# Inference for Numerical Data

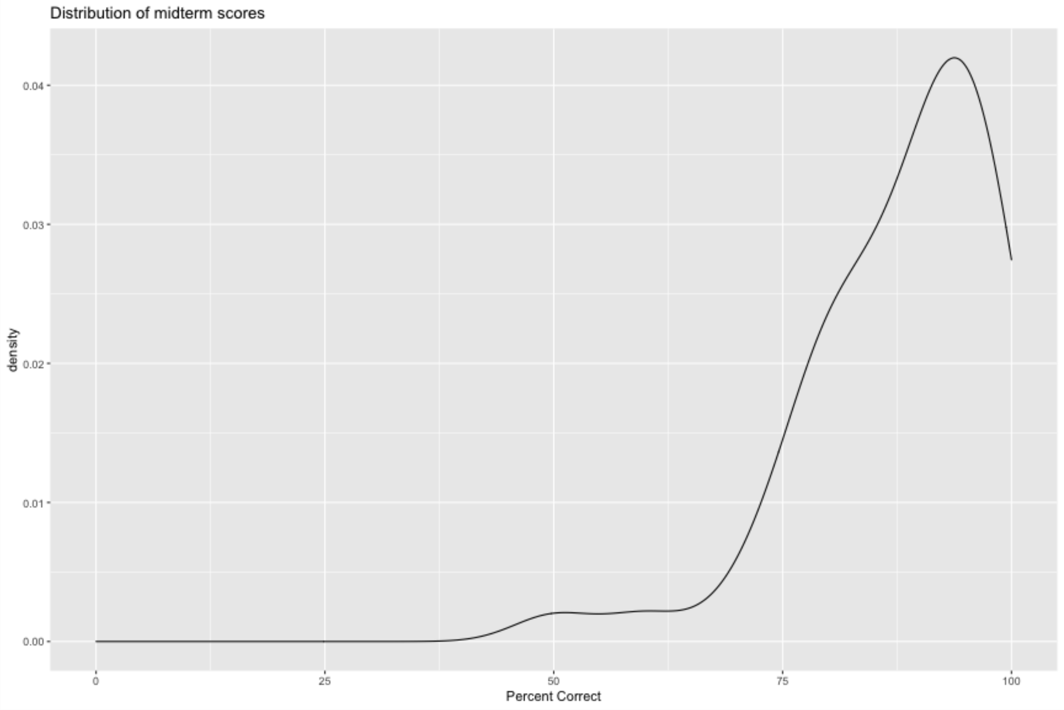
## DATA 606 - Statistics & Probability for Data Analytics

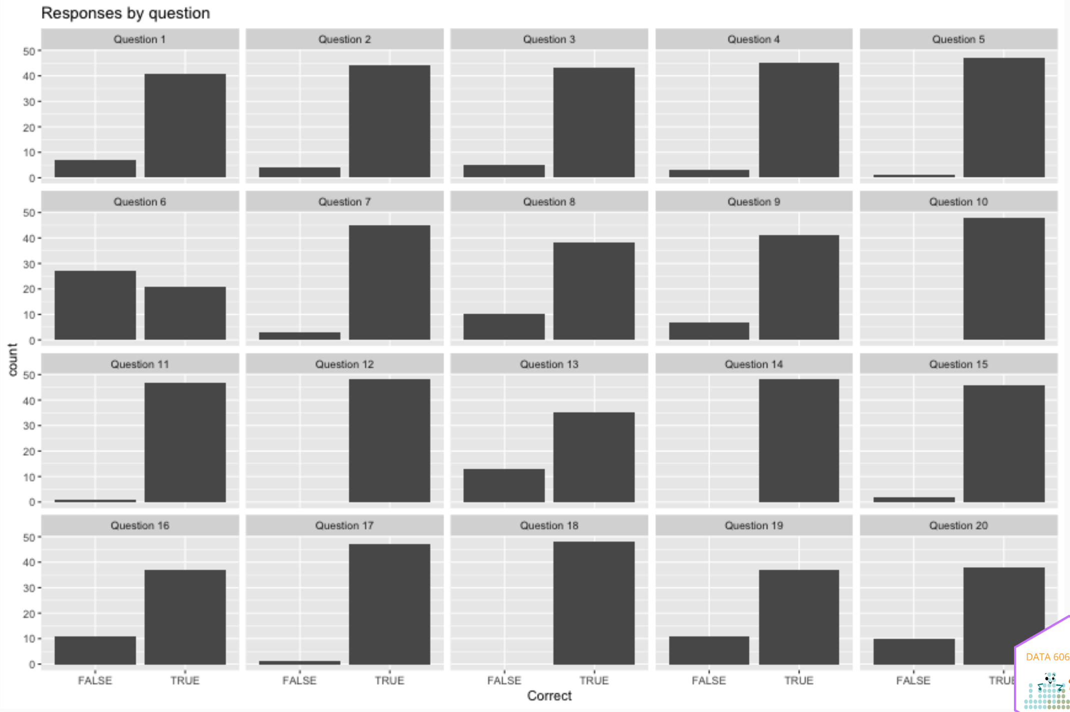
Jason Bryer, Ph.D. and Angela Lui, Ph.D.

October 19, 2022

# Midterm Results

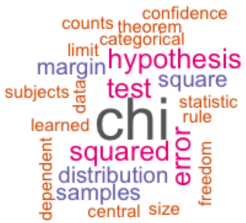
- Results on Blackboard range between 0 and 200. Can divide by 2 to get a percentage score.
- The midterm is worth 10% of your grade.
- We don't want to share the questions and answers but will provide feedback privately (if you ask questions) and will review questions in the next week or two.



# One Minute Paper Results

### What was the most important thing you learned during this class?

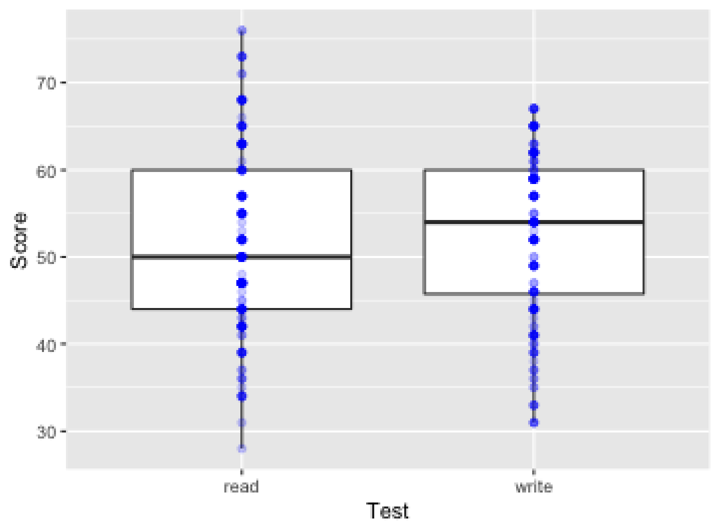



### What important question remains unanswered for you?

# 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

```r
head(hsb2)
```

```
## # A tibble: 6 × 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?

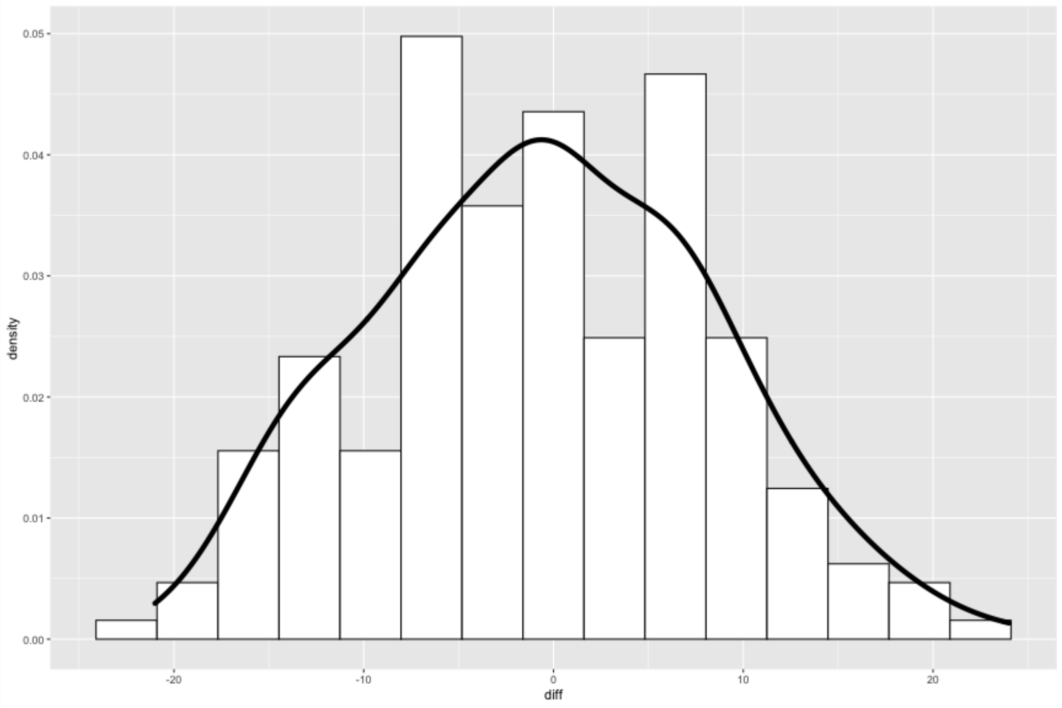
# Analyzing Paired Data

- 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
```

```
ggplot(hsb2, aes(x = diff)) +
    geom_histogram(aes(y = ..density..), bins = 15, col
    geom_density(size = 2)
```



