Analysis of Variance (ANOVA)

IS381 - Statistics and Probability with R

Jason Bryer, Ph.D.

November 17, 2025

One Minute Paper Results

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

What important question remains unanswered for you?

]

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

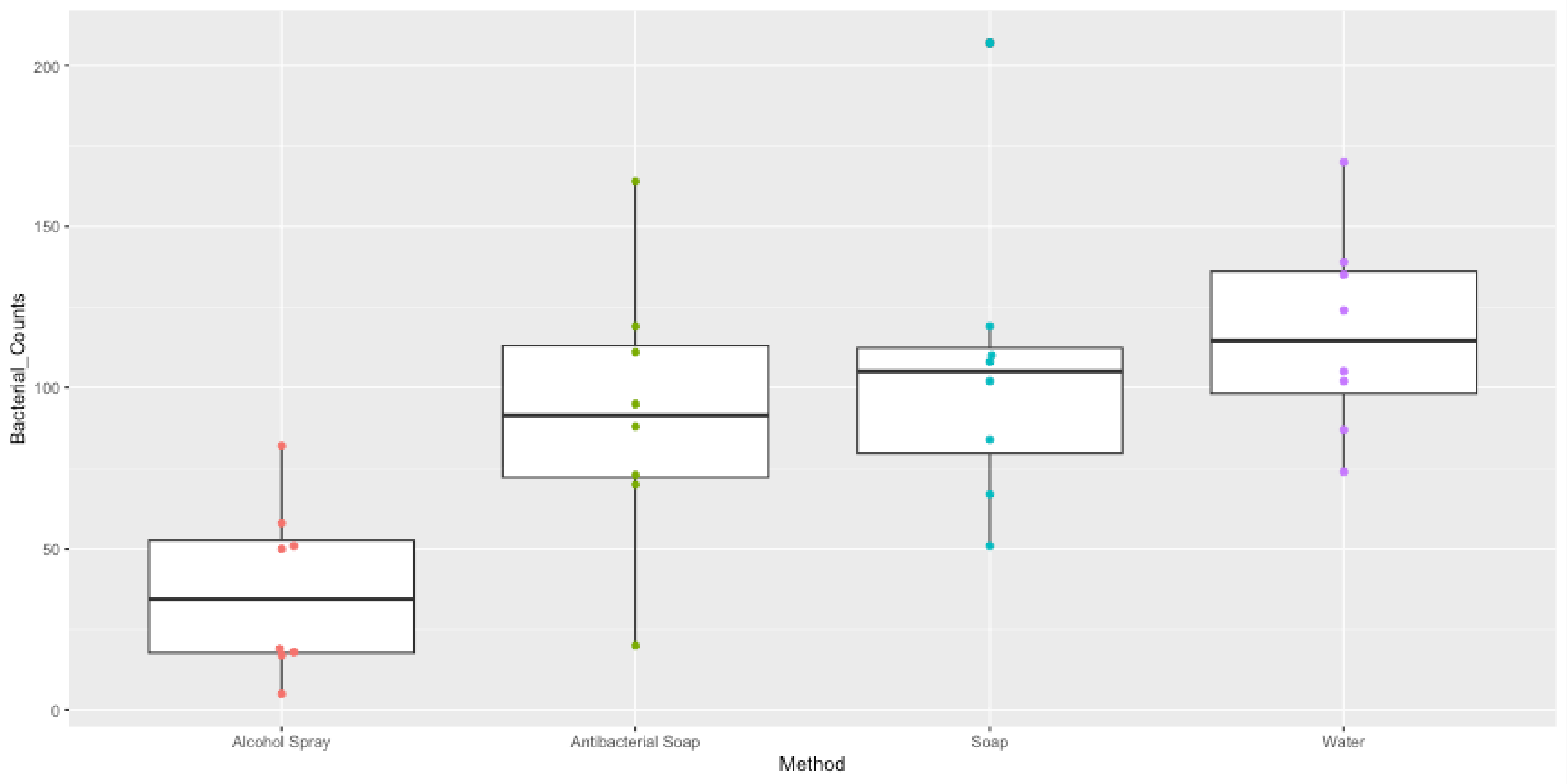
For ANOVA:

