Linear Regression & Analysis of Variance

EPSY 630 - Statistics II

Jason Bryer, Ph.D.

March 9, 2021

Agenda

- Linear regression review
- Analysis of Variance
- New lab
- One minute papers

EPSY 630
Spring 2021

Linear Regression (cont.)

NYS Report Card

NYS publishes data for each school in the state. We will look at the grade 8 math scores for 2012 and 2013. 2013 was the first year the tests were aligned with the Common Core Standards. There was a lot of press about how the passing rates for most schools dropped. Two questions we wish to answer:

1. Did the passing rates drop in a predictable manner?
2. Were the drops different for charter and public schools?

```
load('../course_data/NYSReportCard-Grade7Math.Rda')
names(reportCard)
```

```
##  [1] "BEDSCODE"      "School"        "NumTested2012" "Mean2012"      "Pass2012"
##  [6] "Charter"       "GradeSubject"  "County"        "BOCES"         "NumTested2013"
## [11] "Mean2013"      "Pass2013"
```

EPSY 630 Spring 2021

`reportCard` Data Frame

Show 3 entries   Search:

BEDSCODE	School	NumTested2012	Mean2012	Pass2012	Charter	GradeSubject	County	BOCES	NumTested2013	Mean2013
010100010020	NORTH ALBANY ACADEMY	47	649	13	false	Grade 7 Math	Albany	BOCES ALBANY-SCHOH-SCHENECTADY-SARAT	45	268
010100010030	WILLIAM S HACKETT MIDDLE SCHOOL	212	652	30	false	Grade 7 Math	Albany	BOCES ALBANY-SCHOH-SCHENECTADY-SARAT	250	279
010100010045	STEPHEN AND HARRIET MYERS MIDDLE SCHOOL	262	670	50	false	Grade 7 Math	Albany	BOCES ALBANY-SCHOH-SCHENECTADY-SARAT	256	284

Showing 1 to 3 of 1,362 entries

Previous 1 2 3 4 5 … 454 Next

Descriptive Statistics

```r
summary(reportCard$Pass2012)
```

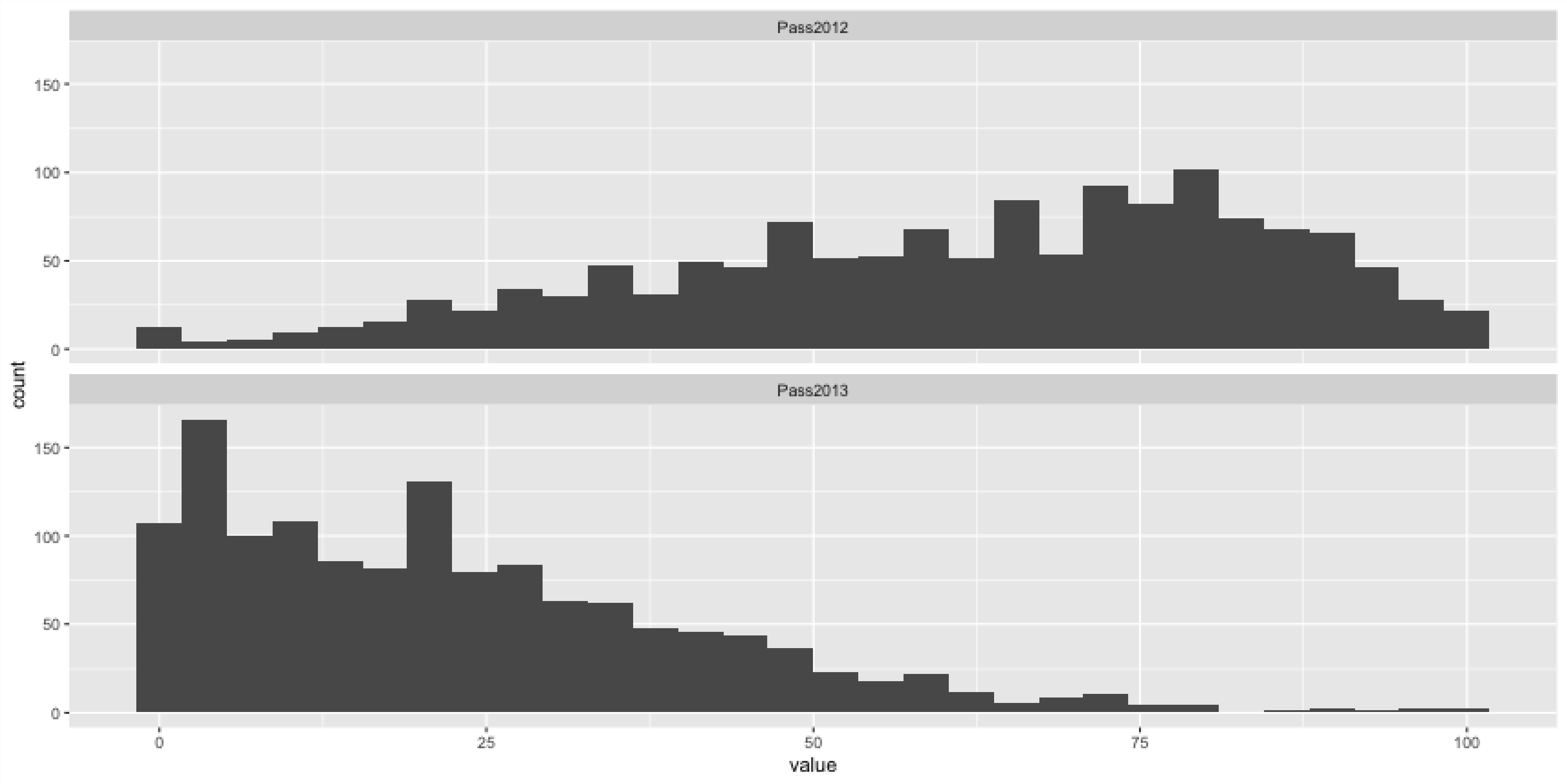
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   46.00   65.00   61.73   80.00  100.00
```

