Null Hypothesis Testing

EPSY 630 - Statistics II

Jason Bryer, Ph.D.

February 23, 2021

Agenda

- Tip for debugging R markdown
- Review CLT / Questions
- Bootstrapping
- Inference for Categorical Variables
- Inference for Numerical Variables
- One minute papers

Debugging R Markdown

- Click Tools -> Global Options
  - Under Workspace, uncheck *Restore .RData into workspace at startup*
  - For *Save workspace to .RData on ext*, choose *Never*
- When you encounter an error, or before you are done, it is good to rerun the document from a clean slate. Click Session -> Restart R. This will unload all packages and any data. You can then start from the top of the document.
  - If there is a particular R chunk that is causing an issue, click this icon to run all R code up to the current R chunk .
  - Then you can click this icon to run the current R chunk
- Getting figures and output to show up. Change the following line towards the top of the RMarkdown:

```
knitr::opts_chunk$set(eval = TRUE, results = FALSE, fig.show = "hide", message = FALSE, warnir
```

to

Central Limit Theorem (CLT)

The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

where SE represents the **standard error**, which is defined as the standard deviation of the sampling distribution. In most cases σ is not known, so use s.

Null Hypothesis Testing

- We start with a null hypothesis (H_0) that represents the status quo.
- We also have an alternative hypothesis (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem.
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Bootstrapping

Bootstrapping

- First introduced by Efron (1979) in *Bootstrap Methods: Another Look at the Jackknife*.
- Estimates confidence of statistics by resampling *with* replacement.
- The *bootstrap sample* provides an estimate of the sampling distribution.
- The `boot` R package provides a framework for doing bootstrapping: https://www.statmethods.net/advstats/bootstrapping.html

Bootstrapping Example (Population)

Define our population with a uniform distribution.

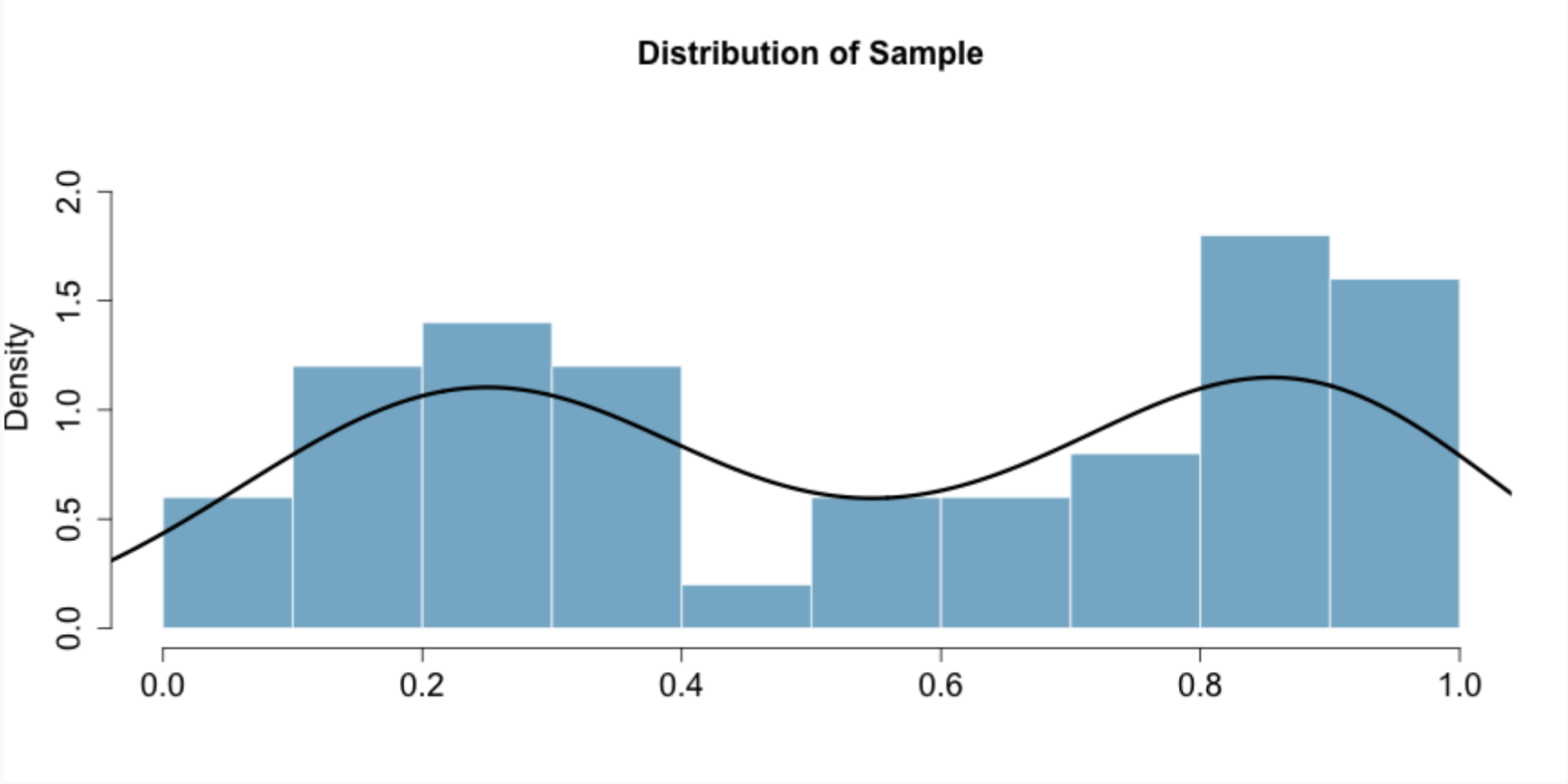
```
n <- 1e5
pop <- runif(n, 0, 1)
mean(pop)
```

```
## [1] 0.5012519
```

Bootstrapping Example (Sample)

We observe one random sample from the population.

```
samp1 <- sample(pop, size = 50)
```



Bootsrapping Example (Estimate)

```
boot.samples <- numeric(1000) # 1,000 bootstrap samples
for(i in seq_along(boot.samples)) {
    tmp <- sample(samp1, size = length(samp1), replace = TRUE)
    boot.samples[i] <- mean(tmp)
}
head(boot.samples)
```

```
## [1] 0.5541522 0.5224922 0.5698420 0.5386104 0.4624742 0.5406412
```

Bootsrapping Example (Distribution)

```
d <- density(boot.samples)
h <- hist(boot.samples, plot=FALSE)
hist(boot.samples, main='Bootstrap Distribution', xlab="", freq=FALSE,
     ylim=c(0, max(d$y, h$density)+.5), col=COL[1,2], border = "white",
     cex.main = 1.5, cex.axis = 1.5, cex.lab = 1.5)
lines(d, lwd=3)
```

95% confidence interval

```
c(mean(boot.samples) - 1.96 * sd(boot.samples),
  mean(boot.samples) + 1.96 * sd(boot.samples))
```

```
## [1] 0.4558440 0.6328191
```

Bootstrapping is not just for means!

```
boot.samples.median <- numeric(1000) # 1,000 bootstrap samples
for(i in seq_along(boot.samples.median)) {
    tmp <- sample(samp1, size = length(samp1), replace = TRUE)
    boot.samples.median[i] <- median(tmp) # NOTICE WE ARE NOW USING THE median FUNCTION!
}
head(boot.samples.median)
```

```
## [1] 0.6991149 0.6623520 0.3587462 0.3587462 0.6623520 0.6623520
```

95% confidence interval for the median

```
c(mean(boot.samples.median) - 1.96 * sd(boot.samples.median),
  mean(boot.samples.median) + 1.96 * sd(boot.samples.median))
```

```
## [1] 0.2810075 0.8015061
```

Inference for Categorical Variables

Example

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1,000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

> - **500 get the drug, 500 don't**

Survey of Americans

The GSS asks the same question, below is the distribution of responses from the 2010 survey:

Response	n
All 1000 get the drug	99
500 get the drug 500 don't	571
Total	**670**

Parameter of Interest

- Parameter of interest: Proportion of *all* Americans who have good intuition about experimental design.

$$p(population\ proportion)$$

- Point estimate: Proportion of *sampled* Americans who have good intuition about experimental design.

$$\hat{p}(sample\ proportion)$$

Inference for a proportion

What percent of all Americans have good intuition about experimental design (i.e. would answer "500 get the drug 500 don't?"