- 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

```r
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 median min max range        se       Var
## X11    1     Alcohol Spray    1 8  37.5 26.55991   34.5   5  82    77  9.390345  705.4286
## X12    2 Antibacterial Soap    1 8  92.5 41.96257   91.5  20 164   144 14.836008 1760.8571
## X13    3              Soap    1 8 106.0 46.95895  105.0  51 207   156 16.602496 2205.1429
## X14    4             Water    1 8 117.0 31.13106  114.5  74 170    96 11.006492  969.1429
```

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

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

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

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

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

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

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

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

```r
( pooled_var <- mean(desc$Var) )
```

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

Contrasts

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

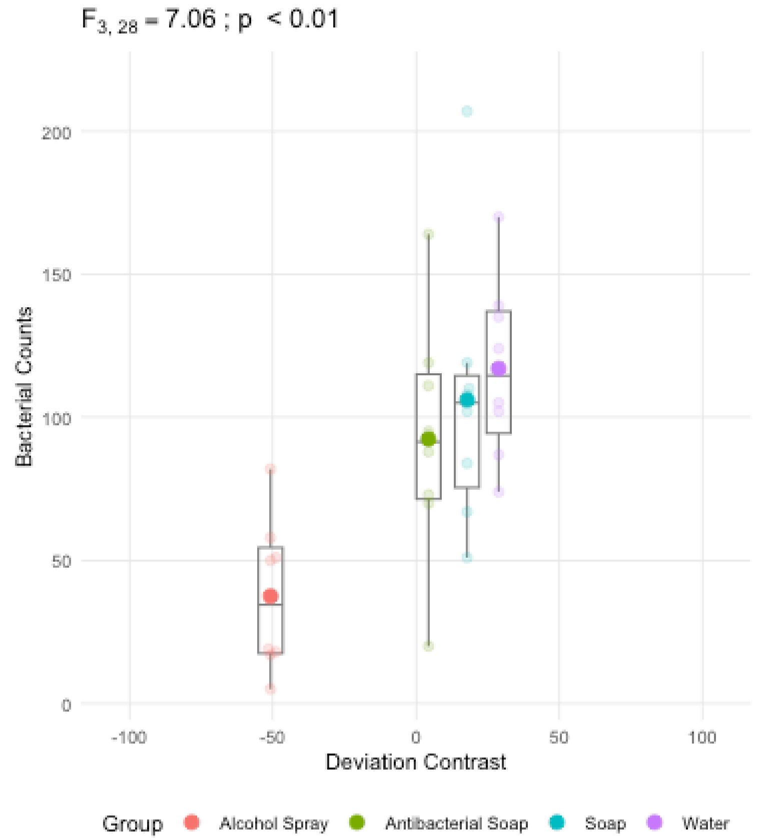
```
desc$contrast <- (desc$mean - mean(desc$mean))
mean(desc$contrast) # Should be 0!
```

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

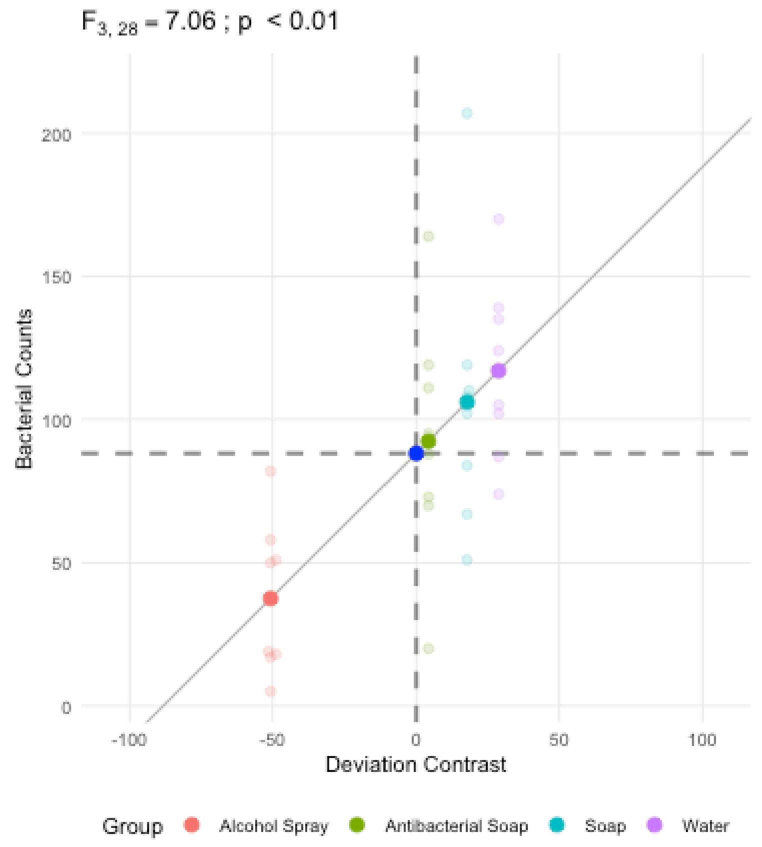
```
desc
```

```
##     item               Method vars n  mean       sd median min max range        se       Var contrast
## X11    1       Alcohol Spray    1 8  37.5 26.55991   34.5   5  82    77  9.390345  705.4286   -50.75
## X12    2 Antibacterial Soap    1 8  92.5 41.96257   91.5  20 164   144 14.836008 1760.8571     4.25
## X13    3                Soap    1 8 106.0 46.95895  105.0  51 207   156 16.602496 2205.1429    17.75
## X14    4               Water    1 8 117.0 31.13106  114.5  74 170    96 11.006492  969.1429    28.75
```