```r
summary(reportCard$Pass2013)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.00   20.00   22.83   33.00   99.00
```

Histograms

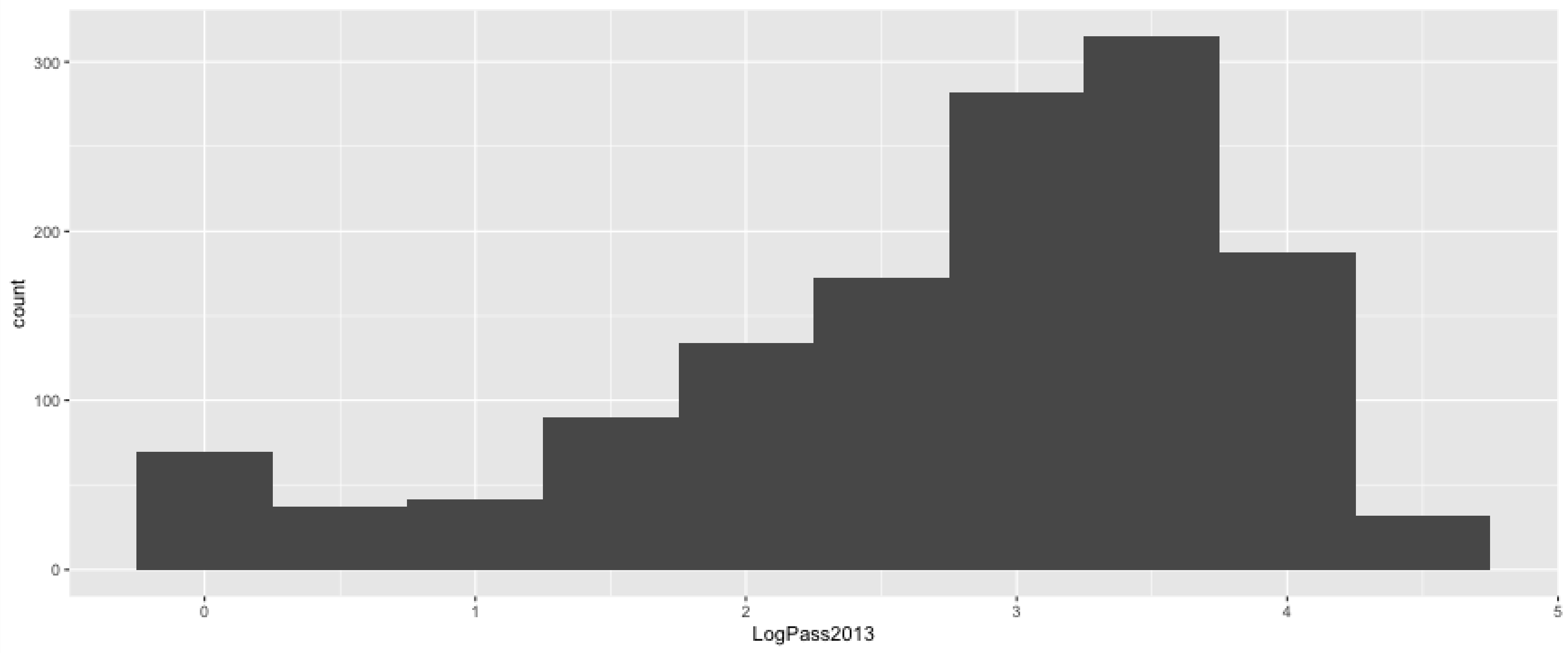
```
melted <- melt(reportCard[,c('Pass2012', 'Pass2013')])
ggplot(melted, aes(x=value)) + geom_histogram() + facet_wrap(~ variable, ncol=1)
```



Log Transformation

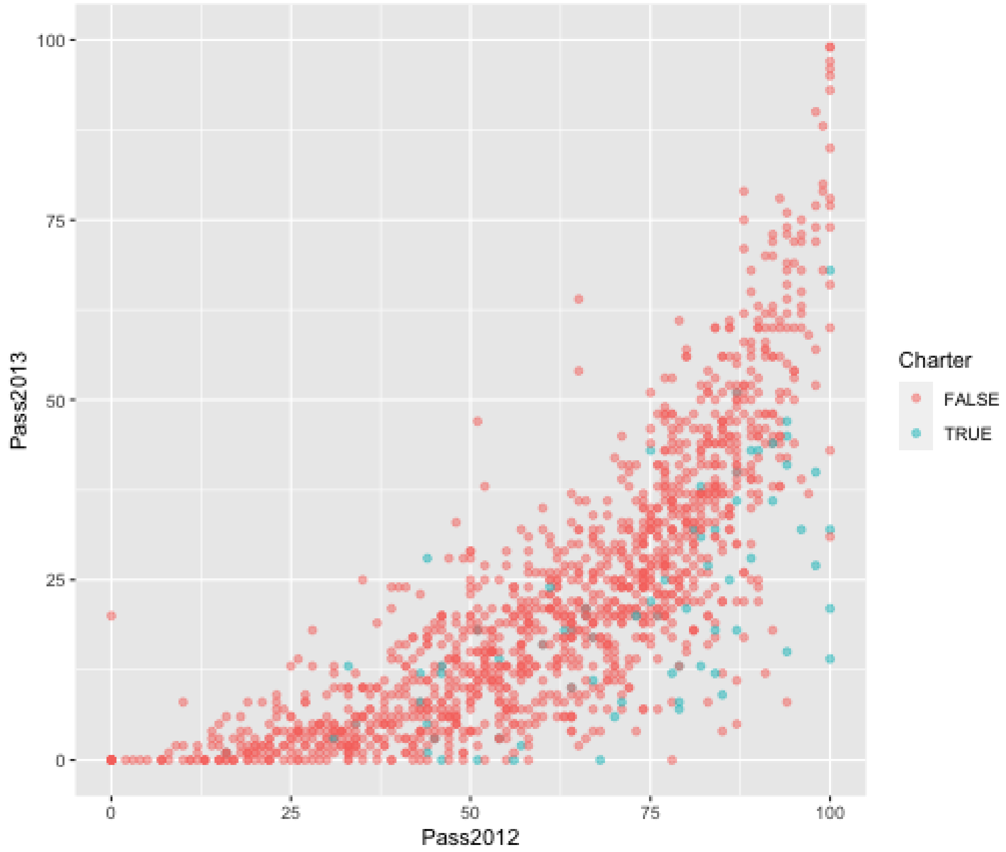
Since the distribution of the 2013 passing rates is skewed, we can log transfor that variable to get a more reasonably normal distribution.

```
reportCard$LogPass2013 <- log(reportCard$Pass2013 + 1)
ggplot(reportCard, aes(x=LogPass2013)) + geom_histogram(binwidth=0.5)
```



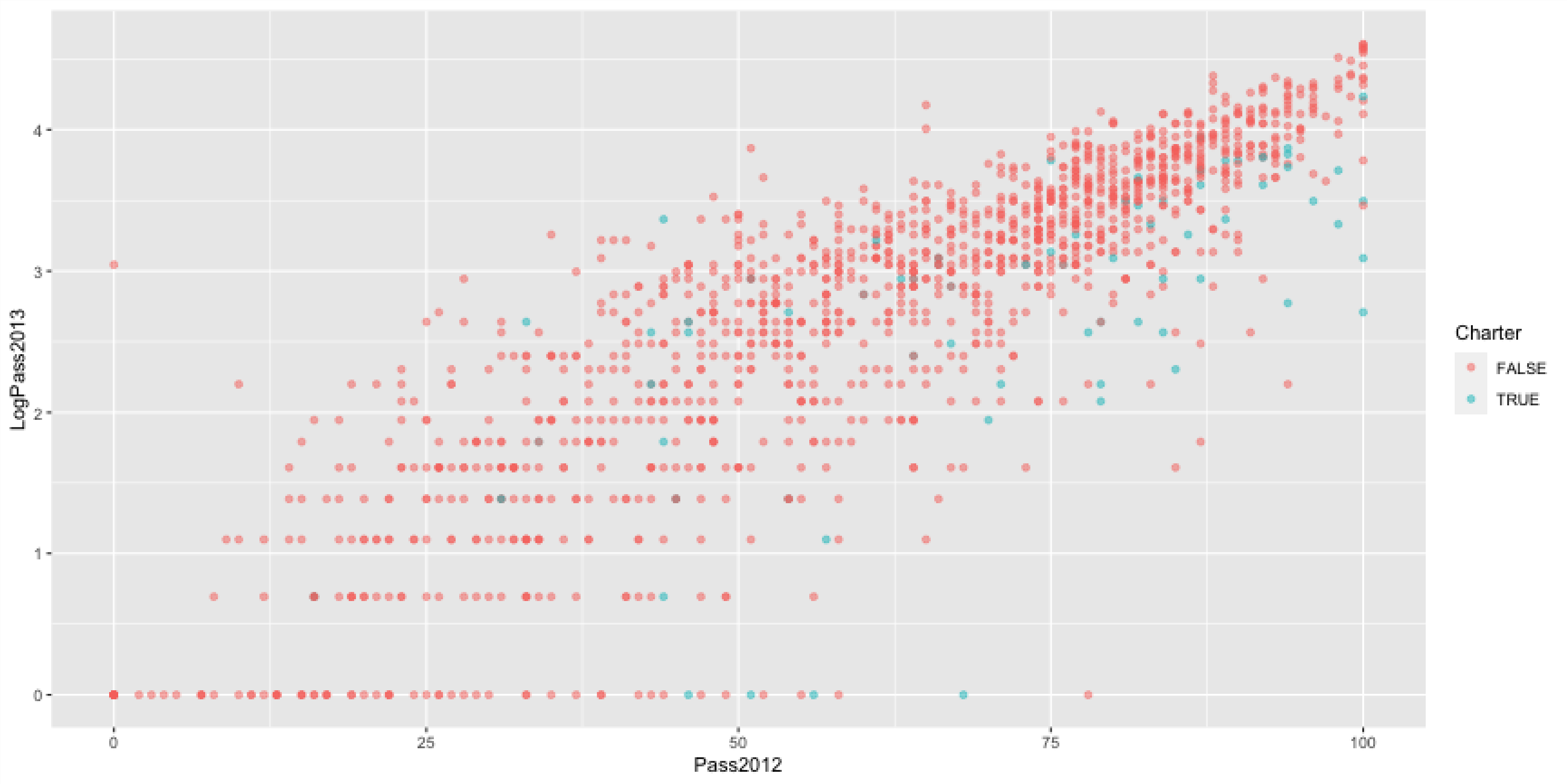
Scatter Plot