- Using a confidence interval

$$point\ estimate \pm ME$$

- We know that ME = critical value x standard error of the point estimate.

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Central limit theoreom for proportions

Sample proportions will be nearly normally distributed with mean equal to the population mean, p, and standard error equal to $\sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

This is true given the following conditions:

- independent observations
- at least 10 successes and 10 failures

Back to the Survey

- 571 out of 670 (85%) of Americans answered the question on experimental design correctly.
- Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

Given: $n = 670$, $\hat{p} = 0.85$.

Conditions:

1. Independence: The sample is random, and 670 < 10% of all Americans, therefore we can assume that one respondent's response is independent of another.

2. Success-failure: 571 people answered correctly (successes) and 99 answered incorrectly (failures), both are greater than 10.

Calculating Confidence Interval

Given: $n = 670$, $\hat{p} = 0.85$.

$$0.85 \pm 1.96\sqrt{\frac{0.85 \times 0.15}{670}} = (0.82,\ 0.88)$$

We are 95% confidence the true proportion of Americans that have a good intuition about experimental designs is betwee 82% and 88%.

How many should we sample?

Suppose you want a 3% margin of error, how many people would you have to survey?

Use $\hat{p} = 0.5$

- If you don't know any better, 50-50 is a good guess
- $\hat{p} = 0.5$ gives the most conservative estimate - highest possible sample size

$$0.03 = 1.96 \times \sqrt{\frac{0.5 \times 0.5}{n}}$$

$$0.03^2 = 1.96^2 \times \frac{0.5 \times 0.5}{n}$$

$$n \approx 1,068$$

Example: Two Proportions

Scientists predict that global warming may have big effects on the polar regions within the next 100 years. One of the possible effects is that the northern ice cap may completely melt. Would this bother you a great deal, some, a little, or not at all if it actually happened?

Response	GSS	Duke
A great deal	454	69
Some	124	40
A little	52	4
Not at all	50	2
Total	680	105

Parameter and Point Estimate

Parameter of interest: Difference between the proportions of *all* Duke students and *all* Americans who would be bothered a great deal by the northern ice cap completely melting.

$$p_{Duke} - p_{US}$$

Point estimate: Difference between the proportions of *sampled* Duke students and *sampled* Americans who would be bothered a great deal by the northern ice cap completely melting.

$$\hat{p}_{Duke} - \hat{p}_{US}$$

Everything else is the same...

- CI: $point\ estimate \pm margin\ of\ error$
- HT: $Z = \frac{point\ estimate - null\ value}{SE}$

Standard error of the difference between two sample proportions

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1\left(1-p_1\right)}{n_1} + \frac{p_2\left(1-p_2\right)}{n_2}}$$

Conditions:

1. Independence within groups: The US group is sampled randomly and we're assuming that the Duke group represents a random sample as well. $n_{Duke} < 10\%$ of all Duke students and $680 < 10\%$ of all Americans.
2. Independence between groups: The sampled Duke students and the US residents are independent of each other.
3. Success-failure: At least 10 observed successes and 10 observed failures in the two groups.

95% Confidence Interval

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap ($p_{Duke} - p_{US}$).

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z* \times \sqrt{\frac{p_{Duke}\left(1 - p_{Duke}\right)}{n_{Duke}} + \frac{p_{US}\left(1 - p_{US}\right)}{n_{US}}}$$

$$(0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} = (-0.108,\ 0.086)$$

Weldon's dice

- Walter Frank Raphael Weldon (1860 - 1906), was an English evolutionary biologist and a founder of biometry. He was the joint founding editor of Biometrika, with Francis Galton and Karl Pearson.
- In 1894, he rolled 12 dice 26,306 times, and recorded the number of 5s or 6s (which he considered to be a success).
    - It was observed that 5s or 6s occurred more often than expected, and Pearson hypothesized that this was probably due to the construction of the dice. Most inexpensive dice have hollowed-out pips, and since opposite sides add to 7, the face with 6 pips is lighter than its opposing face, which has only 1 pip.

Labby's dice

In 2009, Zacariah Labby (U of Chicago), repeated Weldon's experiment using a homemade dice-throwing, pip counting machine. http://www.youtube.com/watch?v=95EErdouO2w

- The rolling-imaging process took about 20 seconds per roll.
  - Each day there were ~150 images to process manually.
  - At this rate Weldon's experiment was repeated in a little more than six full days.
  - Recommended reading: http://galton.uchicago.edu/about/docs/labby09dice.pdf

Summarizing Labby's results

The table below shows the observed and expected counts from Labby's experiment.

Outcome	Observed	Expected
1	53,222	52,612
2	52,118	52,612
3	52,465	52,612
4	52,338	52,612
5	52,244	52,612
6	53,285	52,612
Total	315,672	315,672

Setting the hypotheses

Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

- H_0: There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.

- H_A: There is an inconsistency between the observed and the expected counts. The observed counts **do not** follow the same distribution as the expected counts. There is a bias in which side comes up on the roll of a die.

Evaluating the hypotheses

- To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts.
- Large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis.
- This is called a \hl{goodness of fit} test since we're evaluating how well the observed data fit the expected distribution.

Anatomy of a test statistic

- The general form of a test statistic is:

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

- This construction is based on

  1. identifying the difference between a point estimate and an expected value if the null hypothesis was true, and
  2. standardizing that difference using the standard error of the point estimate.

- These two ideas will help in the construction of an appropriate test statistic for count data.

Chi-Squared

When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the chi-square (χ^2) statistic.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

where k = total number of cells

Outcome	Observed	Expected	$\frac{(O-E)^2}{E}$
1	53,222	52,612	$\frac{(53{,}222-52{,}612)^2}{52{,}612} = 7.07$
2	52,118	52,612	$\frac{(52{,}118-52{,}612)^2}{52{,}612} = 4.64$
3	52,465	52,612	$\frac{(52{,}465-52{,}612)^2}{52{,}612} = 0.41$
4	52,338	52,612	$\frac{(52{,}338-52{,}612)^2}{52{,}612} = 1.43$
5	52,244	52,612	$\frac{(52{,}244-52{,}612)^2}{52{,}612} = 2.57$
6	53,285	52,612	$\frac{(53{,}285-52{,}612)^2}{52{,}612} = 8.61$
Total	315,672	315,672	24.73

Chi-Squared Distribution

Squaring the difference between the observed and the expected outcome does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already looked unusual will become much larger after being squared.

In order to determine if the χ^2 statistic we calculated is considered unusually high or not we need to first describe its distribution.