Plotting using contrasts



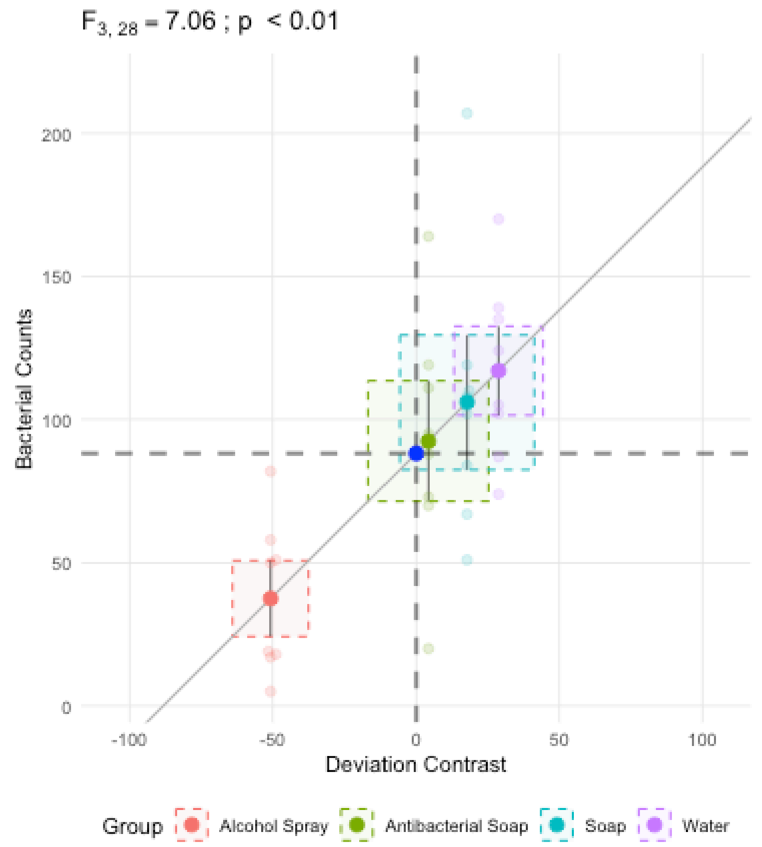
Grade Mean and Unit Line (slope = 1, intercept = $\bar{x}$)