# 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.

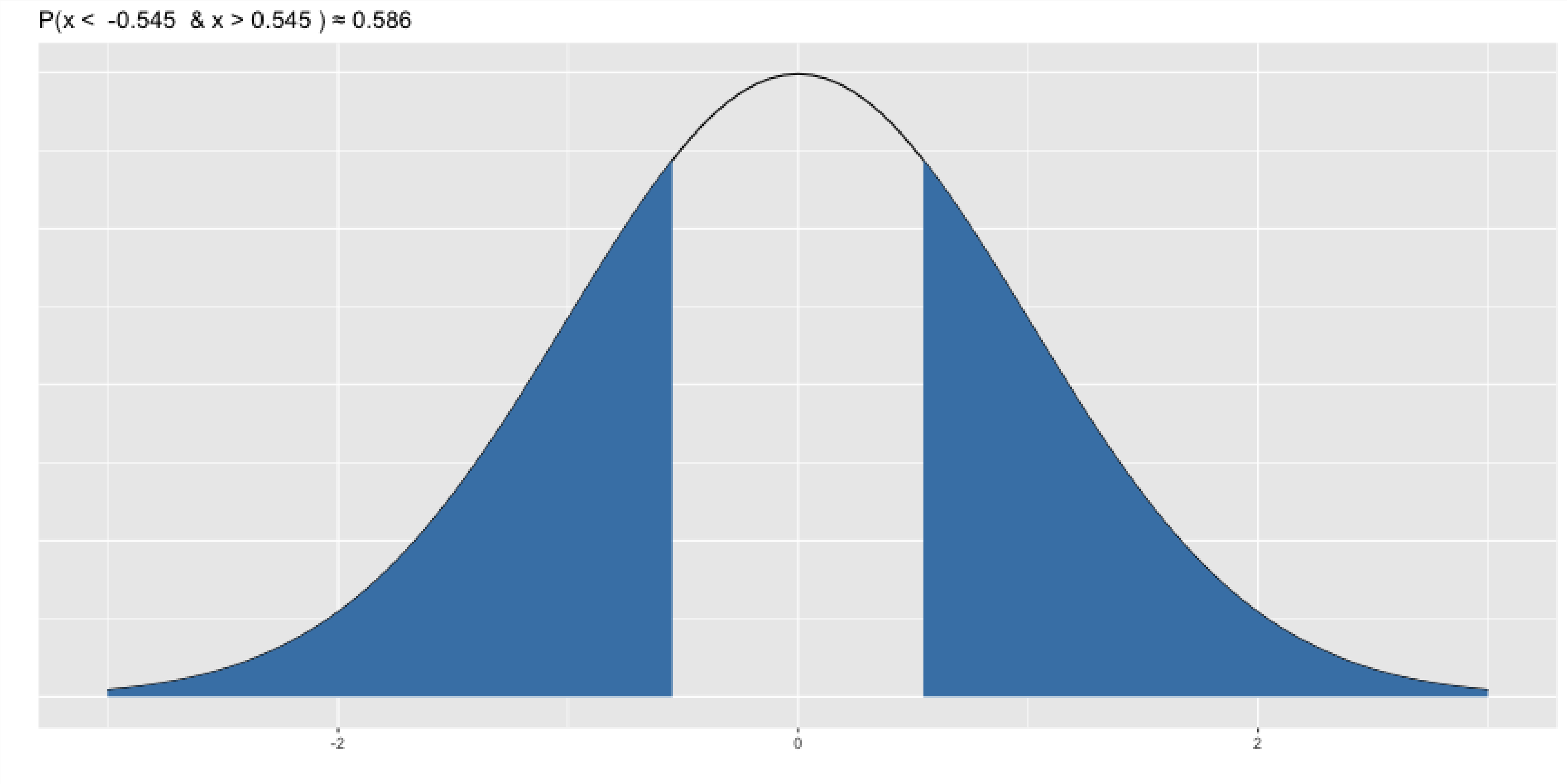
$$\mu_{diff} \neq 0$$

# 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 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the 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
```

# Evaluating the null hypothesis

## 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%.

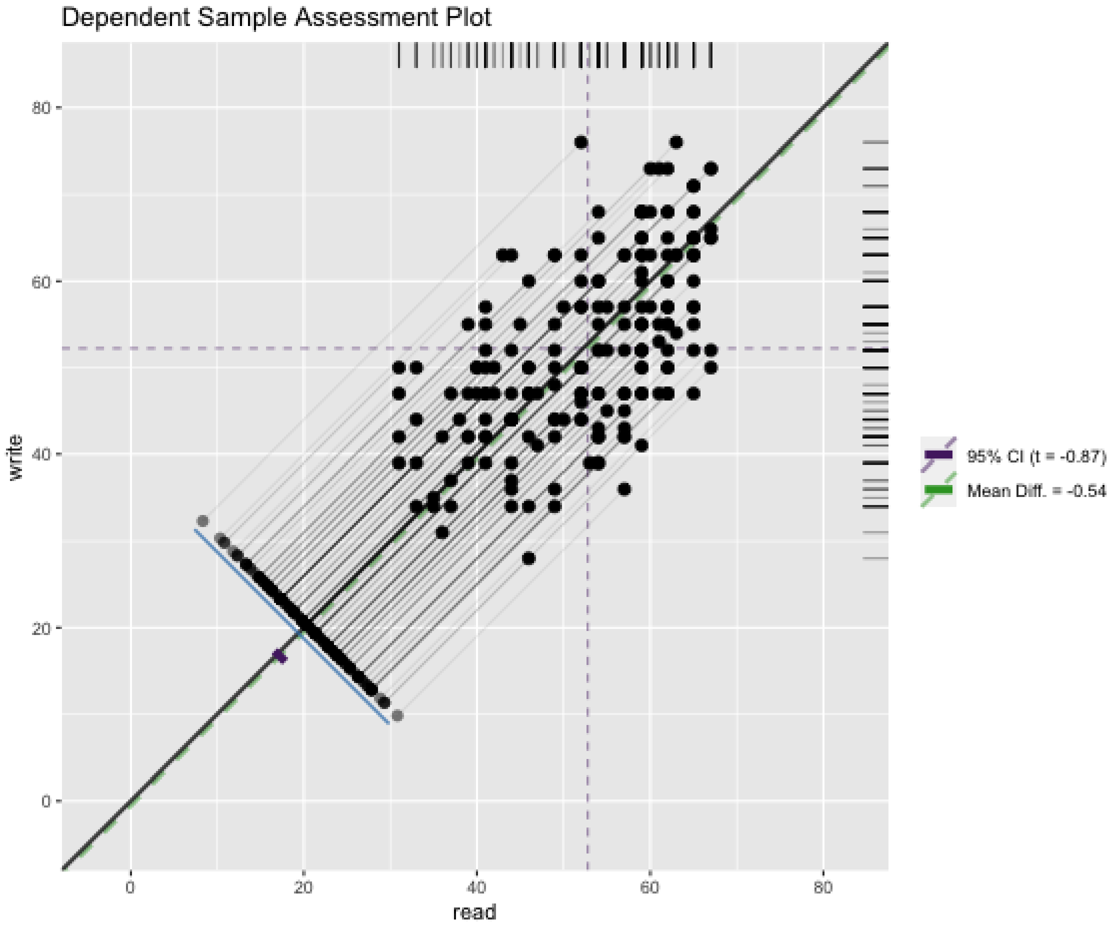
## Calculating 95% Confidence Interval

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

Note that the confidence interval spans zero!

# Visualizing Dependent Sample Tests

```r
library(granovaGG)
granovagg.ds(as.data.frame(hsb2[,c('read', 'write')]))
```



# SAT Scores by Sex