```
ggplot(reportCard, aes(x=Pass2012, y=Pass2013, color=Charter)) +
    geom_point(alpha=0.5) + coord_equal() + ylim(c(0,100)) + xlim(c(0,100))
```



Scatter Plot (log transform)

```
ggplot(reportCard, aes(x=Pass2012, y=LogPass2013, color=Charter)) +
    geom_point(alpha=0.5) + xlim(c(0,100)) + ylim(c(0, log(101)))
```



Correlation

```
cor.test(reportCard$Pass2012, reportCard$Pass2013)
```

```
##
##      Pearson's product-moment correlation
##
## data:  reportCard$Pass2012 and reportCard$Pass2013
## t = 47.166, df = 1360, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7667526 0.8071276
## sample estimates:
##       cor
## 0.7877848
```

Correlation (log transform)

```r
cor.test(reportCard$Pass2012, reportCard$LogPass2013)
```

```
##
##      Pearson's product-moment correlation
##
## data:  reportCard$Pass2012 and reportCard$LogPass2013
## t = 56.499, df = 1360, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8207912 0.8525925
## sample estimates:
##       cor
## 0.8373991
```

Linear Regression

```
lm.out <- lm(Pass2013 ~ Pass2012, data=reportCard)
summary(lm.out)
```

```
##
## Call:
## lm(formula = Pass2013 ~ Pass2012, data = reportCard)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.484  -6.878  -0.478   5.965  51.675
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.68965    0.89378  -18.67   <2e-16 ***
## Pass2012      0.64014    0.01357   47.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.49 on 1360 degrees of freedom
## Multiple R-squared:  0.6206,    Adjusted R-squared:  0.6203
## F-statistic:  2225 on 1 and 1360 DF,  p-value: < 2.2e-16
```

EPSY 630
Spring 2021

Linear Regression (log transform)

```
lm.log.out <- lm(LogPass2013 ~ Pass2012, data=reportCard)
summary(lm.log.out)
```

```
##
## Call:
## lm(formula = LogPass2013 ~ Pass2012, data = reportCard)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3880 -0.2531  0.0776  0.3461  2.7368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.307692   0.046030   6.685 3.37e-11 ***
## Pass2012    0.039491   0.000699  56.499  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5915 on 1360 degrees of freedom
## Multiple R-squared:  0.7012,    Adjusted R-squared:  0.701
## F-statistic:  3192 on 1 and 1360 DF,  p-value: < 2.2e-16
```

EPSY 630
Spring 2021

Did the passing rates drop in a predictable manner?

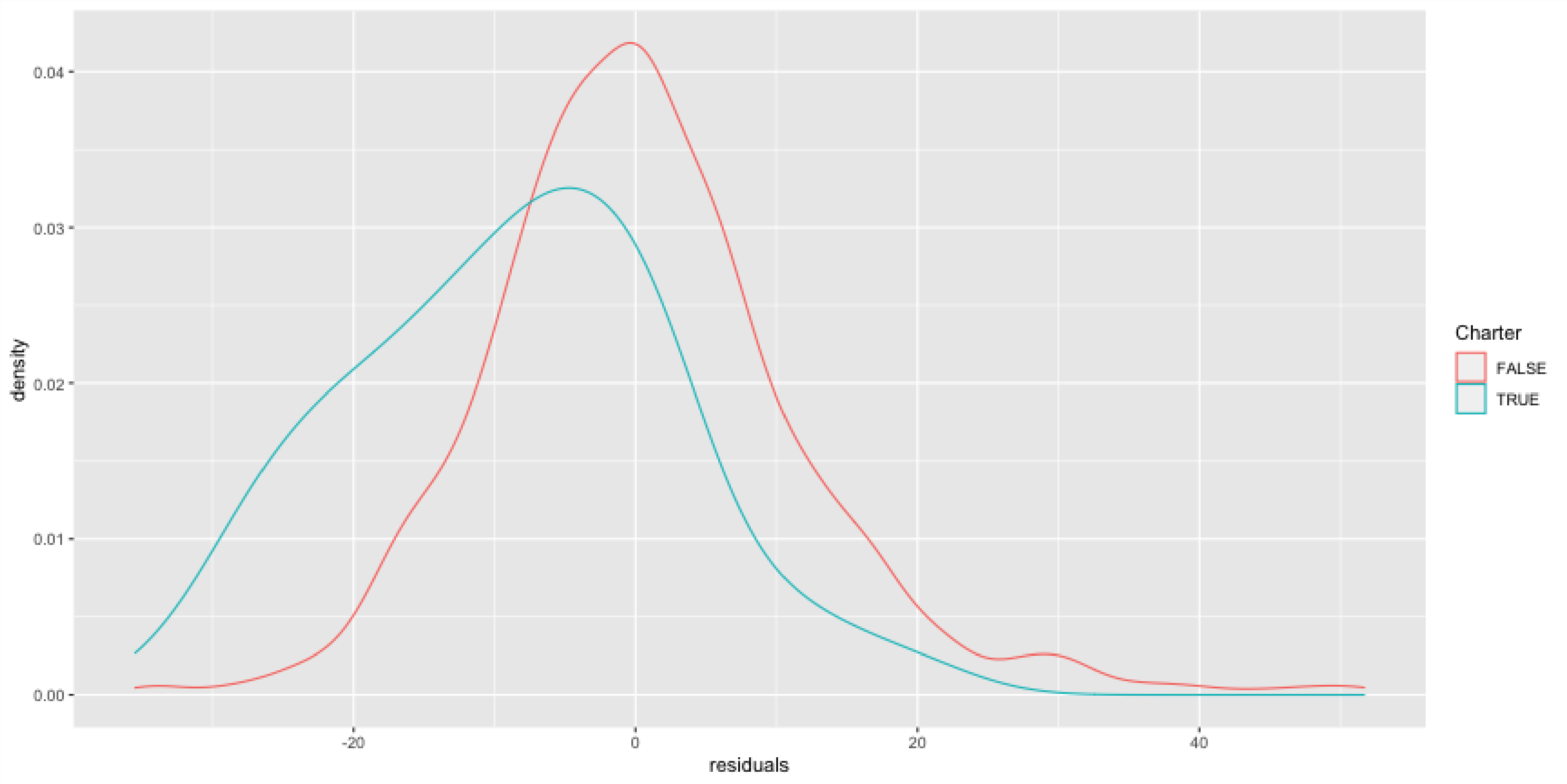
Yes! Whether we log tranform the data or not, the correlations are statistically significant with regression models with R^2 creater than 62%.

To answer the second question, whether the drops were different for public and charter schools, we'll look at the residuals.