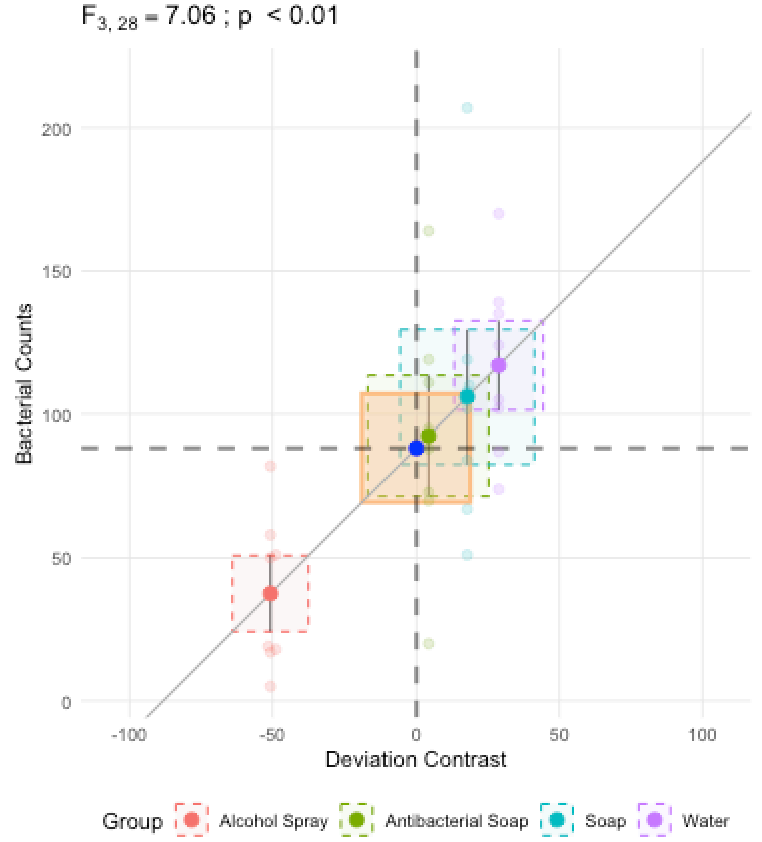
Within Group Variance (error)

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

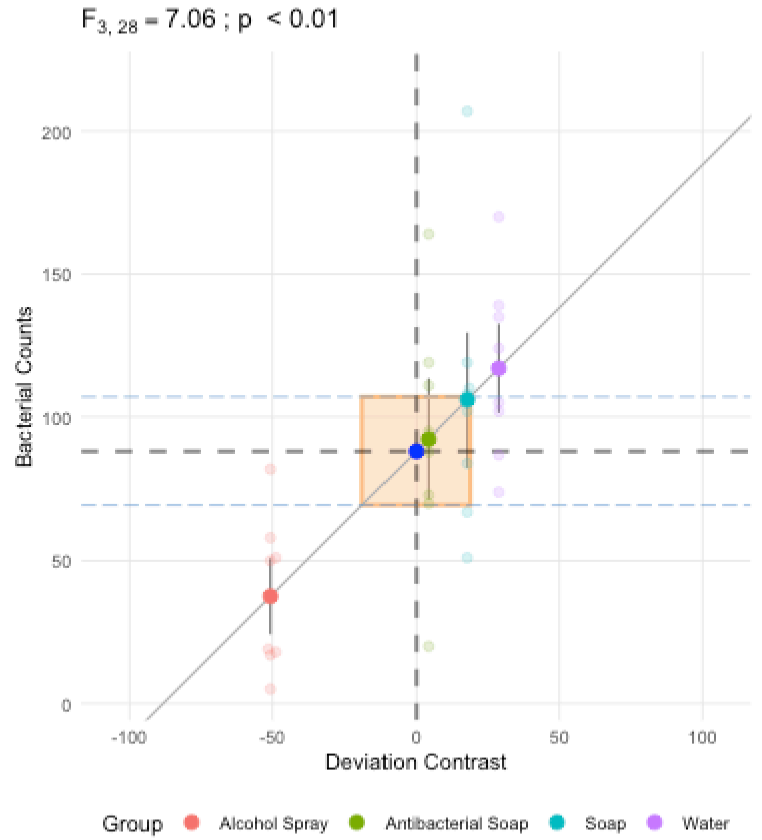
Within Group Variance (error)



Within Group Variance (error)