```r
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
```

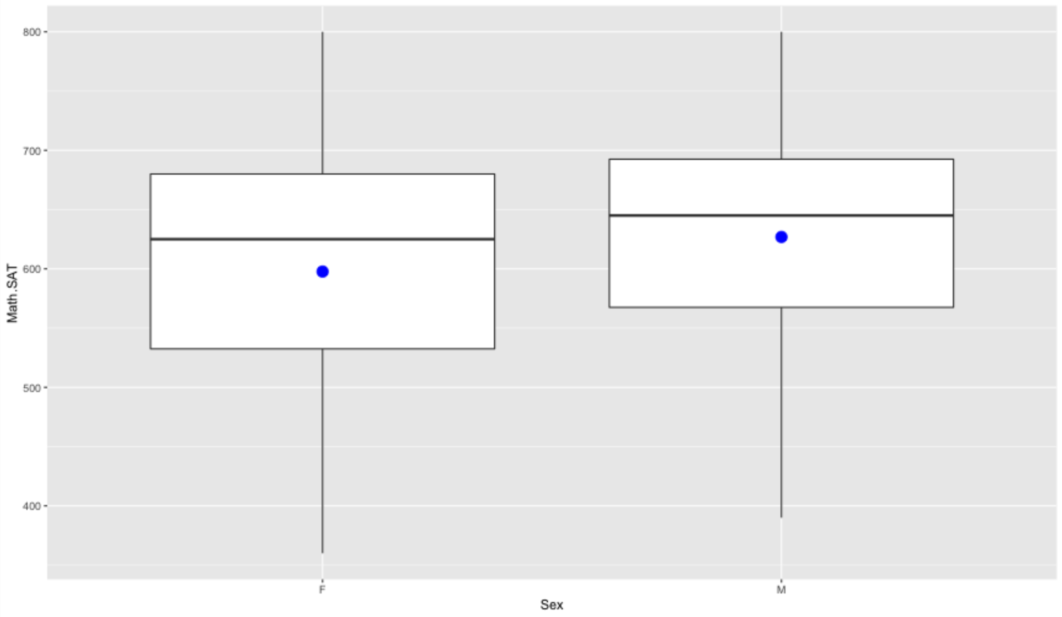
Is there a difference in math scores between males and females?

# SAT Scores by Sex

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

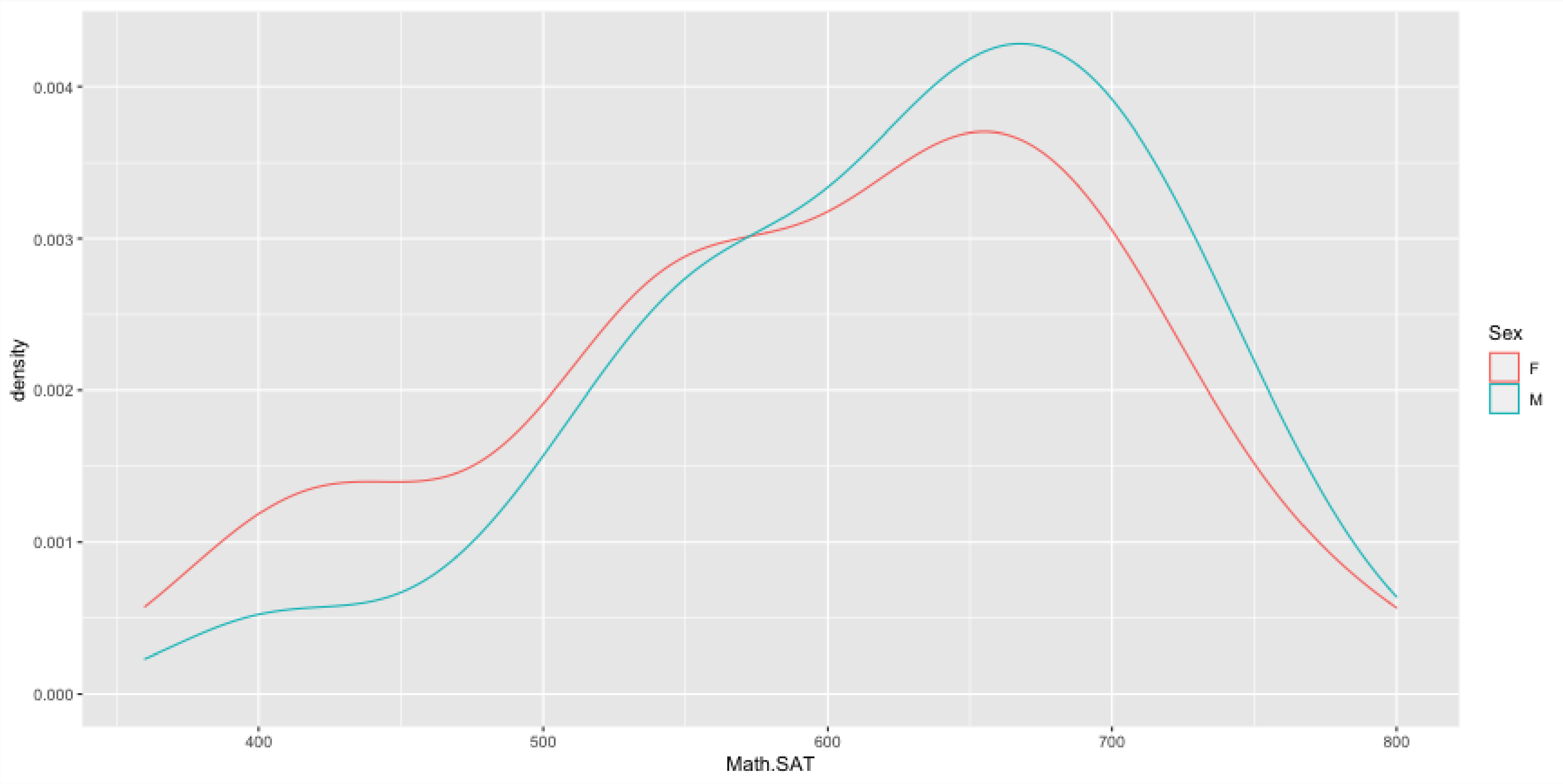
```
##     group1  n     mean        sd min
## X11      F 82 597.6829 103.70065 360
## X12      M 80 626.8750  90.35225 390
```

```
ggplot(sat, aes(x=Sex, y=Math.SAT)) +
    geom_boxplot() +
    geom_point(data = tab, aes(x=group1, y=mean),
               color='blue', size=4)
```



# Distributions