```
reportCard$residuals <- resid(lm.out)
reportCard$residualsLog <- resid(lm.log.out)
```

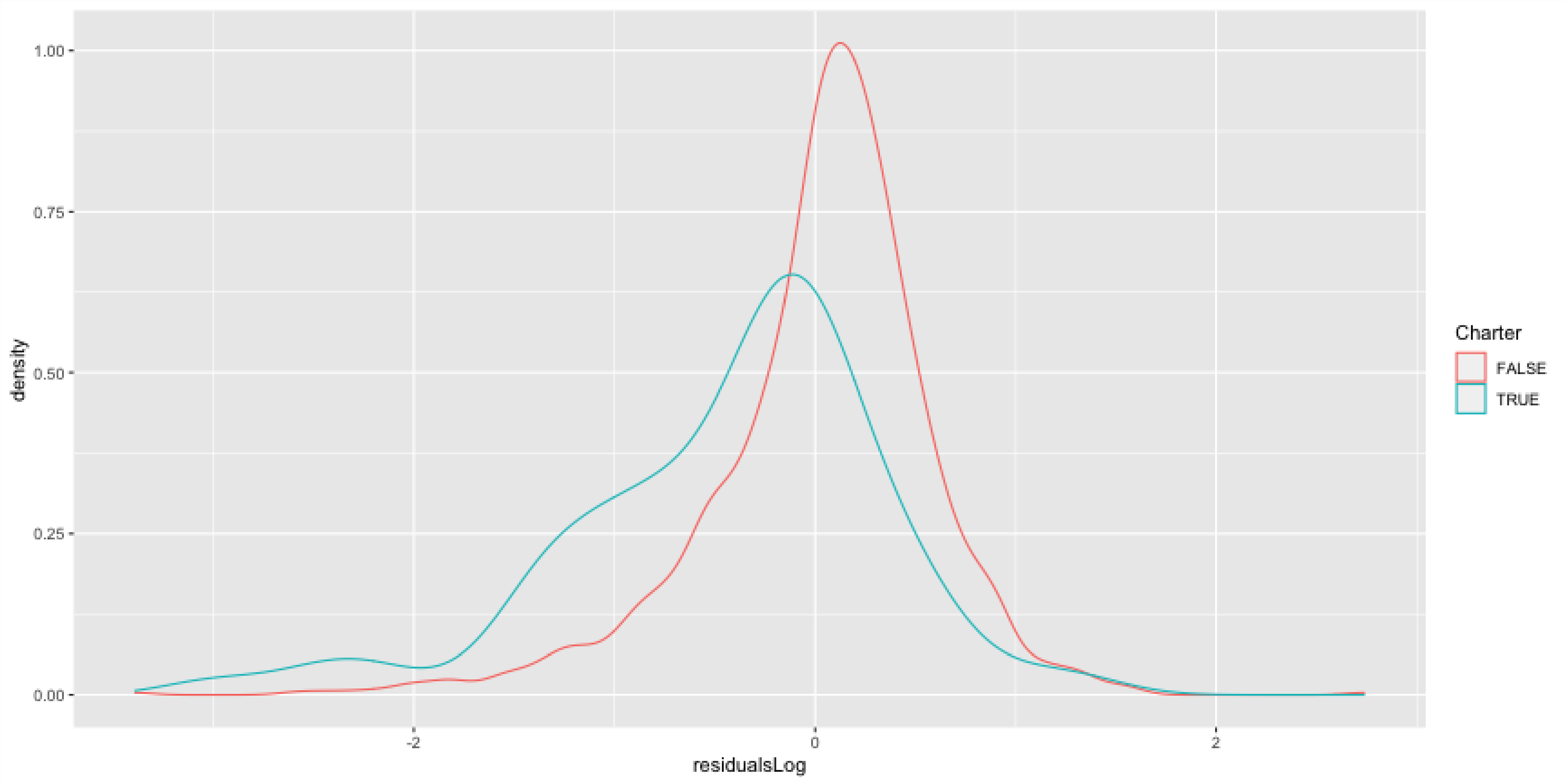
Distribution of Residuals

```
ggplot(reportCard, aes(x=residuals, color=Charter)) + geom_density()
```



Distribution of Residuals

```
ggplot(reportCard, aes(x=residualsLog, color=Charter)) + geom_density()
```



EPSY 630 Spring 2021

Null Hypothesis Testing

H_0: There is no difference in the residuals between charter and public schools.

H_A: There is a difference in the residuals between charter and public schools.

```
t.test(residuals ~ Charter, data=reportCard)
```

```
##
##      Welch Two Sample t-test
##
## data:  residuals by Charter
## t = 6.5751, df = 77.633, p-value = 5.091e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.411064 11.980002
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            0.479356           -8.716177
```

Null Hypothesis Testing (log transform)

```
t.test(residualsLog ~ Charter, data=reportCard)
```

```
##
##      Welch Two Sample t-test
##
## data:  residualsLog by Charter
## t = 4.7957, df = 74.136, p-value = 8.161e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2642811 0.6399761
## sample estimates:
## mean in group FALSE  mean in group TRUE
##          0.02356911         -0.42855946
```

EPSY 630
Spring 2021

Polynomial Models (e.g. Quadratic)

It is possible to fit quatric models fairly easily in R, say of the following form:

$$y = b_1 x^2 + b_2 x + b_0$$

```
quad.out <- lm(Pass2013 ~ I(Pass2012^2) + Pass2012, data=reportCard)
summary(quad.out)$r.squared
```

```
## [1] 0.7065206
```

```
summary(lm.out)$r.squared
```

```
## [1] 0.6206049
```

EPSY 630
Spring 2021

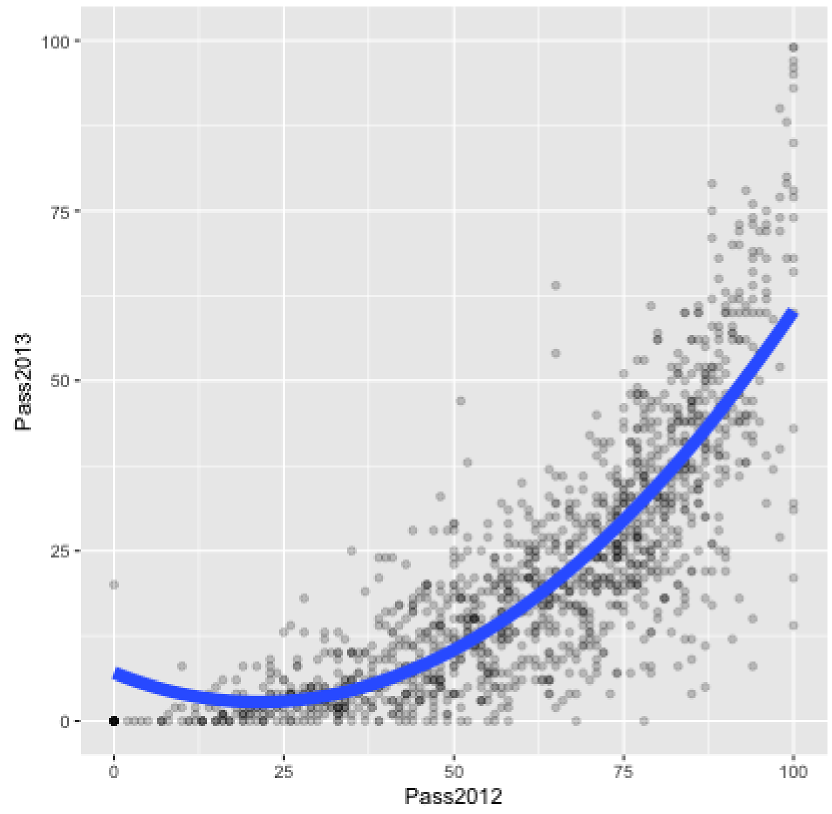
Quadratic Model