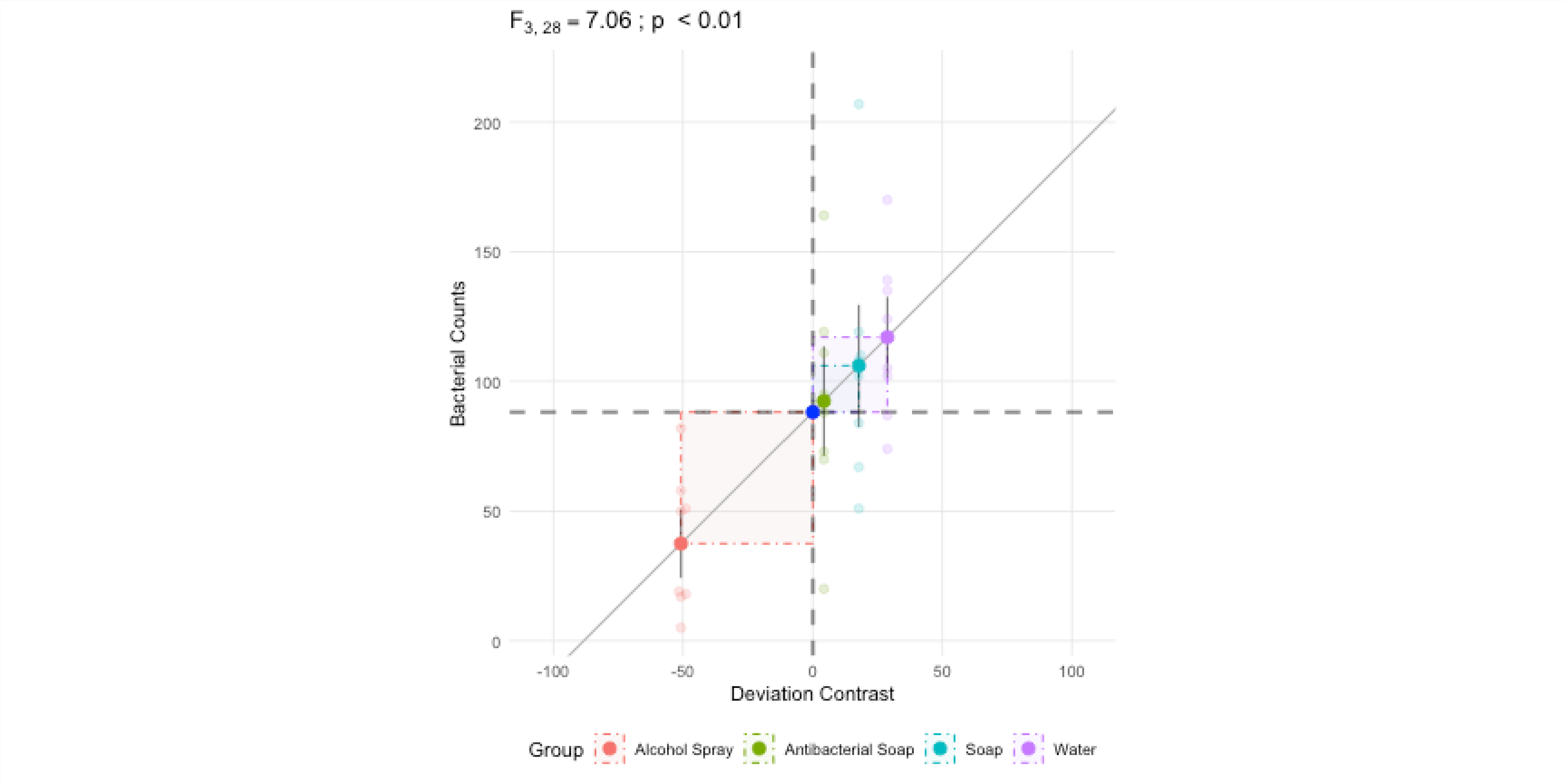
Within Group Variance (error)