```
ggplot(sat, aes(x=Math.SAT, color = Sex)) + geom_density()
```



# 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 ± ME
- And all ME = critical value * 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 for difference in SAT scores

$$SE_{(\bar{x}_M - \bar{x}_F)} = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}}$$

$$SE_{(\bar{x}_M - \bar{x}_F)} = \sqrt{\frac{90.4}{80} + \frac{103.7}{82}} = 1.55$$

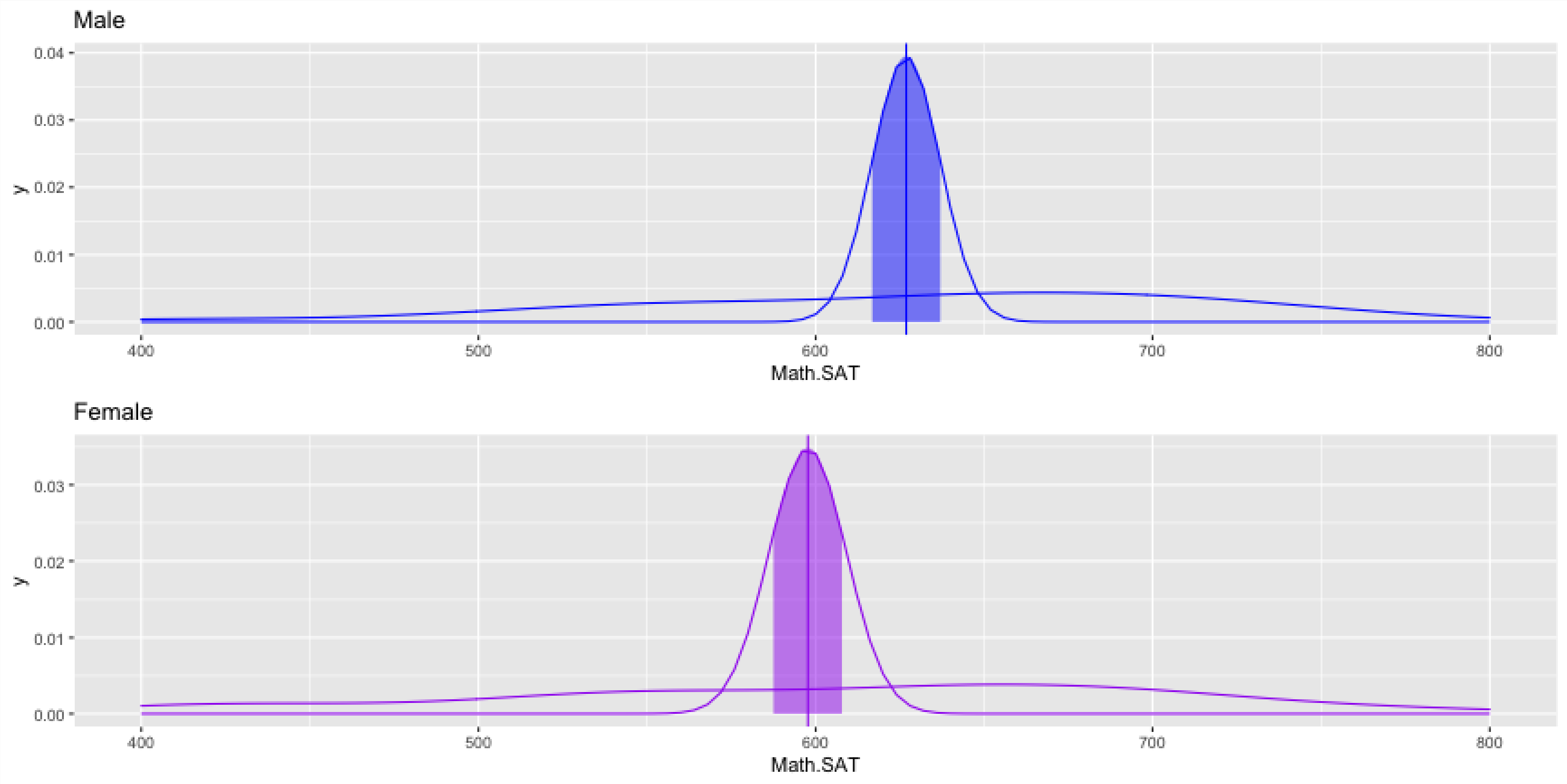
Calculate the 95% confidence interval:

$$(\bar{x}_M - \bar{x}_F) \pm 1.96 SE_{(\bar{x}_M - \bar{x}_F)}$$

$$(626.9 - 597.7) \pm 1.96 \times 1.55$$

$$29.2 \pm 3.038 = (36.162, 42.238)$$

# Visualizing independent sample tests



# What about smaller sample sizes?

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

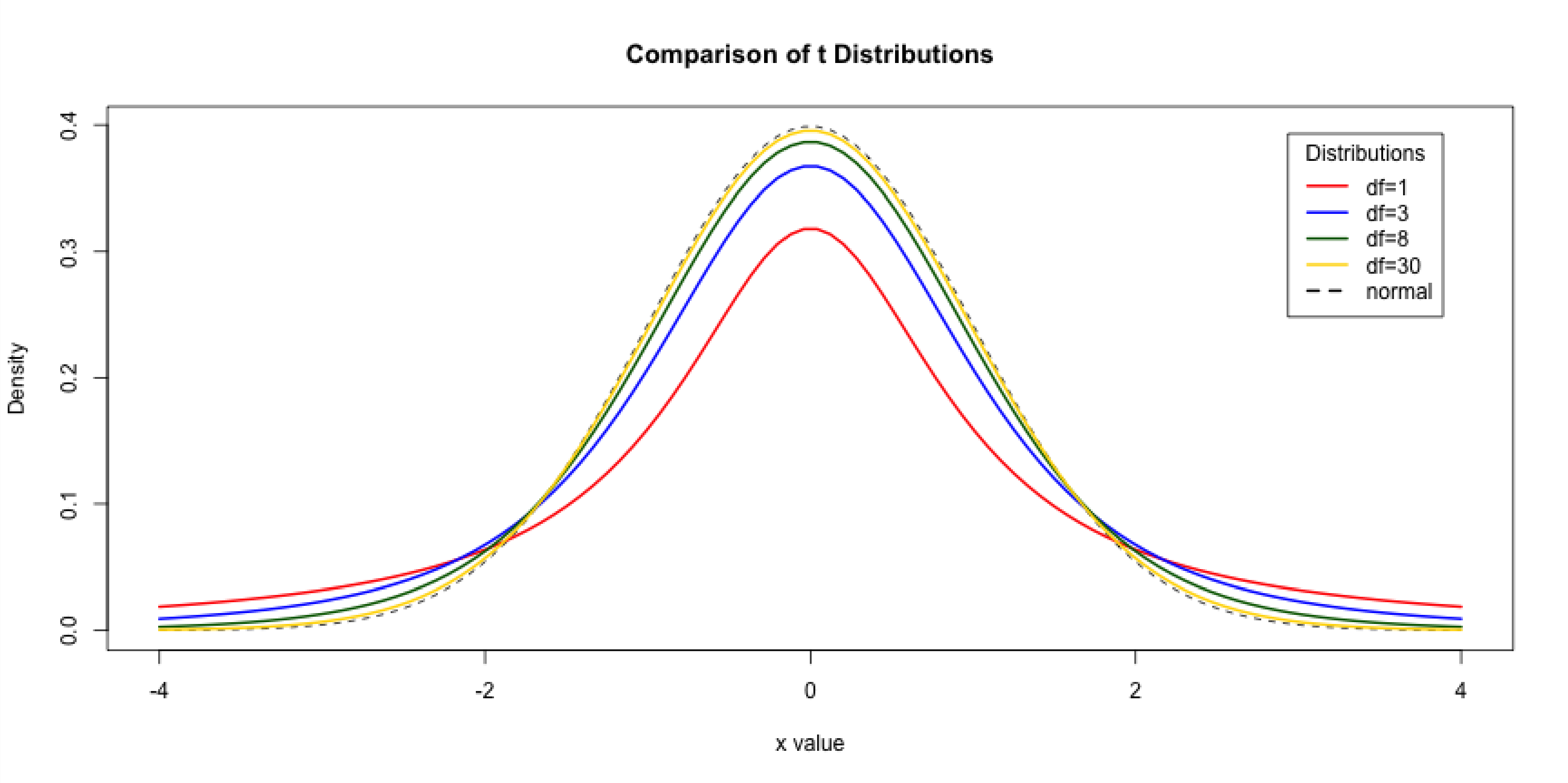
- Sample sizes necessarily need to be small.
- The CLT states that the sampling distribution approximates normal as n -> Infinity
- Need an alternative to the normal distribution.
- The *t* distribution was developed by William Gosset (under the pseudonym *student*) to estimate means when the sample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

Confidence interval is estimated using

$$\overline{x} \pm t^*_{df} SE$$

Where *df* is the degrees of freedom (*df* = *n* -1)

# *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 = 0.05773
## alternative hypothesis: true difference in means bet
## 95 percent confidence interval:
##  -59.3527145   0.9685682
## sample estimates:
## mean in group F mean in group M 
##        597.6829        626.8750
```

# Analysis of Variancne (ANOVA)

# Analysis of Variance (ANOVA)

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

**Hand Washing 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

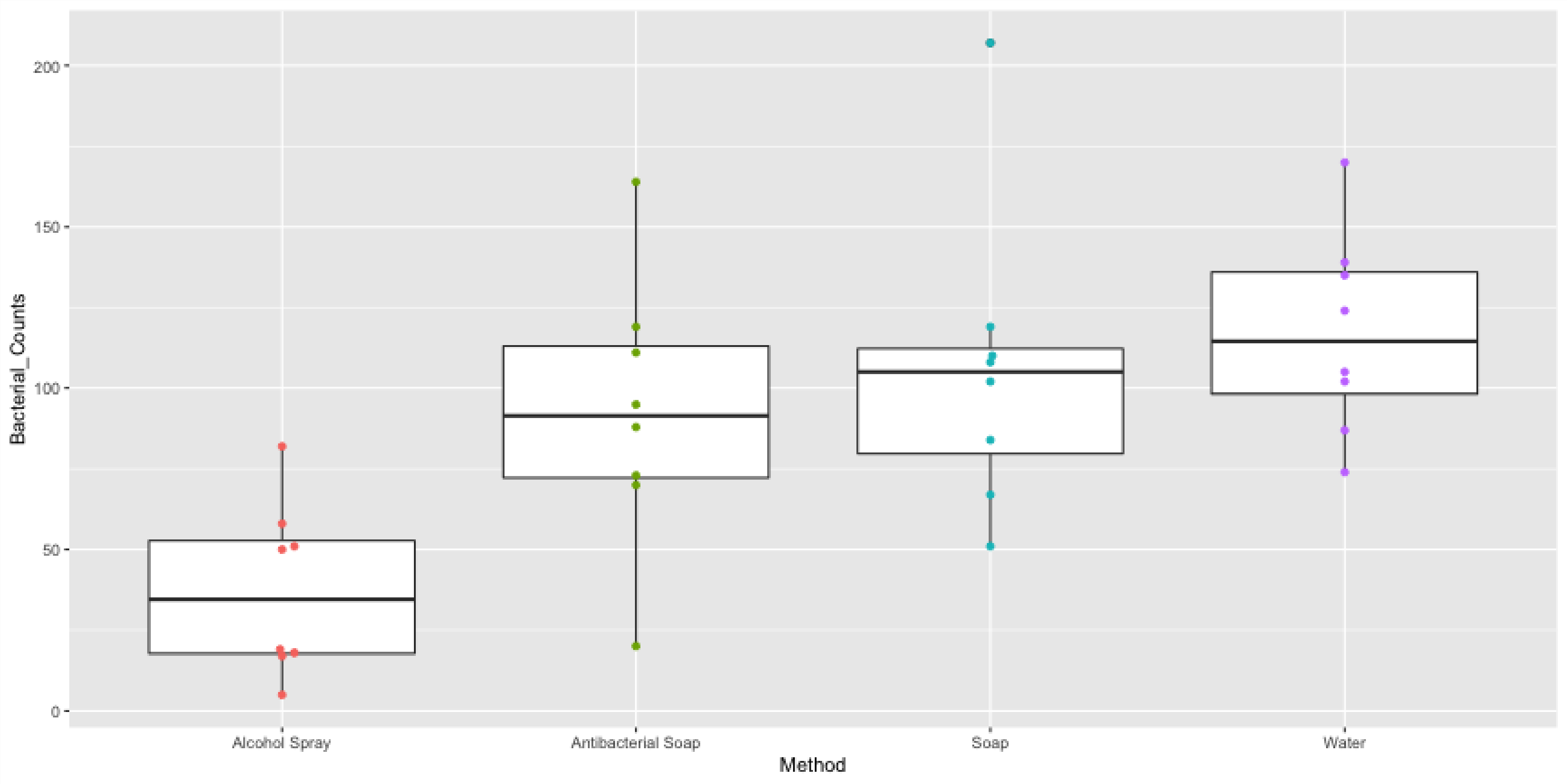
For ANOVA:

- 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.

Source: De Veaux, R.D., Velleman, P.F., & Bock, D.E. (2014). *Intro Stats, 4th Ed.* Pearson.