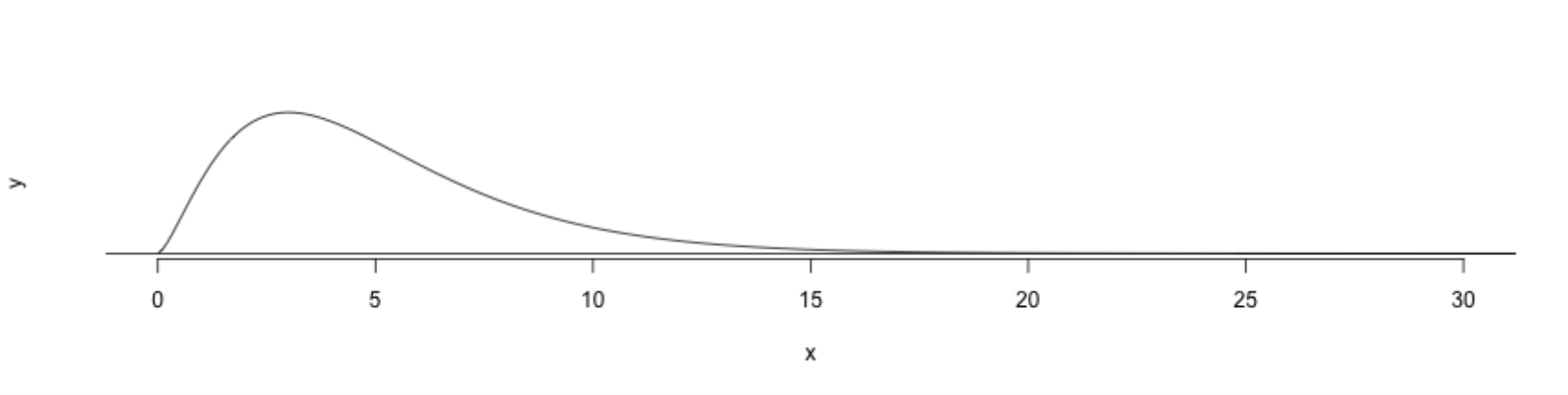
- The chi-square distribution has just one parameter called **degrees of freedom (df)**, which influences the shape, center, and spread of the distribution.

Degrees of freedom for a goodness of fit test

When conducting a goodness of fit test to evaluate how well the observed data follow an expected distribution, the degrees of freedom are calculated as the number of cells (k) minus 1.

$$df = k - 1$$

For dice outcomes, $k = 6$, therefore $df = 6 - 1 = 5$



p-value = $P(\chi^2_{df=5} > 24.67)$ is less than 0.001

Turns out...

- The 1-6 axis is consistently shorter than the other two (2-5 and 3-4), thereby supporting the hypothesis that the faces with one and six pips are larger than the other faces.
- Pearson's claim that 5s and 6s appear more often due to the carved-out pips is not supported by these data.
- Dice used in casinos have flush faces, where the pips are filled in with a plastic of the same density as the surrounding material and are precisely balanced.

Recap: p-value for a chi-square test

- The p-value for a chi-square test is defined as the tail area **above** the calculated test statistic.

- This is because the test statistic is always positive, and a higher test statistic means a stronger deviation from the null hypothesis.

Independence Between Groups

Assume we have a population of 100,000 where groups A and B are independent with $p_A = .55$ and $p_B = .6$ and $n_A = 99,000$ (99% of the population) and $n_B = 1,000$ (1% of the population). We can sample from the population (that includes groups A and B) and from group B of sample sizes of 1,000 and 100, respectively. We can also calculate $\hat{p}$ for group A independent of B.

```
propA <- .55     # Proportion for group A
propB <- .6      # Proportion for group B
pop.n <- 100000 # Population size
sampleA.n <- 1000
sampleB.n <- 100
```

```
pop <- data.frame(
    group = c(rep('A', pop.n * 0.99),
              rep('B', pop.n * 0.01) ),
    response = c(
        sample(c(1,0),
               size = pop.n * 0.99,
               prob = c(propA, 1 - propA),
               replace = TRUE),
        sample(c(1,0),
               size = pop.n * 0.01,
               prob = c(propB, 1 - propB),
               replace = TRUE) )
)
sampA <- pop[sample(nrow(pop),
                    size = sampleA.n),]
sampB <- pop[sample(which(pop$group == 'B'),
                    size = sampleB.n),]
```

Independence Between Groups (cont.)

$\hat{p}$ for the population sample

```
mean(sampA$response)
```

```
## [1] 0.534
```

$\hat{p}$ for the population sample, excluding group B

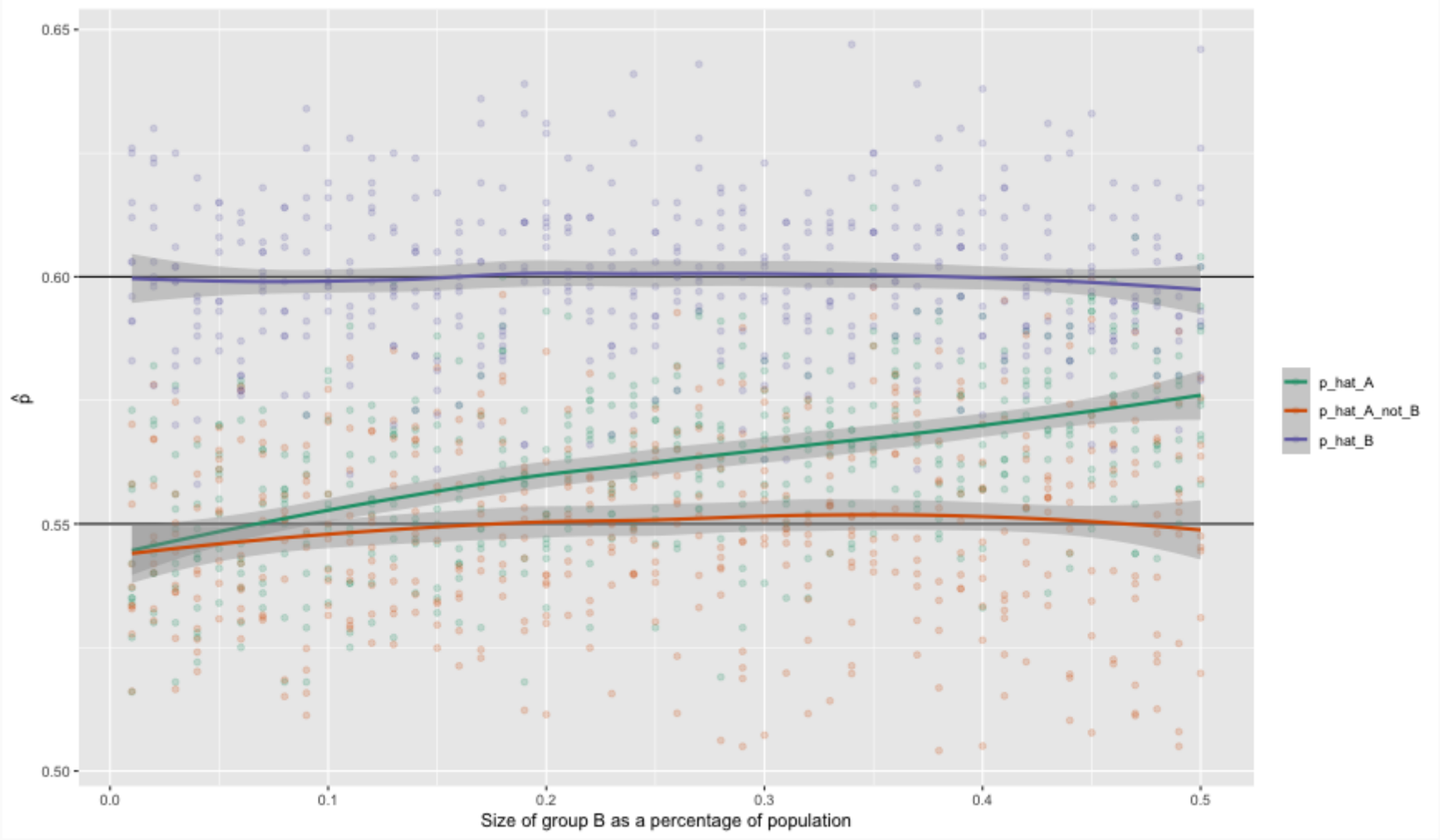
```
mean(sampA[sampA$group == 'A',]$response)
```

```
## [1] 0.5308392
```

$\hat{p}$ for group B sample

```
mean(sampB$response)
```

Independence Between Groups (cont.)