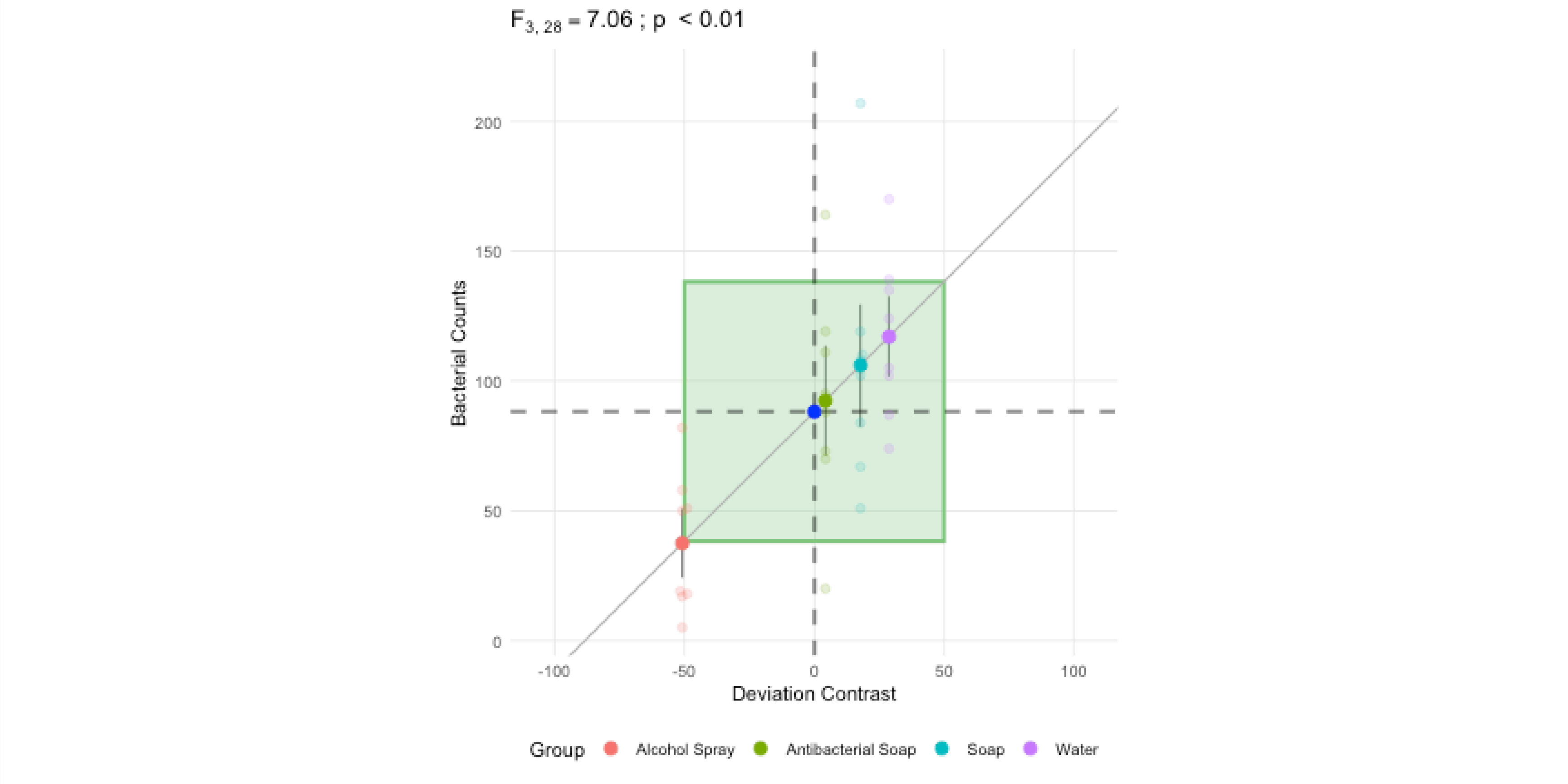
Between Group Variance

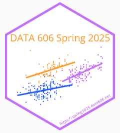
$$SS_{between} = \sum_{k} n_k (\bar{x}_k - \bar{x})^2$$