# Boxplot

```r
ggplot(hand_washing, aes(x = Method, y = Bacterial_Counts)) +  geom_boxplot() +
    geom_beeswarm(aes(color = Method)) + theme(legend.position = 'none')
```



# Descriptive Statistics

```
desc <- psych::describeBy(hand_washing$Bacterial_Counts, group = hand_washing$Method, mat = TRUE, skew = FALSE)
names(desc)[2] <- 'Method' # Rename the grouping column
desc$Var <- desc$sd^2 # We will need the variance latter, so calculate it here
desc
```

```
##     item              Method vars n  mean       sd min max range        se
## X11    1       Alcohol Spray    1 8  37.5 26.55991   5  82    77  9.390345
## X12    2 Antibacterial Soap    1 8  92.5 41.96257  20 164   144 14.836008
## X13    3                Soap    1 8 106.0 46.95895  51 207   156 16.602496
## X14    4               Water    1 8 117.0 31.13106  74 170    96 11.006492
##           Var
## X11  705.4286
## X12 1760.8571
## X13 2205.1429
## X14  969.1429
```

```
( k <- length(unique(hand_washing$Method)) )
```

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

```
( n <- nrow(hand_washing) )
```

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

```
( grand_mean <- mean(hand_washing$Bacterial_Counts) )
```

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

```
( grand_var <- var(hand_washing$Bacterial_Counts) )
```

```
## [1] 2237.613
```

# Contrasts

A contrast is a linear combination of two or more factor level means with coefficients that sum to zero.

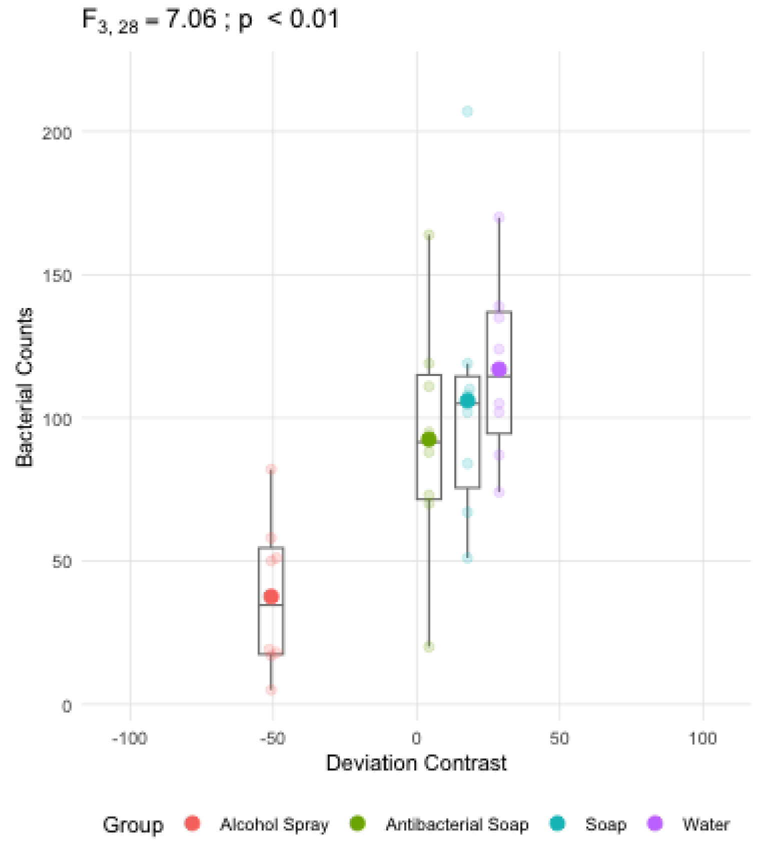
```
desc$contrast <- (desc$mean - mean(desc$mean))
mean(desc$contrast) # Should be 0!
```

```
## [1] 0
```

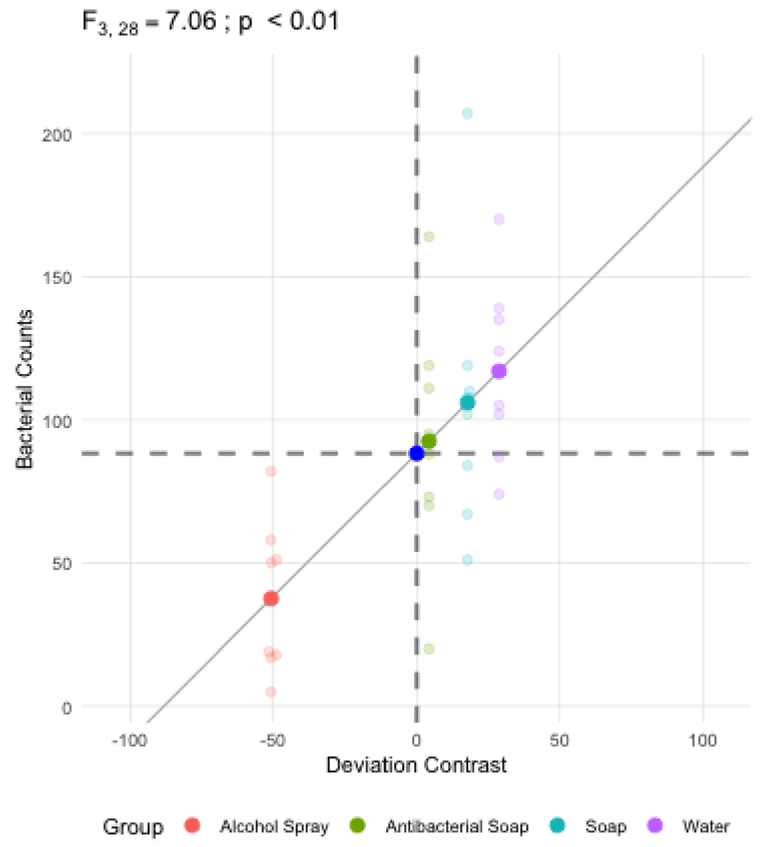
```
desc
```

```
##     item              Method vars n  mean       sd min max range        se
## X11    1       Alcohol Spray    1 8  37.5 26.55991   5  82    77  9.390345
## X12    2 Antibacterial Soap    1 8  92.5 41.96257  20 164   144 14.836008
## X13    3                Soap    1 8 106.0 46.95895  51 207   156 16.602496
## X14    4               Water    1 8 117.0 31.13106  74 170    96 11.006492
##           Var contrast
## X11  705.4286   -50.75
## X12 1760.8571     4.25
## X13 2205.1429    17.75
## X14  969.1429    28.75
```

# Plotting using contrasts



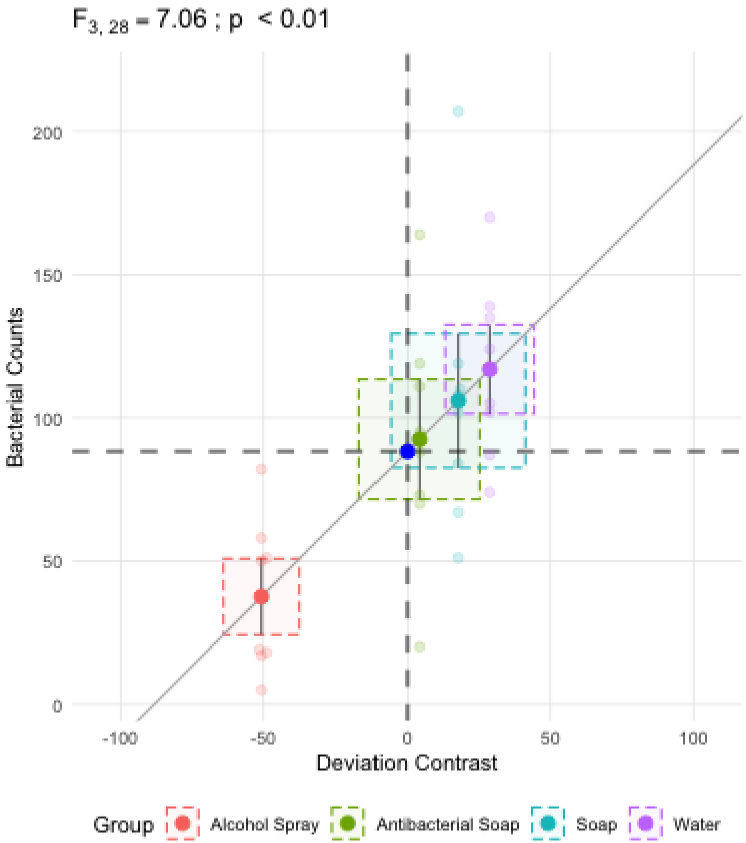
# Grade Mean and Unit Line (slope = 1, intercept = $\bar{x}$)