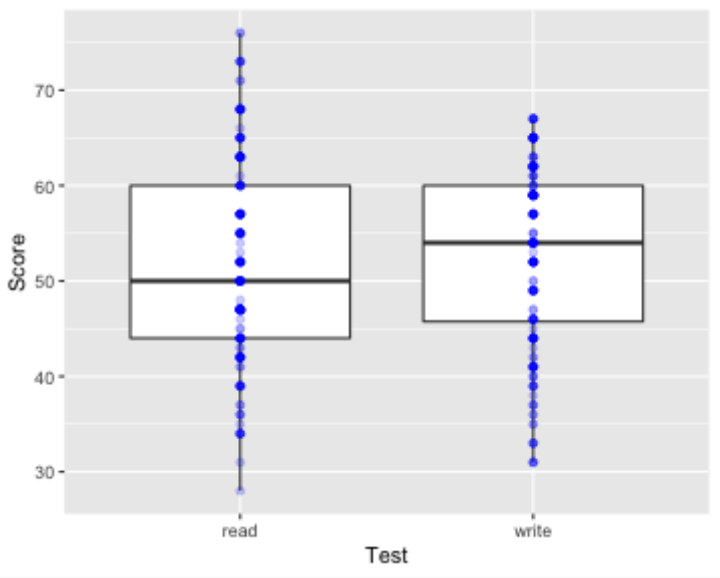


Inference for Numerical Variables

High School & Beyond Survey

200 randomly selected students completed the reading and writing test of the High School and Beyond survey. The results appear to the right. Does there appear to be a difference?

```
data(hsb2) # in openintro package
hsb2.melt <- melt(hsb2[,c('id','read', 'write')], id='id')
ggplot(hsb2.melt, aes(x=variable, y=value)) +     geom_boxplot() +
    geom_point(alpha=0.2, color='blue') + xlab('Test') + ylab('Score')
```



High School & Beyond Survey

```
head(hsb2)
```

```
## # A tibble: 6 x 11
##      id gender race  ses    schtyp prog        read write  math science socst
##   <int> <chr>  <chr> <fct>  <fct>  <fct>      <int> <int> <int>   <int> <int>
## 1    70 male   white low    public general       57    52    41      47    57
## 2   121 female white middle public vocational    68    59    53      63    61
## 3    86 male   white high   public general       44    33    54      58    31
## 4   141 male   white high   public vocational    63    44    47      53    56
## 5   172 male   white middle public academic      47    52    57      53    61
## 6   113 male   white middle public academic      44    52    51      63    61
```

Are the reading and writing scores of each student independent of each other?

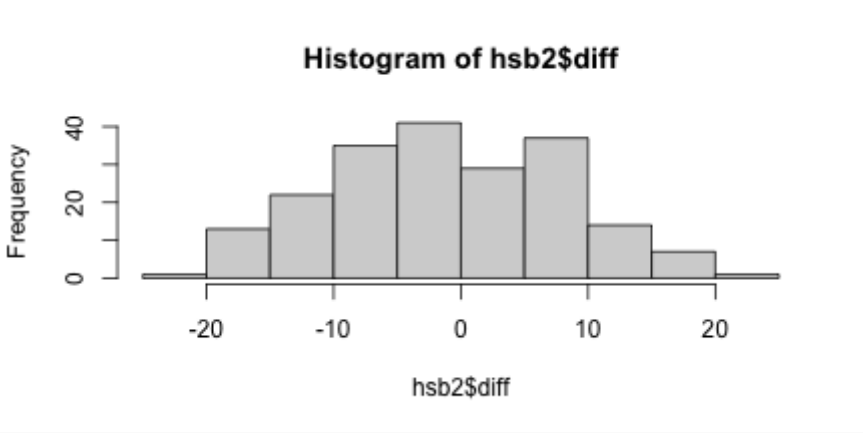
Analyzing Paired Data

- When two sets of observations are not independent, they are said to be paired.
- To analyze these type of data, we often look at the difference.

```
hsb2$diff <- hsb2$read - hsb2$write
head(hsb2$diff)
```

```
## [1]  5  9 11 19 -5 -8
```

```
hist(hsb2$diff)
```



Setting the Hypothesis

What are the hypothesis for testing if there is a difference between the average reading and writing scores?

H_0: There is no difference between the average reading and writing scores.

$$\mu_{diff} = 0$$

H_A: There is a difference between the average reading and writing score.