```
summary(quad.out)
```

```
##
## Call:
## lm(formula = Pass2013 ~ I(Pass2012^2) + Pass2012, data = reportCard)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.258  -4.906  -0.507   5.430  43.509
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.0466153  1.4263773   4.940 8.77e-07 ***
## I(Pass2012^2)  0.0092937  0.0004659  19.946  < 2e-16 ***
## Pass2012      -0.3972481  0.0533631  -7.444 1.72e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.11 on 1359 degrees of freedom
## Multiple R-squared:  0.7065,    Adjusted R-squared:  0.7061
## F-statistic:  1636 on 2 and 1359 DF,  p-value: < 2.2e-16
```

Scatter Plot

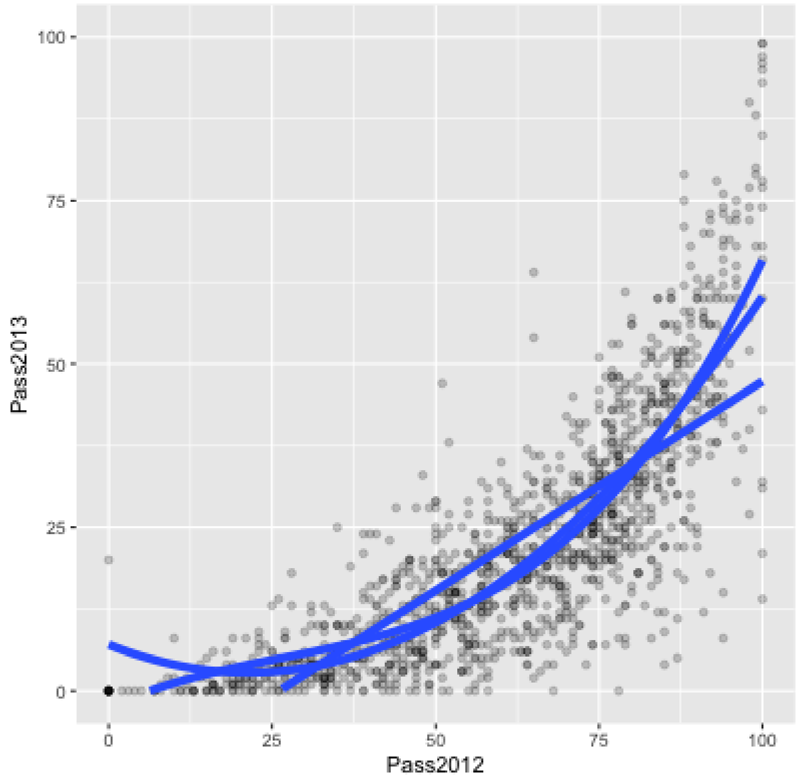
```
ggplot(reportCard, aes(x=Pass2012, y=Pass2013)) + geom_point(alpha=0.2) +
    geom_smooth(method='lm', formula=y~poly(x,2,raw=TRUE), size=3, se=FALSE) +
    coord_equal() + ylim(c(0,100)) + xlim(c(0,100))
```



Let's go crazy, cubic!

```
cube.out <- lm(Pass2013 ~ I(Pass2012^3) + I(Pass2012^2) + Pass2012, data=reportCard)
summary(cube.out)$r.squared
```

```
## [1] 0.7168206
```



Shiny App

```
shiny::runGitHub('NYSchools','jbryer',subdir='NYSReportCard')
```

See also the Github repository for more information: https://github.com/jbryer/NYSchools

Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA)

The goal of ANOVA is to test whether there is a discernible difference between the means of several groups.

EPSY 630
Spring 2021

Example

Is there a difference between washing hands with: water only, regular soap, antibacterial soap (ABS), and antibacterial spray (AS)?

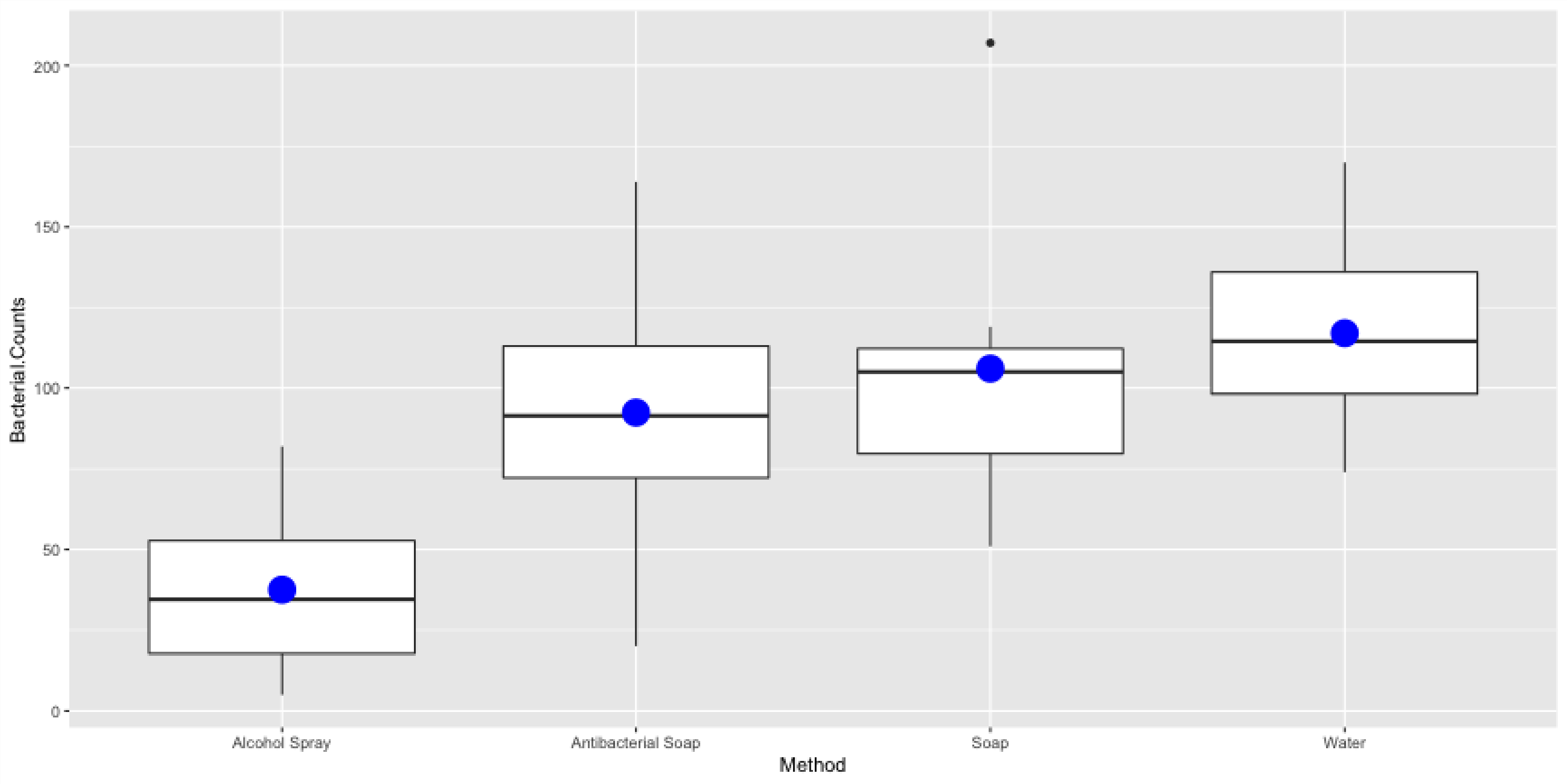
- Each tested with 8 replications
- Treatments randomly assigned

For ANOVA:

- The means all differ.
- Is this just natural variability?
- Null hypothesis: All the means are the same.
- Alternative hypothesis: The means are not all the same.

Hand Washing Comparison

```
ggplot(hand, aes(x=Method, y=Bacterial.Counts)) + geom_boxplot() +
    stat_summary(fun = mean, color = 'blue', size = 1.5)
```



Hand Washing Comparison (cont.)