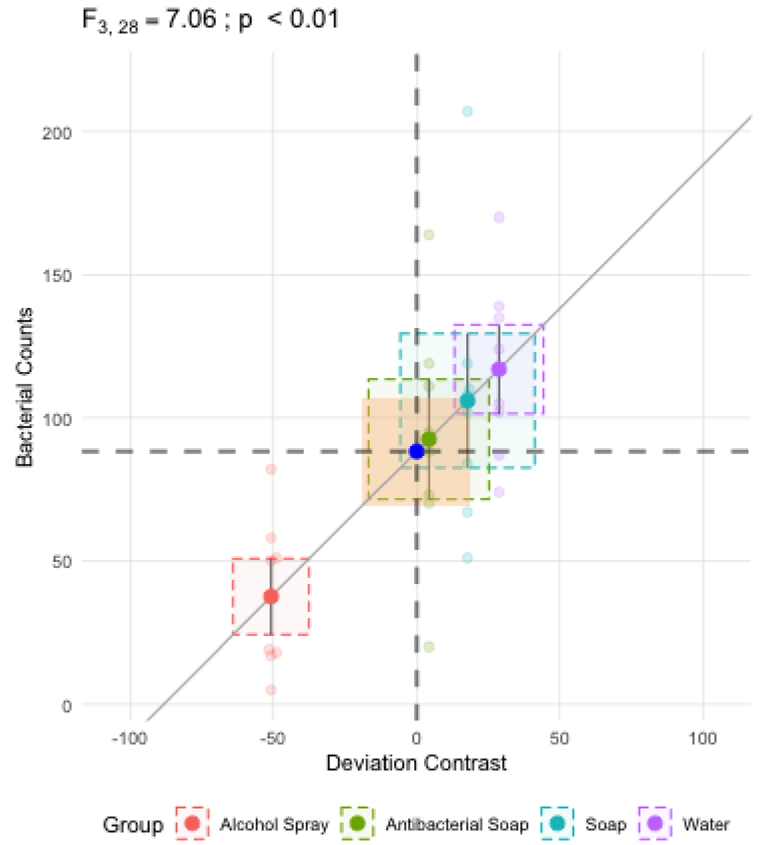
# Within Group Variance (error)

$$SS_{within} = \sum_k \sum_i (\bar{x}_{ik} - \bar{x}_k)^2$$

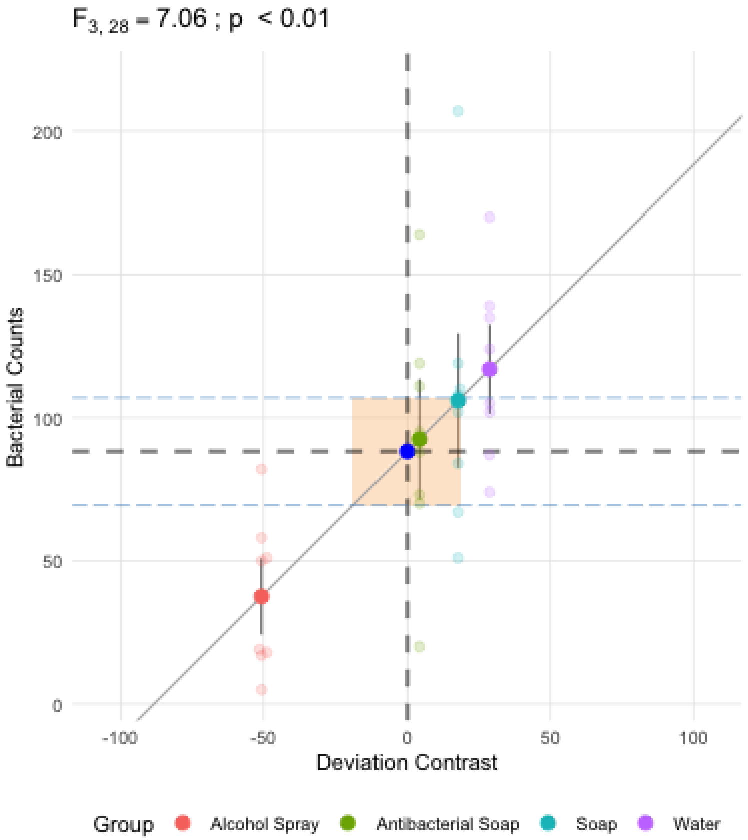
# Within Group Variance (error)



# Within Group Variance (error)



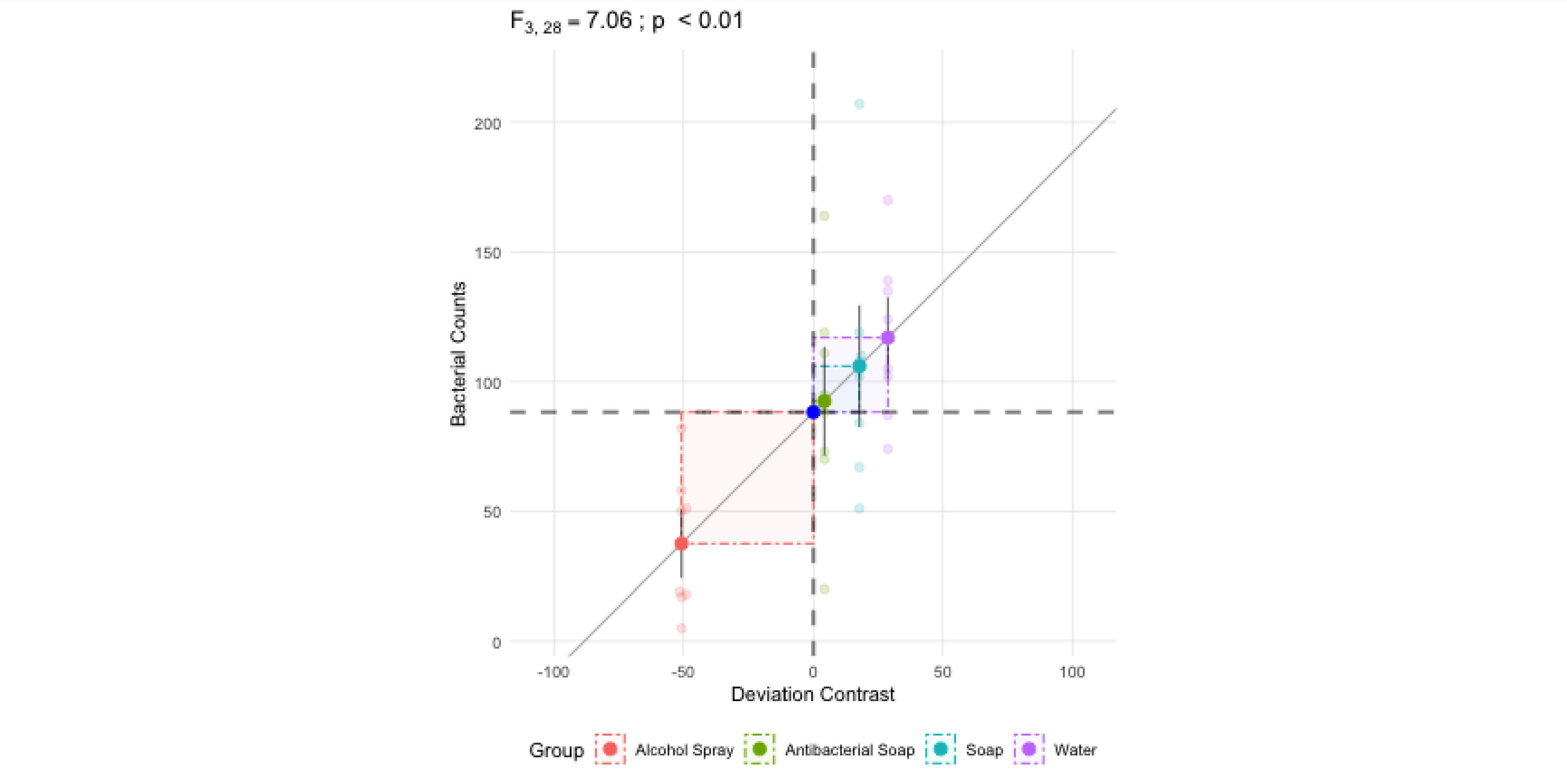
# Within Group Variance (error)