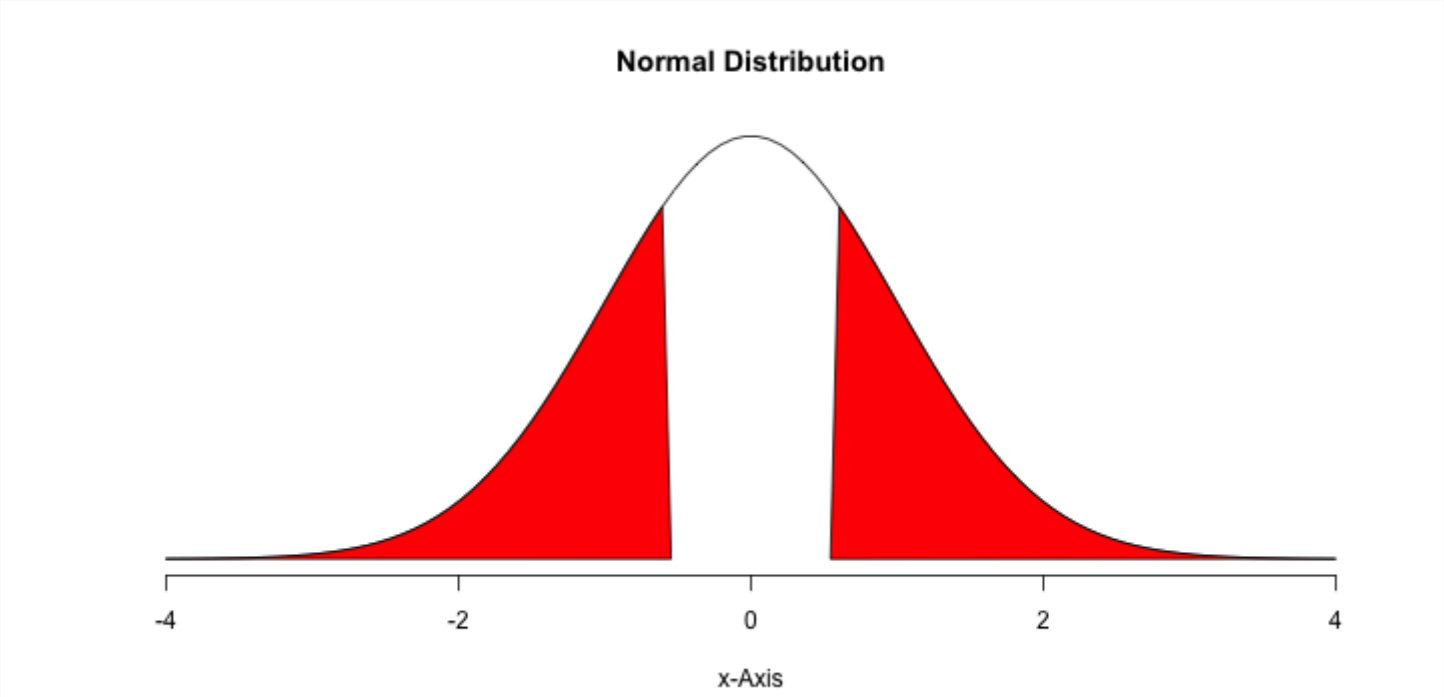
$$\mu_{diff} \neq 0$$

Nothing new here...

- The analysis is no different that what we have done before.
- We have data from one sample: differences.
- We are testing to see if the average difference is different that 0.

Calculating the test-statistic and the p-value

The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.8866664 points. Do these data provide confincing evidence of a difference between the average scores ont eh two exams (use $\alpha = 0.05$)?



Calculating the test-statistic and the p-value

$$Z = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = \frac{-0.545}{0.628} = -0.87$$

$$p - value = 0.1949 \times 2 = 0.3898$$

Since p-value > 0.05, we fail to reject the null hypothesis. That is, the data do not provide evidence that there is a statistically significant difference between the average reading and writing scores.

```
2 * pnorm(mean(hsb2$diff), mean=0, sd=sd(hsb2$diff)/sqrt(nrow(hsb2)))
```

```
## [1] 0.3857741
```

Interpretation of the p-value

The probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the score is 0, is 38%.

Calculating 95% Confidence Interval

$$-0.545 \pm 1.96 \frac{8.887}{\sqrt{200}} = -0.545 \pm 1.96 \times 0.628 = (-1.775, 0.685)$$

Note that the confidence interval spans zero!

SAT Scores by Gender

```
data(sat)
head(sat)
```

```
##   Verbal.SAT Math.SAT Sex
## 1        450      450   F
## 2        640      540   F
## 3        590      570   M
## 4        400      400   M
## 5        600      590   M
## 6        610      610   M
```

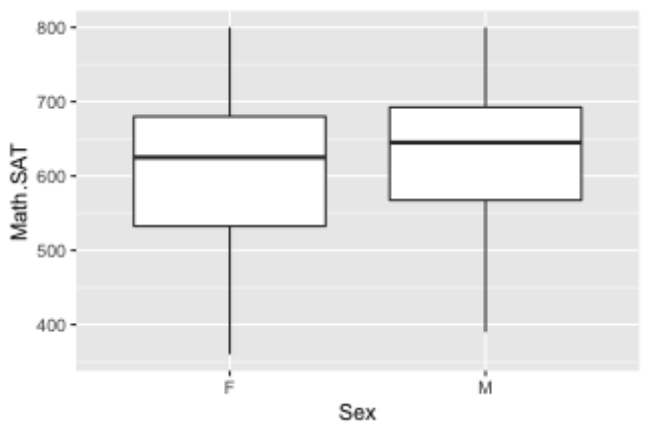
Is there a difference in math scores between males and females?

SAT Scores by Gender

```
describeBy(sat$Math.SAT, group=sat$Sex, mat=TRUE, skew=FALSE)[,c(2,4:7)]
```

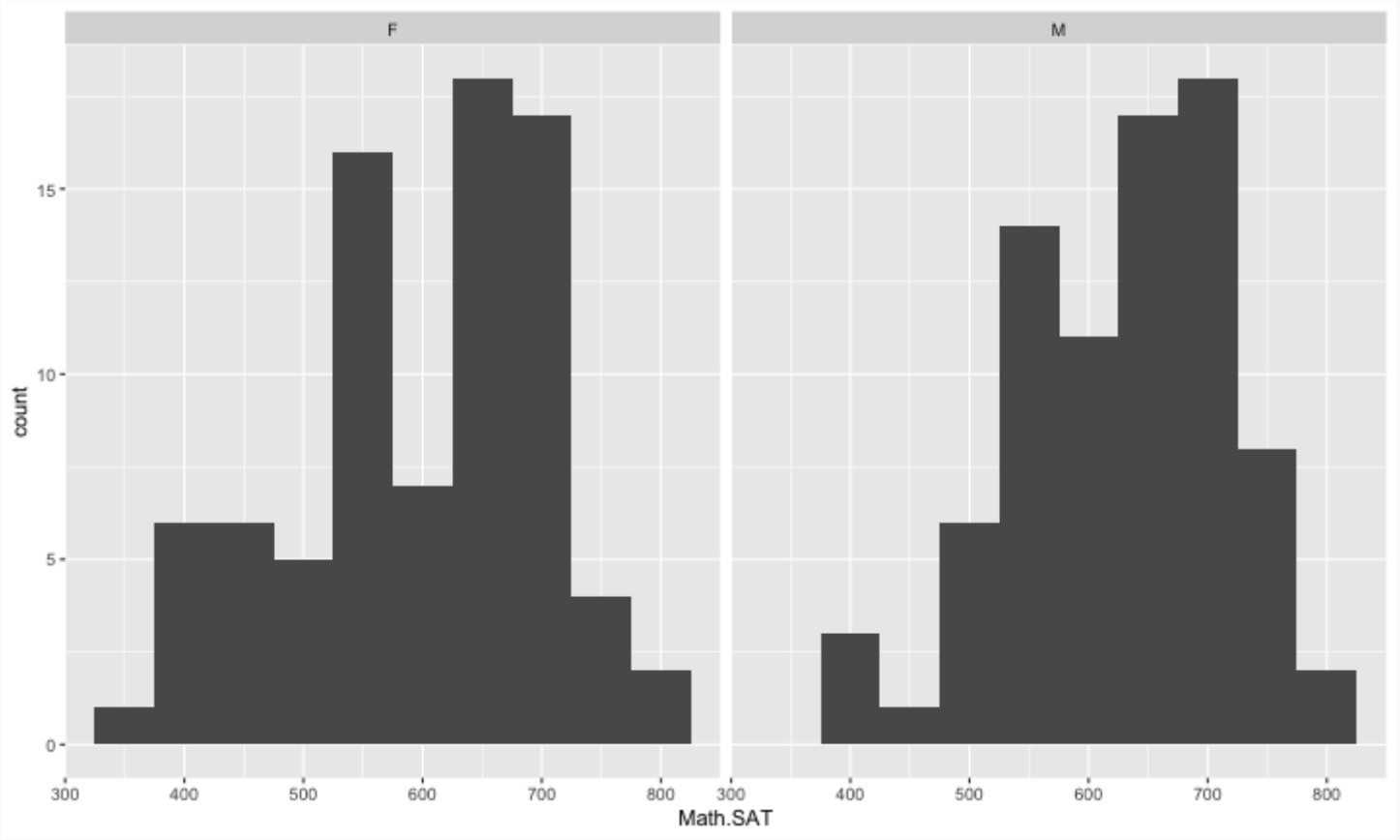
```
##     group1  n     mean        sd min
## X11      F 82 597.6829 103.70065 360
## X12      M 80 626.8750  90.35225 390
```

```
ggplot(sat, aes(x=Sex, y=Math.SAT)) + geom_boxplot()
```



Distributions

```
ggplot(sat, aes(x=Math.SAT)) + geom_histogram(binwidth=50) + facet_wrap(~ Sex)
```



95% Confidence Interval

We wish to calculate a 95% confidence interval for the average difference between SAT scores for males and females.