```
desc <- describeBy(hand$Bacterial.Counts, hand$Method, mat=TRUE)[,c(2,4,5,6)]
desc$Var <- desc$sd^2
print(desc, row.names=FALSE)
```

```
##              group1 n  mean       sd       Var
##       Alcohol Spray 8  37.5 26.55991  705.4286
##  Antibacterial Soap 8  92.5 41.96257 1760.8571
##                Soap 8 106.0 46.95895 2205.1429
##               Water 8 117.0 31.13106  969.1429
```

Washing type all the same?

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
- By Central Limit Theorem:

$$Var(\bar{y}) = \frac{\sigma^2}{n} = \frac{\sigma^2}{8}$$

- Variance of {26.56, 41.96, 46.96, 31.13} is 1410.14.
- $\frac{\sigma^2}{8} = 1410.14$
- $\sigma^2 = 9960.64$
- This estimate for σ^2 is called the Treatment Mean Square, Between Mean Square, or MS_T
- Is this very high compared to what we would expect?

EPSY 630
Spring 2021

How can we decide what σ^2 should be?

- Assume each washing method has the same variance.
- Then we can pool them all together to get the pooled variance s_p^2
- Since the sample sizes are all equal, we can average the four variances: $s_p^2 = 1410.14$
- Other names for s_p^2: Error Mean Square, Within Mean Square, MS_E.

EPSY 630
Spring 2021

Comparing MS_T (between) and MS_E (within)

MS_T

- Estimates s^2 if H_0 is true
- Should be larger than s^2 if H_0 is false

MS_E

- Estimates s^2 whether H_0 is true or not
- If H_0 is true, both close to s^2, so MS_T is close to MS_E

Comparing

- If H_0 is true, $\frac{MS_T}{MS_E}$ should be close to 1
- If H_0 is false, $\frac{MS_T}{MS_E}$ tends to be > 1

The F-Distribution

- How do we tell whether $\frac{MS_T}{MS_E}$ is larger enough to not be due just to random chance
- $\frac{MS_T}{MS_E}$ follows the F-Distribution
  - Numerator df: k - 1 (k = number of groups)
  - Denominator df: k(n - 1)
  - n = # observations in each group
- $F = \frac{MS_T}{MS_E}$ is called the F-Statistic.

A Shiny App by Dr. Dudek to explore the F-Distribution: https://shiny.rit.albany.edu/stat/fdist/

EPSY 630
Spring 2021

The F-Distribution (cont.)

```
df.numerator <- 4 - 1
df.denominator <- 4 * (8 - 1)
plot(function(x)(df(x,df1=df.numerator,df2=df.denominator)),
     xlim=c(0,5), xlab='x', ylab='f(x)', main='F-Distribution')
```

Back to Bacteria

Source	Sum of Squares	*df*	MS	F	p
Between Group (Factor)	$\sum_k n_k(\bar{x}_k - \bar{x})^2$	k - 1	$\frac{SS_{between}}{df_{between}}$	$\frac{MS_{between}}{MS_{within}}$	area to right of $F_{k-1,n-k}$
Within Group (Error)	$\sum_k \sum_i (\bar{x}_{ik} - \bar{x}_k)^2$	n - k	$\frac{SS_{within}}{df_{within}}$		
Total	$\sum_k \sum_i (\bar{x}_{ik} - \bar{x})^2$	n - 1			

EPSY 630 Spring 2021

ANOVA Steps

```r
(grand.mean <- mean(hand$Bacterial.Counts))
```

```
## [1] 88.25
```

```r
(n <- nrow(hand))
```

```
## [1] 32
```

```r
(k <- length(unique(hand$Method)))
```

```
## [1] 4
```

```r
(ss.total <- sum((hand$Bacterial.Counts - grand.mean)^2))
```

```
## [1] 69366
```

ANOVA Steps

Between Groups

```
(df.between <- k - 1)
```

```
## [1] 3
```

```
(ss.between <- sum(desc$n * (desc$mean - grand.mean)^2))
```

```
## [1] 29882
```

```
(MS.between <- ss.between / df.between)
```

```
## [1] 9960.667
```

Within Groups

```
(df.within <- n - k)
```

```
## [1] 28
```

```
(ss.within <- ss.total - ss.between)
```

```
## [1] 39484
```

```
(MS.within <- ss.within / df.within)
```

```
## [1] 1410.143
```

F Statistic

- $MS_T = 9960.67$
- $MS_E = 1410.14$
- Numerator df = 4 - 1 = 3
- Denominator df = 4(8 - 1) = 28.

```r
(f.stat <- 9960.64 / 1410.14)
```

```
## [1] 7.063582
```

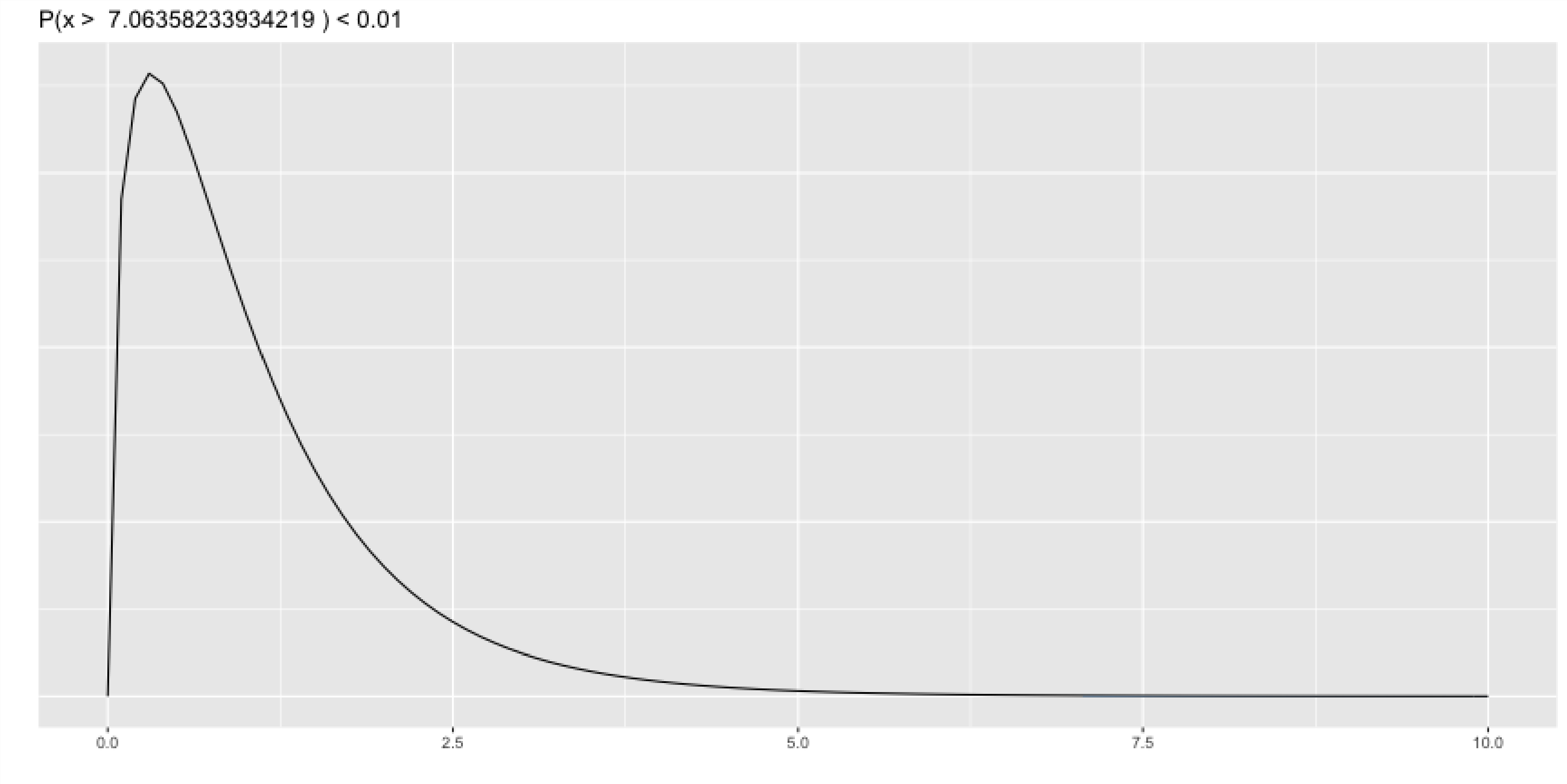
```r
1 - pf(f.stat, 3, 28)
```