Between Group Variance



Between Group Variance



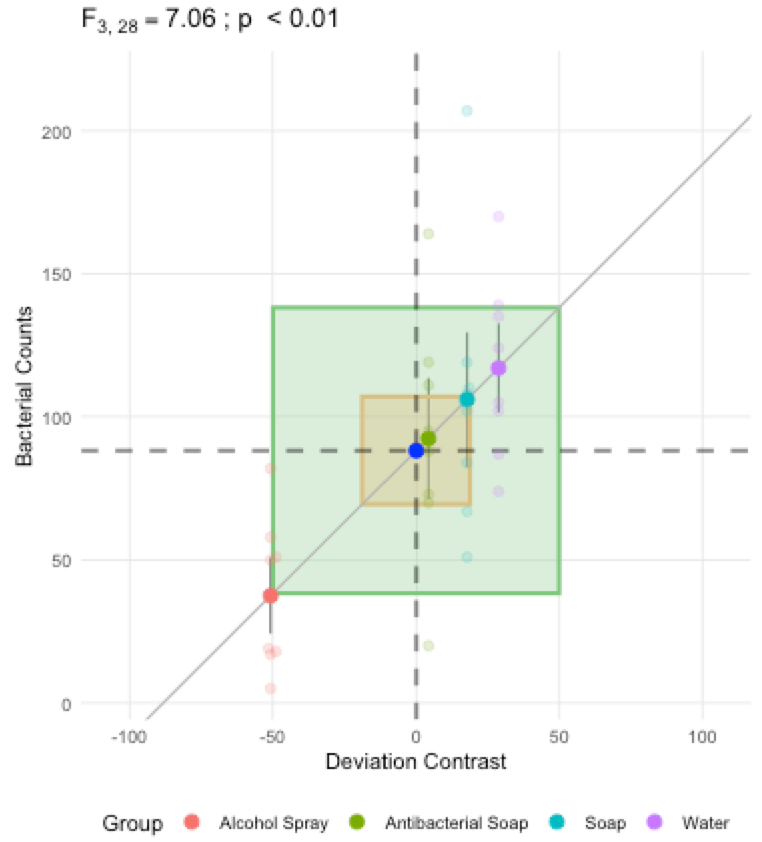


Mean Square

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	$\sum_n (x_n - \bar{x})^2$	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 MSE_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