Assumptions:

1. Independence within groups.
2. Independence between groups.
3. Sample size/skew

Confidence Interval for Difference Between Two Means

- All confidence intervals have the same form: point estimate X ME
- And all ME = critical value X SE of point estimate
- In this case the point estimate is $\bar{x}_1 - \bar{x}_2$ Since the sample sizes are large enough, the critical value is z* So the only new concept is the standard error of the difference between two means...

Standard error of the difference between two sample means

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence Interval for Difference in SAT Scores

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}} = \sqrt{\frac{90.4}{80} + \frac{103.7}{82}} = 1.55$$

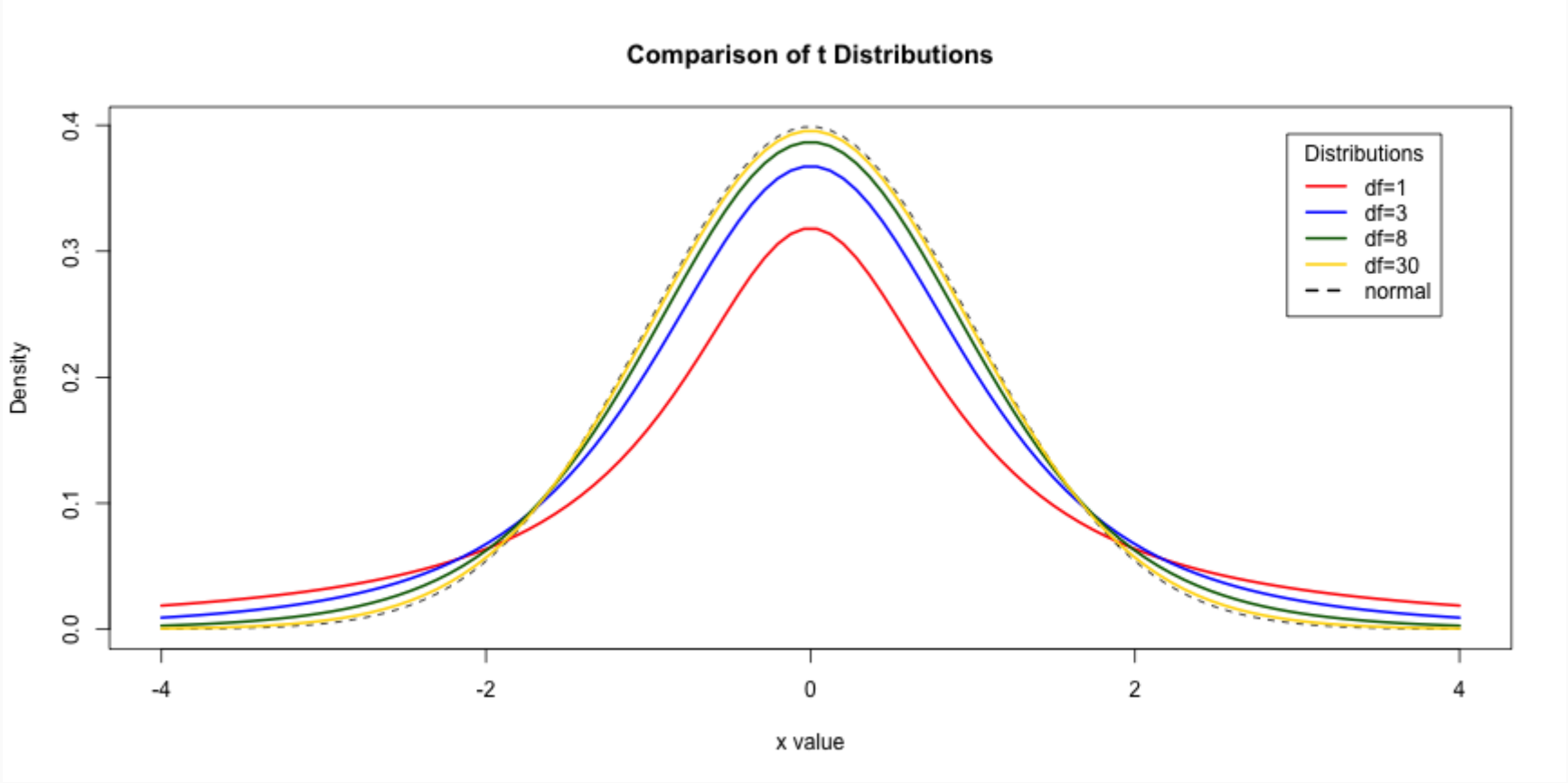
Student's *t*-Distribution

What if you want to compare the quality of one batch of Guinness beer to the next?

- Sample sizes necessarily need to be small.
- The CLT states that the sampling distribution approximates normal as n -> Infinity
- Need an alternative to the normal distribution.
- The *t* distribution was developed by William Gosset (under the pseudonym *student*) to estimate means when the sample size is small.

Confidence interval is estimated using

t-Distributions



t-test in R

The `pt` and `qt` will give you the *p*-value and critical value from the *t*-distribution, respectively.

Critical value for p = 0.05, degrees of freedom = 10

```
qt(0.025, df = 10)
```

```
## [1] -2.228139
```