```
## [1] 0.001111464
```

P-value for $F_{3,28} = 0.0011$

F Distribution

```
DATA606::F_plot(df.numerator, df.denominator, cv = f.stat)
```



Assumptions and Conditions

- To check the assumptions and conditions for ANOVA, always look at the side-by-side boxplots.
  - Check for outliers within any group.
  - Check for similar spreads.
  - Look for skewness.
  - Consider re-expressing.
- Independence Assumption
  - Groups must be independent of each other.
  - Data within each group must be independent.
  - Randomization Condition
- Equal Variance Assumption
  - In ANOVA, we pool the variances. This requires equal variances from each group: Similar Spread Condition.

EPSY 630
Spring 2021

ANOVA in R

```
aov.out <- aov(Bacterial.Counts ~ Method, data=hand)
summary(aov.out)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Method       3  29882    9961   7.064 0.00111 **
## Residuals   28  39484    1410
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Graphical ANOVA

```r
hand.anova <- granova.1w(hand$Bacterial.Counts, group=hand$Method)
```

Graphical ANOVA

```r
hand.anova
```

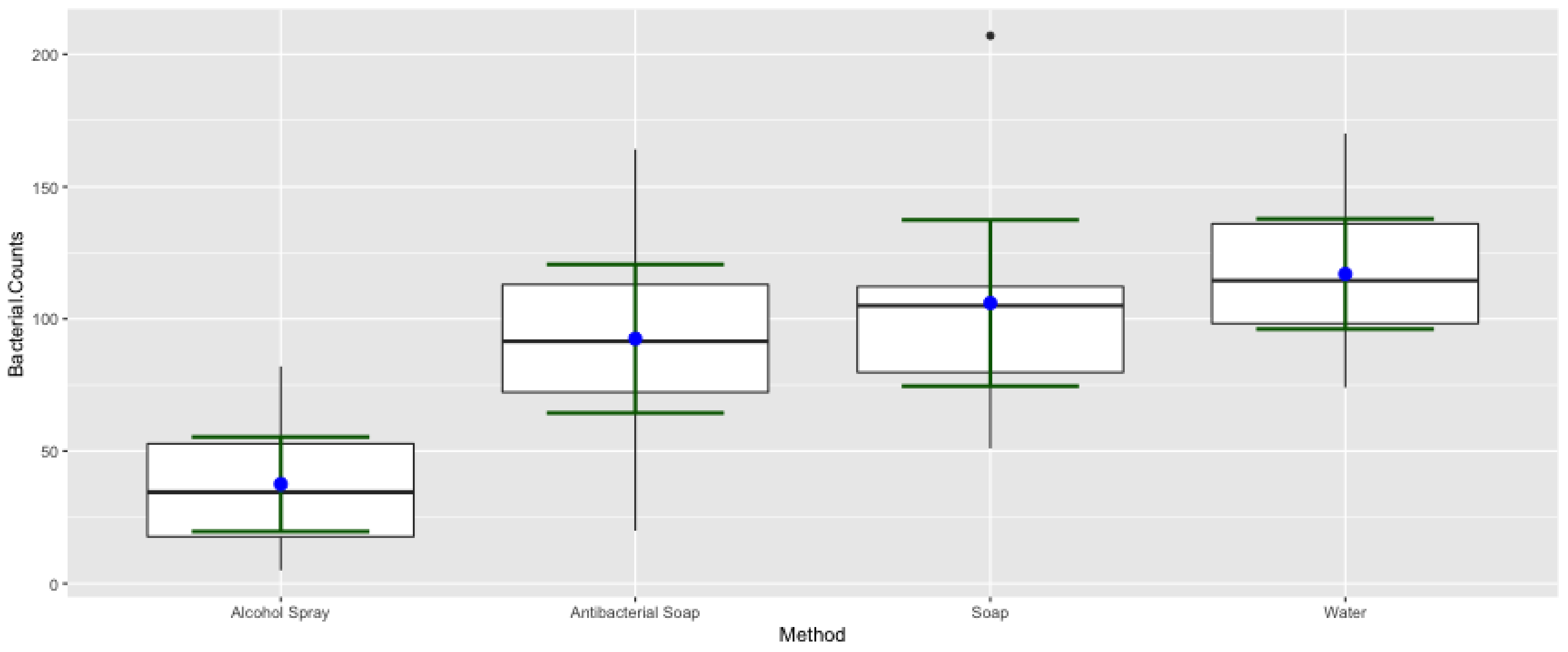
```
## $grandsum
##      Grandmean           df.bet          df.with          MS.bet          MS.with          F.stat
##          88.25             3.00            28.00         9960.67          1410.14            7.06
##         F.prob SS.bet/SS.tot
##           0.00          0.43
##
## $stats
##                      Size Contrast Coef Wt'd Mean  Mean Trim'd Mean    Var. St. Dev.
## Alcohol Spray           8        -50.75      37.5  37.5        35.50  705.43    26.56
## Antibacterial Soap      8          4.25      92.5  92.5        92.67 1760.86    41.96
## Soap                    8         17.75     106.0 106.0        98.33 2205.14    46.96
## Water                   8         28.75     117.0 117.0       115.33  969.14    31.13
```

What Next?

- P-value large -> Nothing left to say
- P-value small -> Which means are large and which means are small?
- We can perform a t-test to compare two of them.
- We assumed the standard deviations are all equal.
- Use s_p, for pooled standard deviations.
- Use the Students t-model, df = N - k.
- If we wanted to do a t-test for each pair:
  - P(Type I Error) = 0.05 for each test.
  - Good chance at least one will have a Type I error.
- **Bonferroni to the rescue!**
  - Adjust a to α/J where J is the number of comparisons.
  - 95% confidence (1 - 0.05) with 3 comparisons adjusts to $(1 - 0.05/3) \approx 0.98333$.
  - Use this adjusted value to find t**.

