galesburg

There are several hypotheses being asked. The first is that the male proportion testing positive on the AAS is 10.5%.

```
genM = rbinom(1e6, size=50, prob=0.105)   ## Based on H0
mean(genM >= 24)

genM = rbinom(1e6, size=50, prob=24/50)   ## Based on data
quantile(genM, c(0.025,0.975))/50
```

The p-value from the simulation test is much, much less than our usual alpha-value of 0.05. As such, we conclude that the rate of a positive AAS assessment for males at Knox College is larger than the national average.

We are 95% confident that the proportion of males at Knox who test positive on AAS are from 34% to 62%.

Example 4: A Tree Grows in Galesburg

The second hypothesis is that the female proportion testing positive on the AAS is 5.1%.

```
genF = rbinom(1e6, size=25, prob=0.051)    ## Based on H0
mean(genF >= 6)

genF = rbinom(1e6, size=25, prob=6/25)     ## Based on data
quantile(genF, c(0.025,0.975))/25
```

The p-value from the simulation test (0.001325) is less than our usual alpha-value of 0.05. As such, we can conclude that the rate of a positive AAS assessment for females at Knox College is larger than the national average.

We are 95% confident that the proportion of females at Knox who test positive on AAS are from 8% to 40%. The accepted proportion of 5.1% is not in this interval.

Example 4: A Tree Grows in Galesburg

Alternatively, you can use the built-in Binomial test:

```
binom.test(xm,nm, p=0.105)
binom.test(xf,nf, p=0.051)
```

These lines provide the same substantive results, although the confidence intervals will be slightly different:

> *We are 95% confident that the proportion of males at Knox College who are AAS-positive is between 33.7% and 62.6%; of females, between 9.4% and 45.1%.*

Example 4: A Tree Grows in Galesburg

The third embedded hypothesis is that males are twice as likely as females to be AAS-positive.

One way of testing this is simply to double the female successes in the sample and/or doubling the female success rate in the population. This is not the "optimal" method, but it usually does a sufficient job. A second method is to halve the number of "successes" in the male-identifying population.

What I tend to do is both of these. If the two methods tell the same story, then I am confident in my conclusion. If, on the other hand, they give contradicting results, then I report both test results and conclude that, *if* there is a detected effect, then it is quite minor.

Example 4: A Tree Grows in Galesburg

Here is the code and the output from the first method.

```
prop.test(c(xm,xf*2), c(nm,nf))
```

```
2-sample test for equality of proportions without continuity correction

data:  c(xm, xf * 2) out of c(nm, nf)
X-squared = 0, df = 1, p-value = 1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2398535  0.2398535
sample estimates:
prop 1 prop 2
  0.48   0.48
```

I leave it as an exercise to show that the second method gives similar results.

Example 4: A Tree Grows in Galesburg

Conclusion:

According to SAMHSA, the proportion of males who test positive on the AAS assessment is about double that of females. The male proportion is 10.5%; the female, 5.1%.

For Knox College, the proportion of males testing positive on the AAS Assessment is significantly higher, from 34% to 62%. Similarly, the proportion of females testing positive is also significantly higher, from 8% to 40%.

The SAMHSA noted that the proportion of males testing positive is about twice the proportion that females tested positive. In this, Knox College students seem to be typical. There is no evidence that the ratio is anything other than what SAMHSA found in the population at large.

All of this suggests that the relationship between male and female problematic drinking holds at Knox College, even if the actual rates of AAS-positive findings are higher.

Today's Objectives

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

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

Since we used R to perform the calculations, we were better able to focus on the interpretation than on the tedious calculations.

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

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
- `binom.test(x, n)` performs the Binomial test for one proportion
- `onevar.test(x, s)` performs the one-sample χ^2-test for variance

- `t.test(x, y)` performs the two-sample t-test
- `wilcox.test(x, y, conf.int=TRUE)` performs the Mann-Whitney test
- `prop.test(x=c(x1,x2), n=c(n1,n2))` performs the proportions test for comparing two proportions
- `var.test(x, y)` performs the two-sample F-test for variance

Supplemental Activities

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

- SCA 9a
- SCA 9b
- SCA-11
- SCA-12
- SCA-13
- SCA-21
- SCA-22
- SCA-23

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

In addition to the SCAs, there is **Laboratory Activity F**, which also looks at 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 understanding the theory of hypothesis testing:

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

- Wikipedia: Hypothesis Testing