p-value for a critical value of 2, degrees of freedom = 10

```
pt(2, df=10)
```

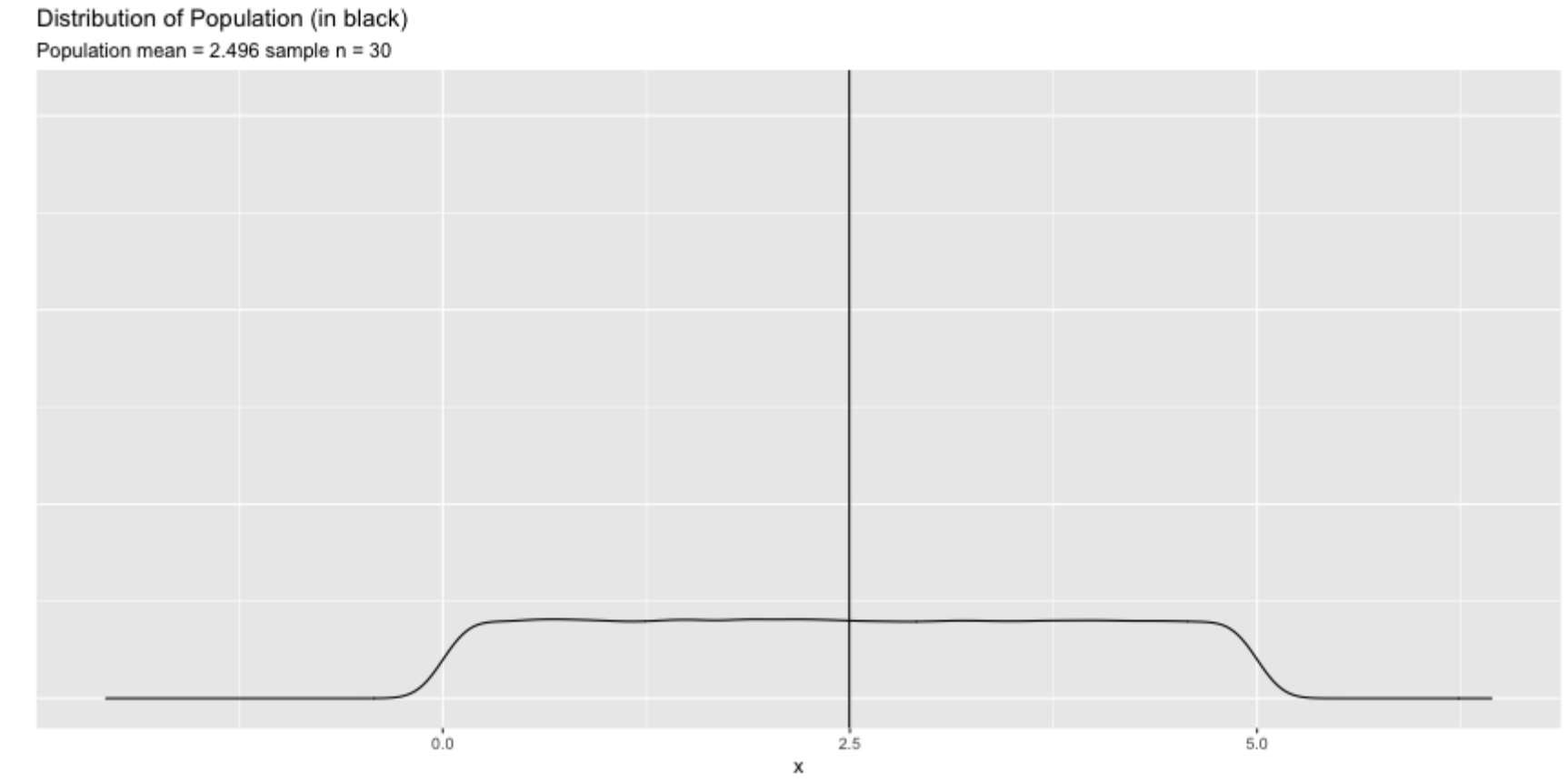
```
## [1] 0.963306
```

The `t.test` function will calculate a null hyphothesis test using the *t*-distribution.

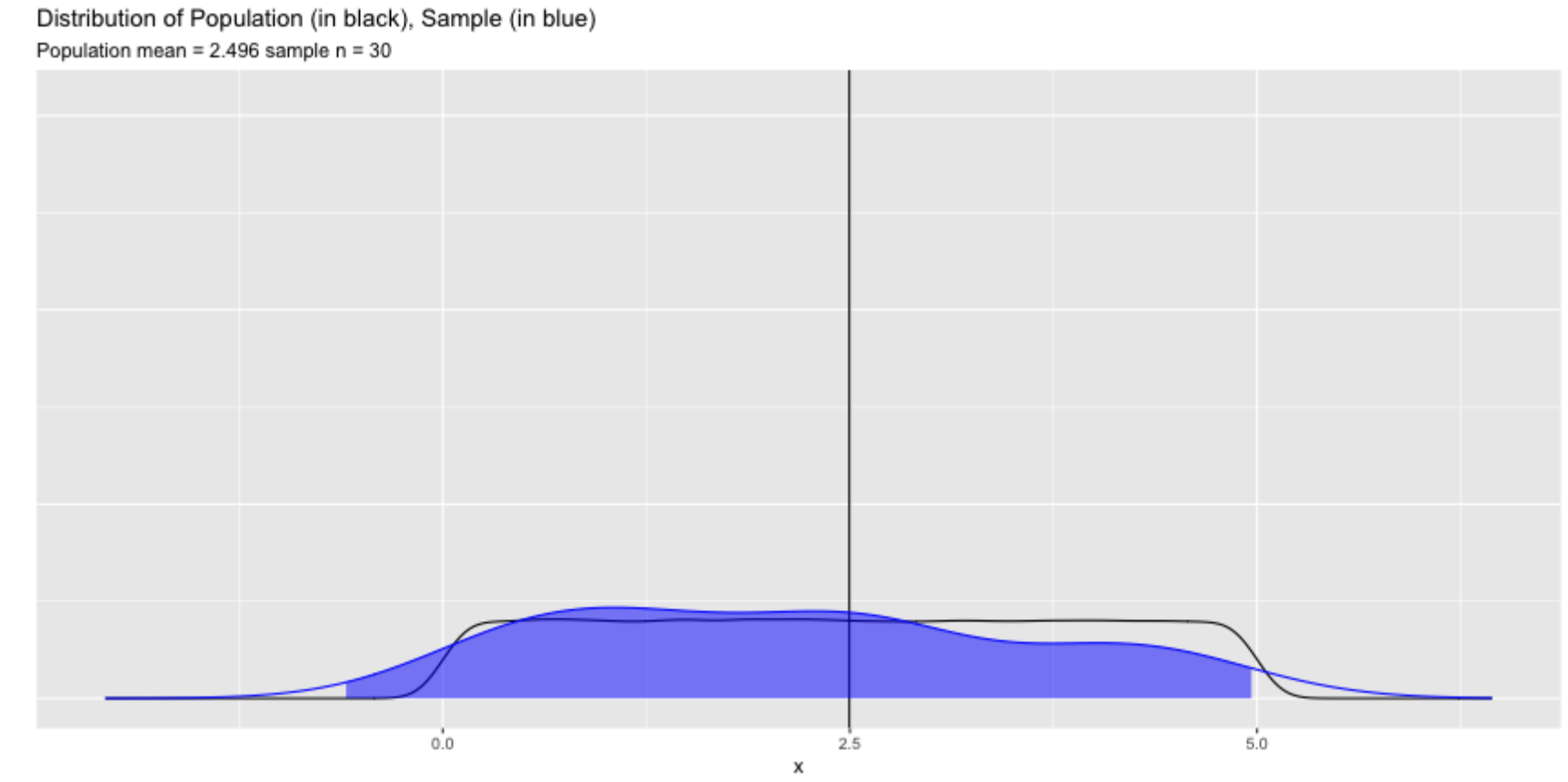
```
t.test(Math.SAT ~ Sex, data = sat)
```

```
##
##     Welch Two Sample t-test
##
## data:  Math.SAT by Sex
## t = -1.9117, df = 158.01, p-value = 0.057
## alternative hypothesis: true difference
## 95 percent confidence interval:
##  -59.3527145   0.9685682
## sample estimates:
## mean in group F mean in group M
```