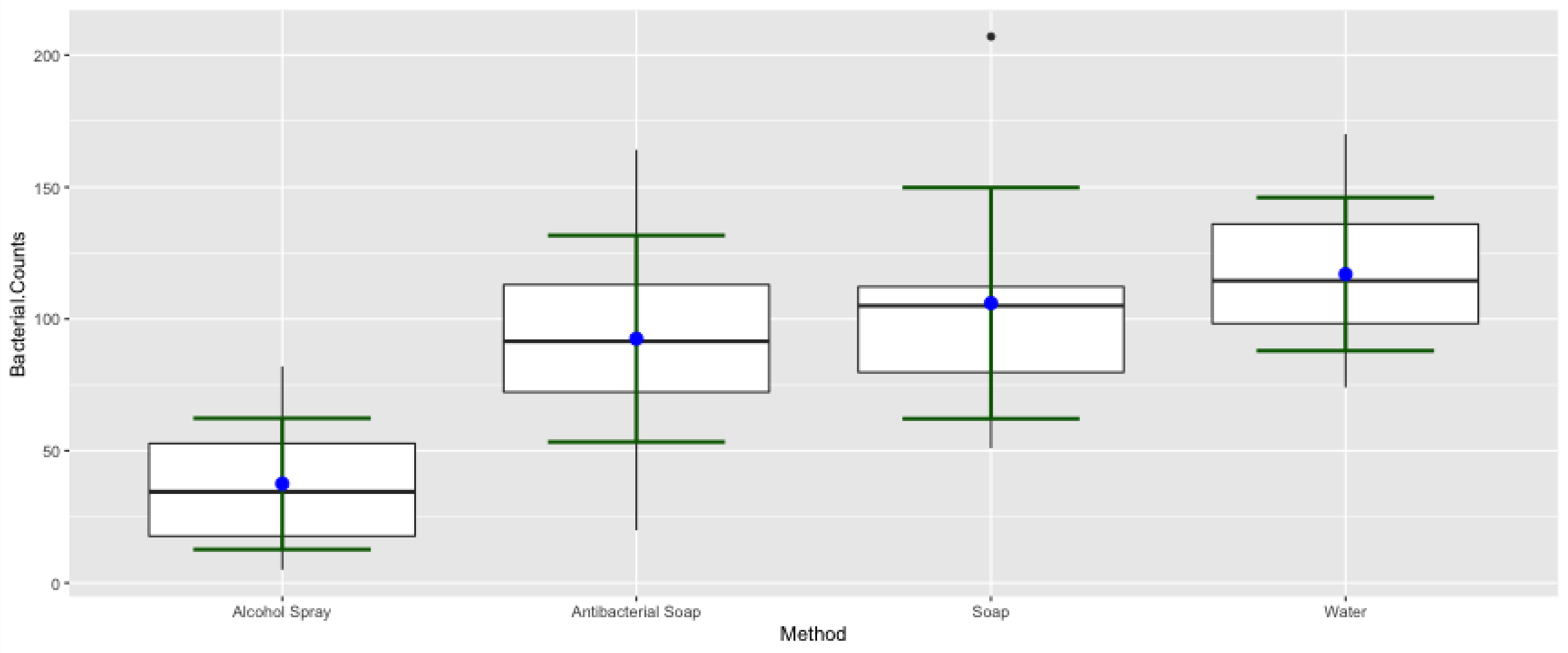
Multiple Comparisons (no Bonferroni adjustment)

```
cv <- qt(0.05, df = 7)
tab <- describeBy(hand$Bacterial.Counts, group = hand$Method, mat = TRUE)
ggplot(hand, aes(x = Method, y = Bacterial.Counts)) + geom_boxplot() +
    geom_errorbar(data = tab, aes(x = group1, y = mean,
                              ymin = mean - cv * se, ymax = mean + cv * se),
                  color = 'darkgreen', width = 0.5, size = 1) +
    geom_point(data = tab, aes(x = group1, y = mean), color = 'blue', size = 3)
```



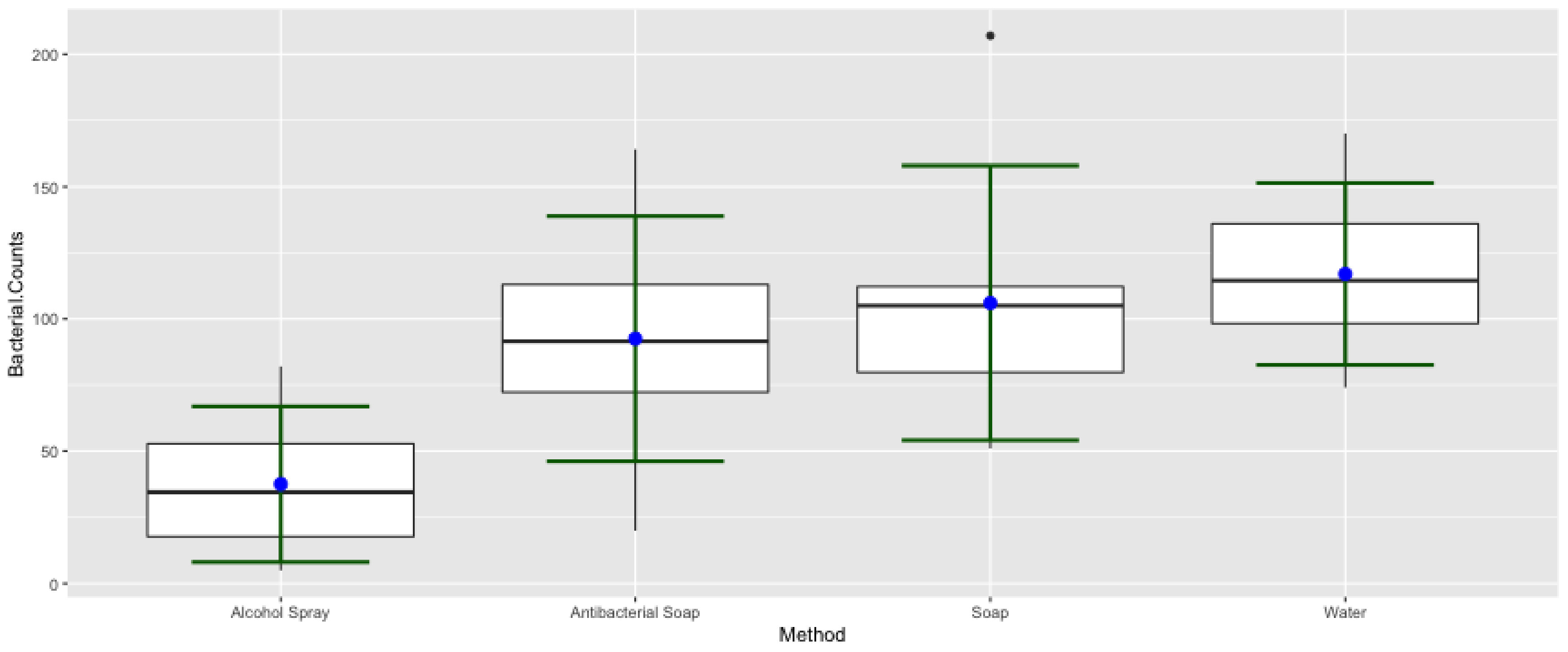
Multiple Comparisons (3 paired tests)

```r
cv <- qt(0.05 / 3, df = 7)
tab <- describeBy(hand$Bacterial.Counts, group = hand$Method, mat = TRUE)
ggplot(hand, aes(x = Method, y = Bacterial.Counts)) + geom_boxplot() +
    geom_errorbar(data = tab, aes(x = group1, y = mean,
                                  ymin = mean - cv * se, ymax = mean + cv * se),
                  color = 'darkgreen', width = 0.5, size = 1) +
    geom_point(data = tab, aes(x = group1, y = mean), color = 'blue', size = 3)
```



Multiple Comparisons (6 paired tests)

```
cv <- qt(0.05 / choose(4, 2), df = 7)
tab <- describeBy(hand$Bacterial.Counts, group = hand$Method, mat = TRUE)
ggplot(hand, aes(x = Method, y = Bacterial.Counts)) + geom_boxplot() +
    geom_errorbar(data = tab, aes(x = group1, y = mean,
                                  ymin = mean - cv * se, ymax = mean + cv * se ),
                  color = 'darkgreen', width = 0.5, size = 1) +
    geom_point(data = tab, aes(x = group1, y = mean), color = 'blue', size = 3)
```



Assignments

ANOVA lab.

```
DATA606::startLab('Lab7b') # https://r.bryer.org/shiny/Lab7a/
```

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?