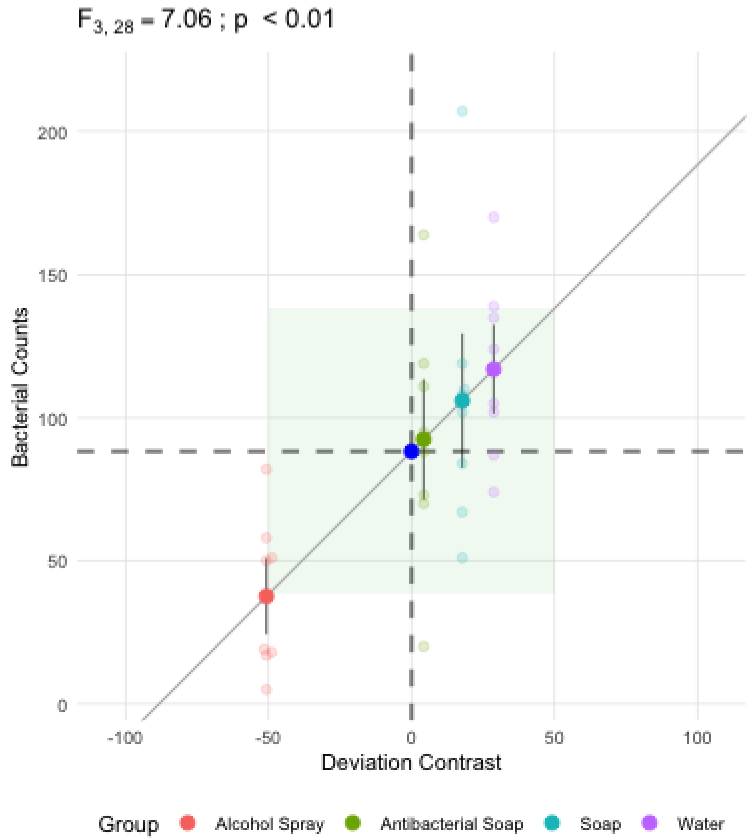
# Between Group Variance

$$SS_{between} = \sum_k n_k(\bar{x}_k - \bar{x})^2$$

# Between Group Variance



# Between Group Variance

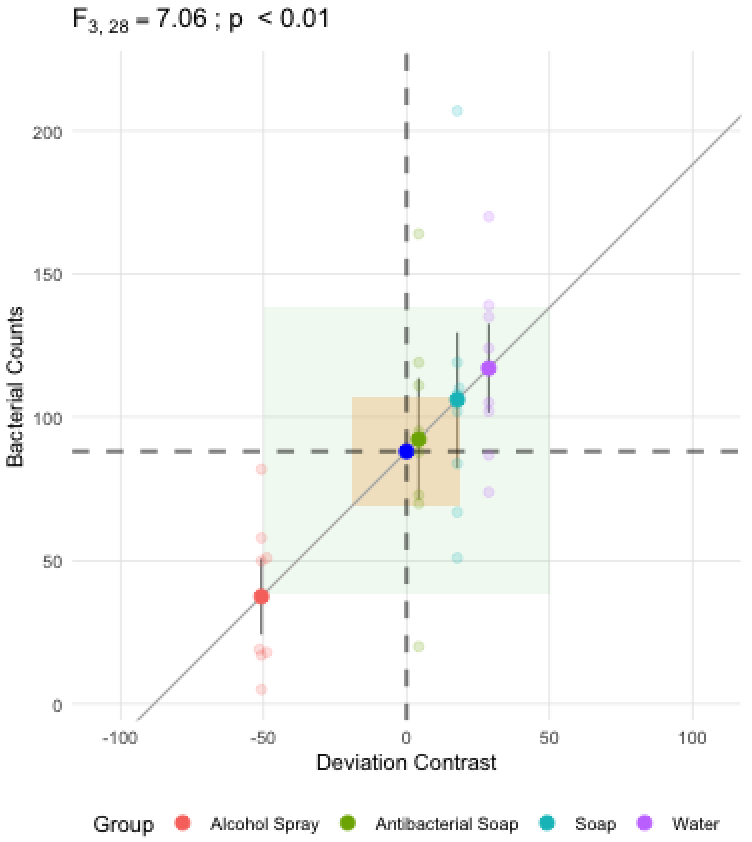


# Mean Square

| Source | Sum of Squares | *df* | MS |
|---|---|---|---|
| Between Group (Treatment) | $\sum_k n_k(\bar{x}_k - \bar{x})^2$ | k - 1 | $\frac{SS_{between}}{df_{between}}$ |
| 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 | |

# $MS_{Between}/MS_{Within}$ = F-Statistic

Mean squares can be represented as squares, hence the ratio of area of the two rectagles is equal to $\frac{MS_{Between}}{MS_{Within}}$ which is the F-statistic.



# Washing type all the same?

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Variance components we need to evaluate the null hypothesis:

- Between Sum of Squares: $SS_{between} = \sum_k n_k(\bar{x}_k - \bar{x})^2$
- Within Sum of Squares: $SS_{within} = \sum_k \sum_i (\bar{x}_{ik} - \bar{x}_k)^2$
- Between degrees of freedom: $df_{between} = k - 1$ (k = number of groups)
- Within degrees of freedom: $df_{within} = k(n - 1)$
- Mean square between (aka treatment): $MS_T = \frac{SS_{between}}{df_{between}}$
- Mean square within (aka error): $MS_E = \frac{SS_{within}}{df_{within}}$

# Comparing $MS_T$ (between) and $MS_E$ (within)

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: $s_p^2 = 1410.14$

```
mean(desc$Var)
```

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

$MS_T$

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

$MS_E$

- Estimates $s_p^2$ whether $H_0$ is true or not
- If $H_0$ is true, both close to $s_p^2$, 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?

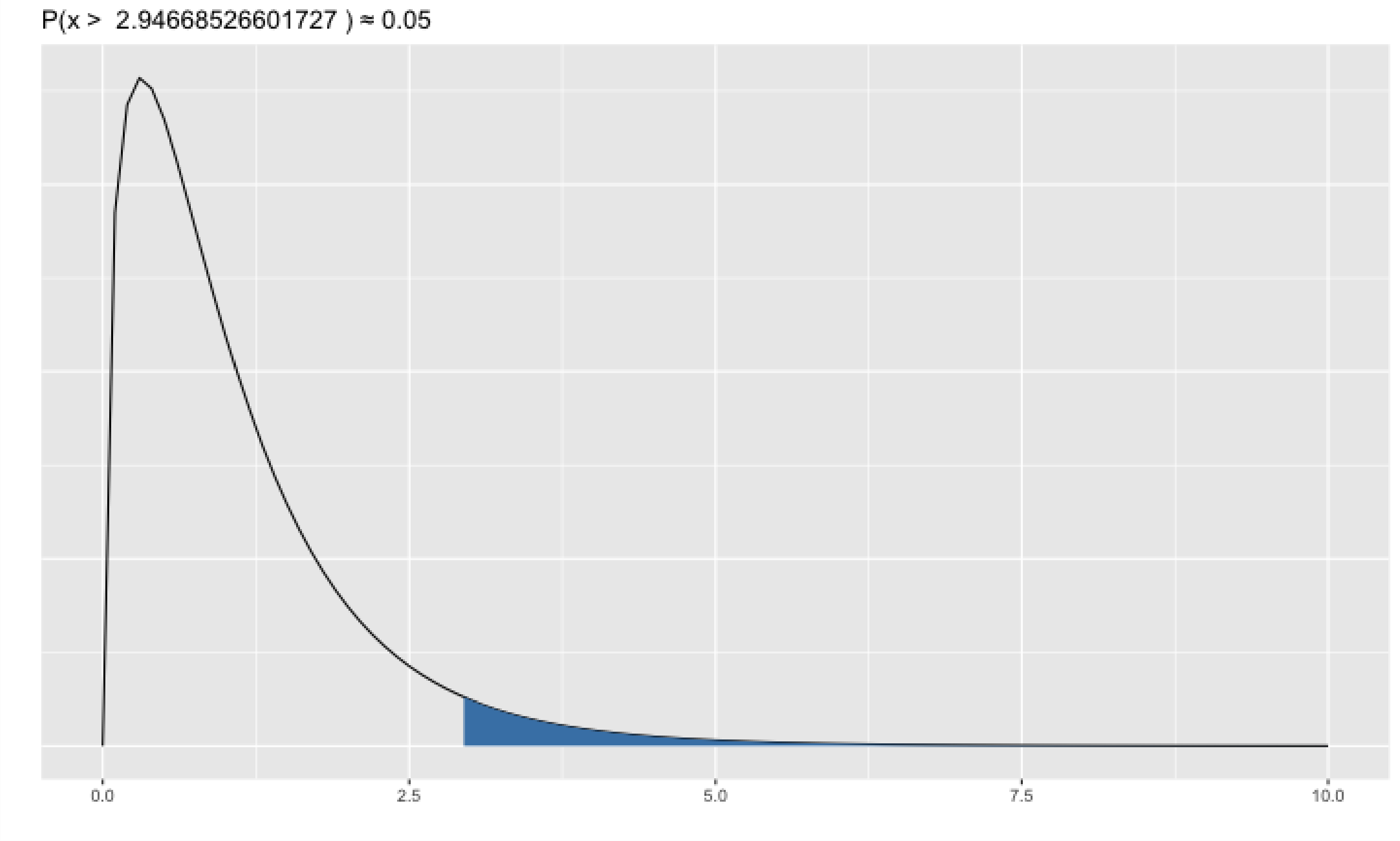
- $\frac{MS_T}{MS_E}$ follows the F-Distribution
    - Numerator df: k - 1 (k = number of groups)
    - Denominator df: k(n - 1)
    - n = # observations in each group

- $F = \frac{MS_T}{MS_E}$ is called the F-Statistic.

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

# The F-Distribution (cont.)

```
df.numerator <- 4 - 1
df.denominator <- 4 * (8 - 1)
DATA606::F_plot(df.numerator, df.denominator, cv = qf(0.95, df.numerator, df.denominator))
```



# ANOVA Table

| Source | Sum of Squares | *df* | MS | F | p |
|---|---|---|---|---|---|
| Between Group (Treatment) | $\sum_k n_k(\bar{x}_k - \bar{x})^2$ | 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 | | | |

# 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.

# More Information

ANOVA Vignette in the `VisualStats` package:
https://jbryer.github.io/VisualStats/articles/anova.html

The plots were created using the `VisualStats::anova_vis()` function.

Shiny app:

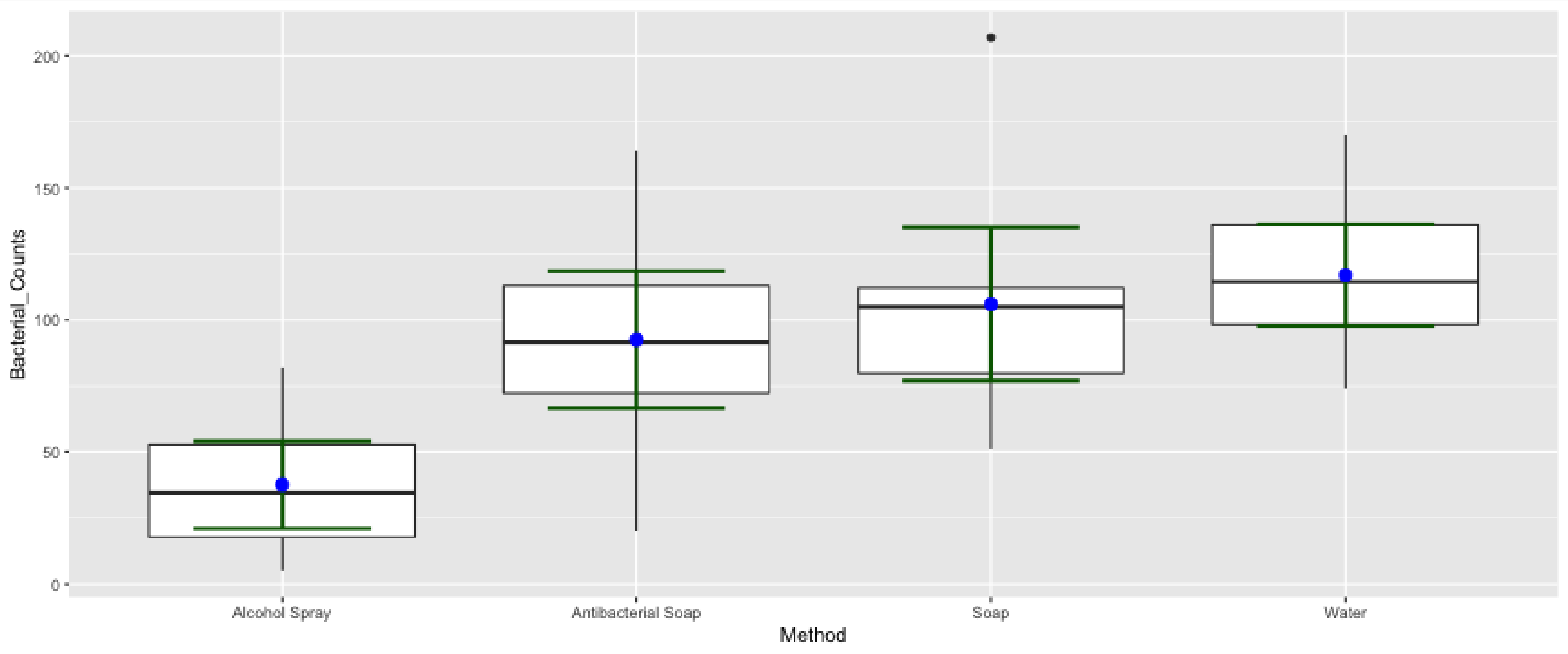
```
remotes::install_github('jbryer/VisualStats')
library(VisualStats)
shiny_demo('anova', package = 'VisualStats')
```

# 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 $\alpha/J$ 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**.

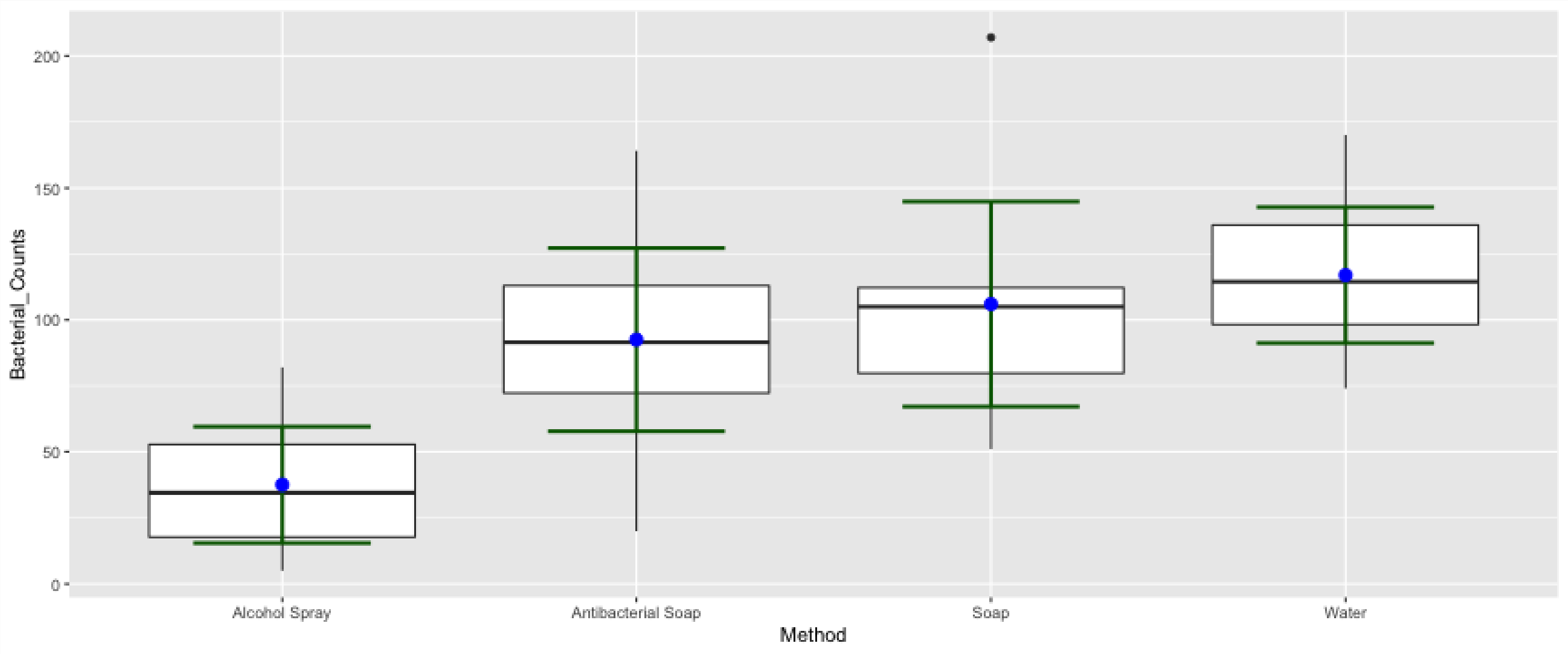
# Multiple Comparisons (no Bonferroni adjustment)

```r
cv <- qt(0.05, df = 15)
tab <- describeBy(hand_washing$Bacterial_Counts, group = hand_washing$Method, mat = TRUE)
ggplot(hand_washing, 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)
```



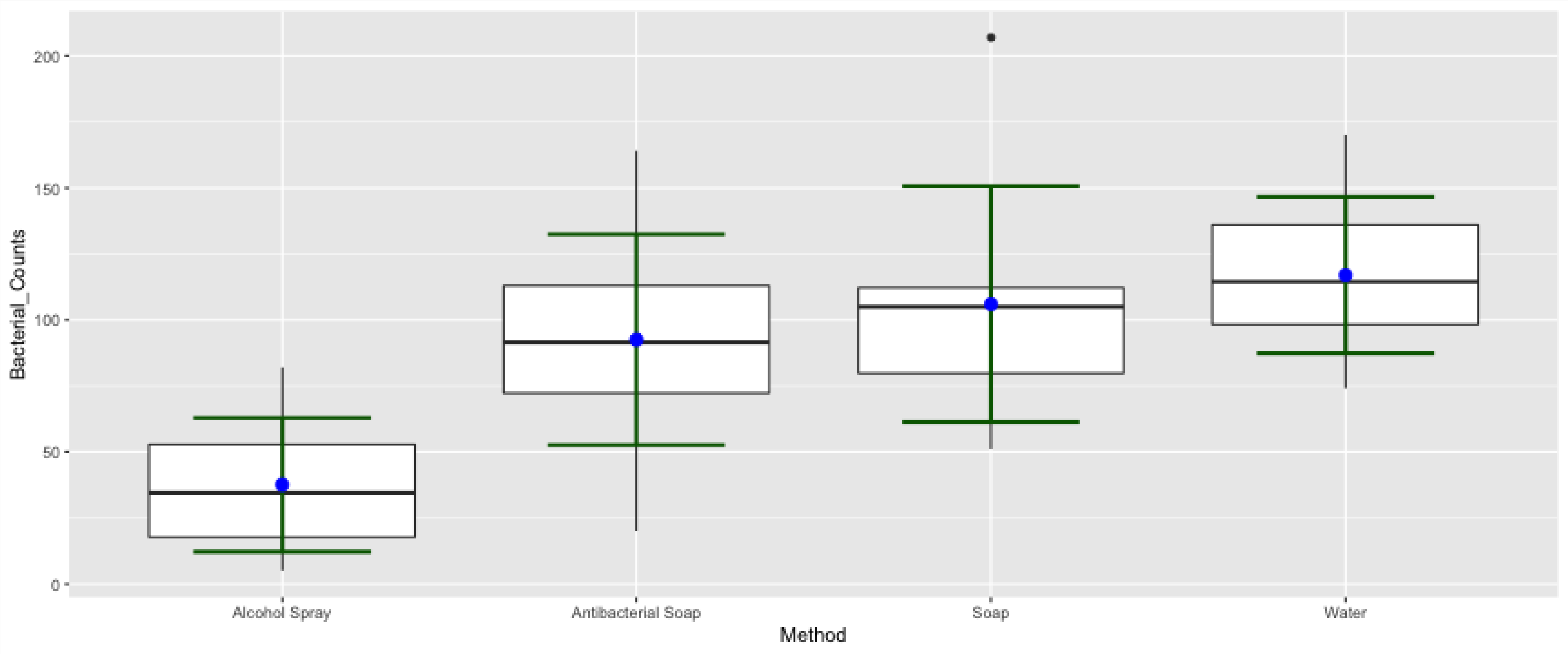
# Multiple Comparisons (3 paired tests)

```
cv <- qt(0.05 / 3, df = 15)
tab <- describeBy(hand_washing$Bacterial_Counts, group = hand_washing$Method, mat = TRUE)
ggplot(hand_washing, 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 = 15)
tab <- describeBy(hand_washing$Bacterial_Counts, group = hand_washing$Method, mat = TRUE)
ggplot(hand_washing, 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)
```



# One Minute Paper

Complete the one minute paper:

https://forms.gle/CA1dbnMtqQ7Zyj5Y8

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