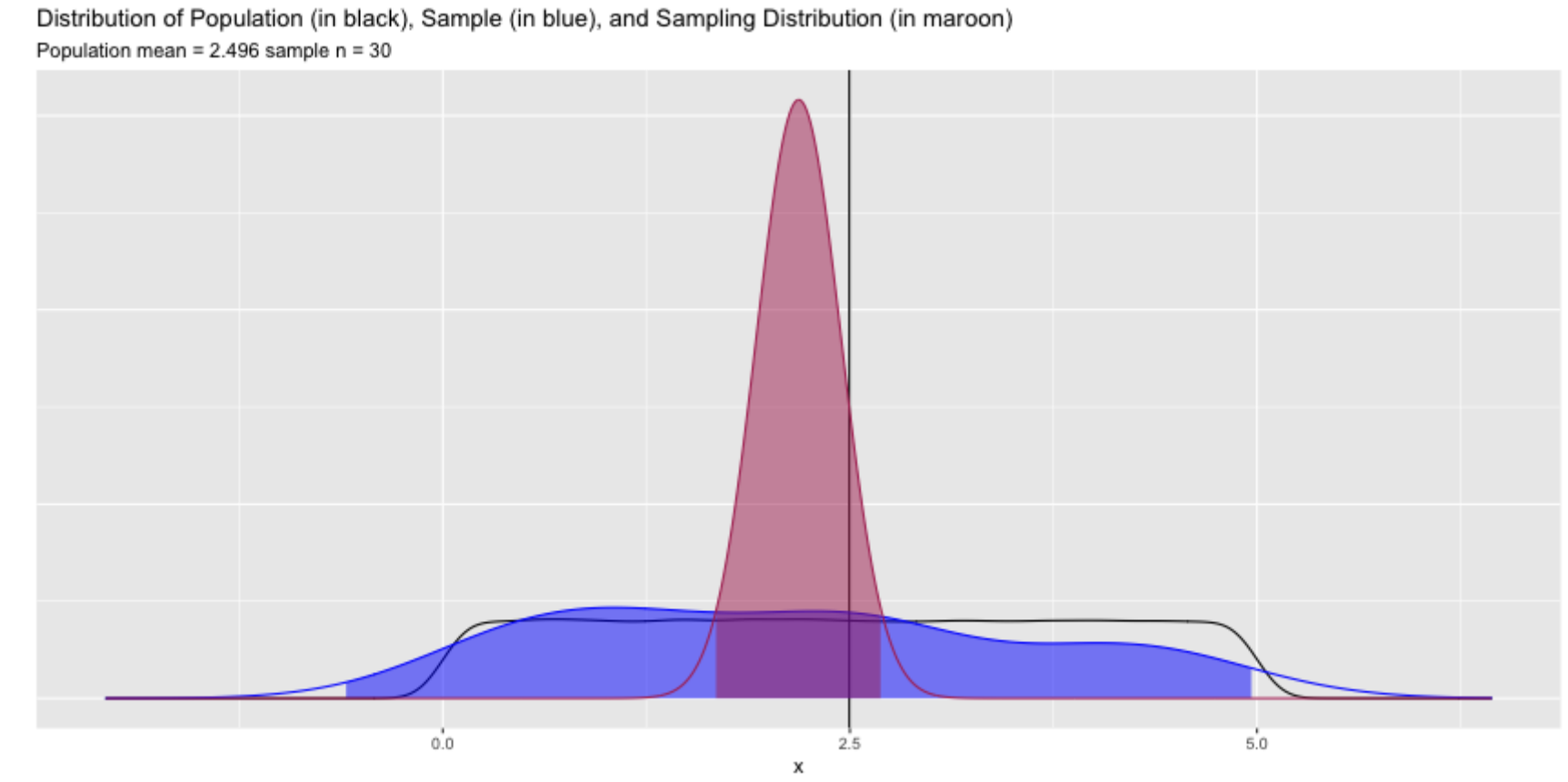
Review: Sampling Distribution



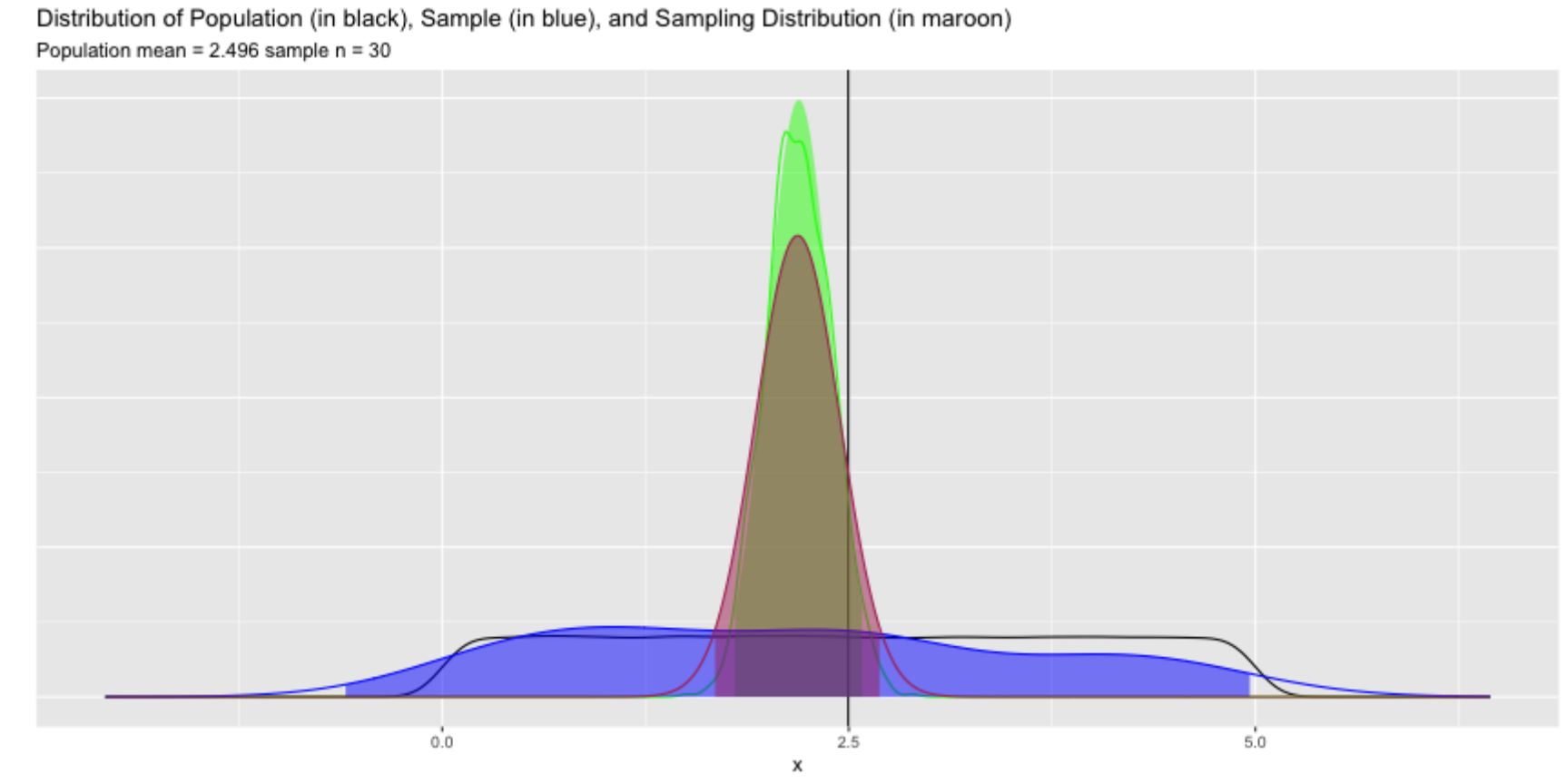
Review: Sampling Distribution



Review: Sampling Distribution



Review: Add Bootstrap Distribution



Assignments

Lab 5 is in two parts: A) Sampling Distributions and B) Confidence Levels. To get started, run the following commands:

```
DATA606::startLab('Lab5a') # https://r.bryer.org/shiny/Lab5a/
DATA606::startLab('Lab5b') # https://r.bryer.org/shiny/Lab5b/
```

One Minute Paper

Complete the one minute paper:

https://forms.gle/yB3ds6MYE89Z1pURA

1. What was the most important thing you learned during this class?
2. What important question remains unanswered for you?