MSE_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 MSE_E

Comparing

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

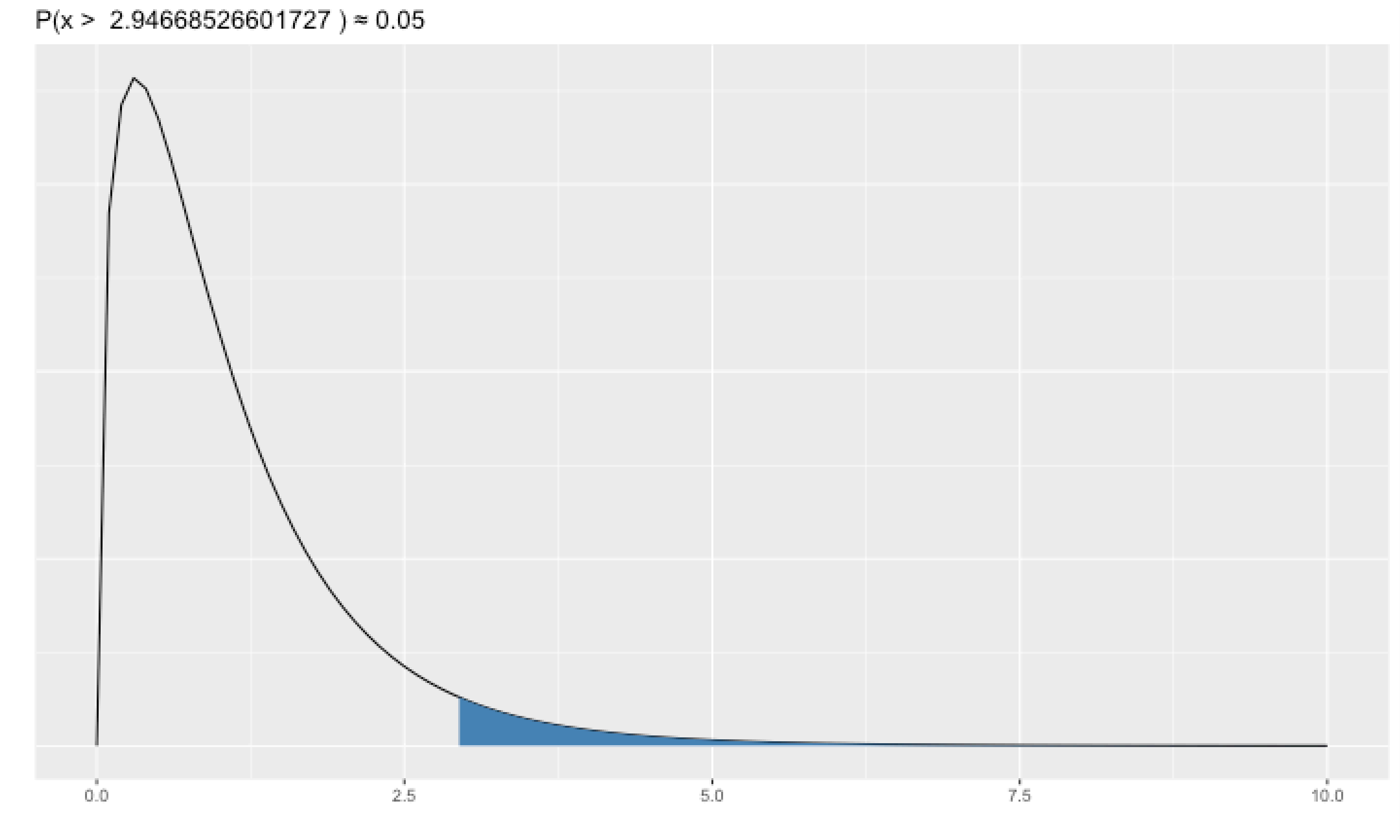
The F-Distribution

- 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	$\sum_n (x_n - \bar{x})^2$	n - 1			

```r
aov(Bacterial_Counts ~ Method, data = hand_washing) |> summary()
```

```
##             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
```

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

ANVA 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:

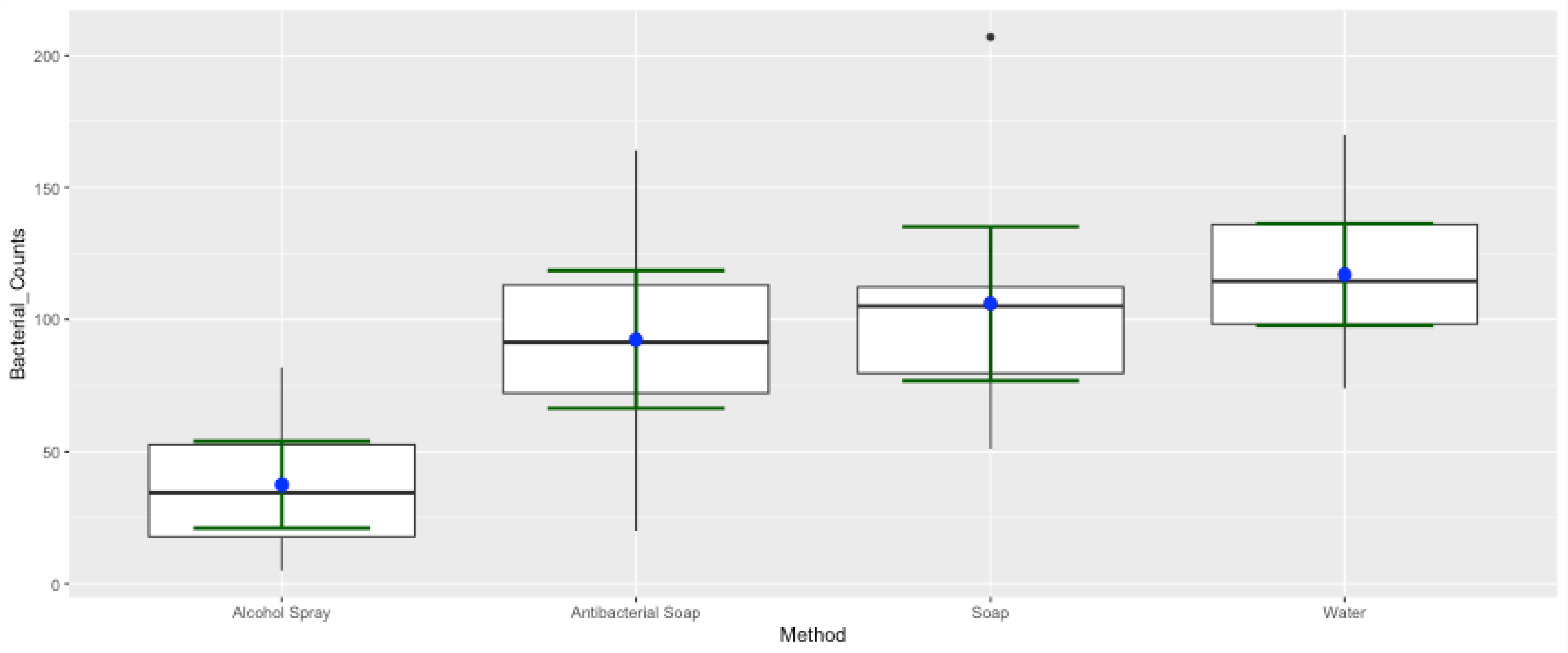
```
# remotes::install_github('jbryer/VisualStats')
library(VisualStats)
VisualStats::anova_shiny()
```

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**.

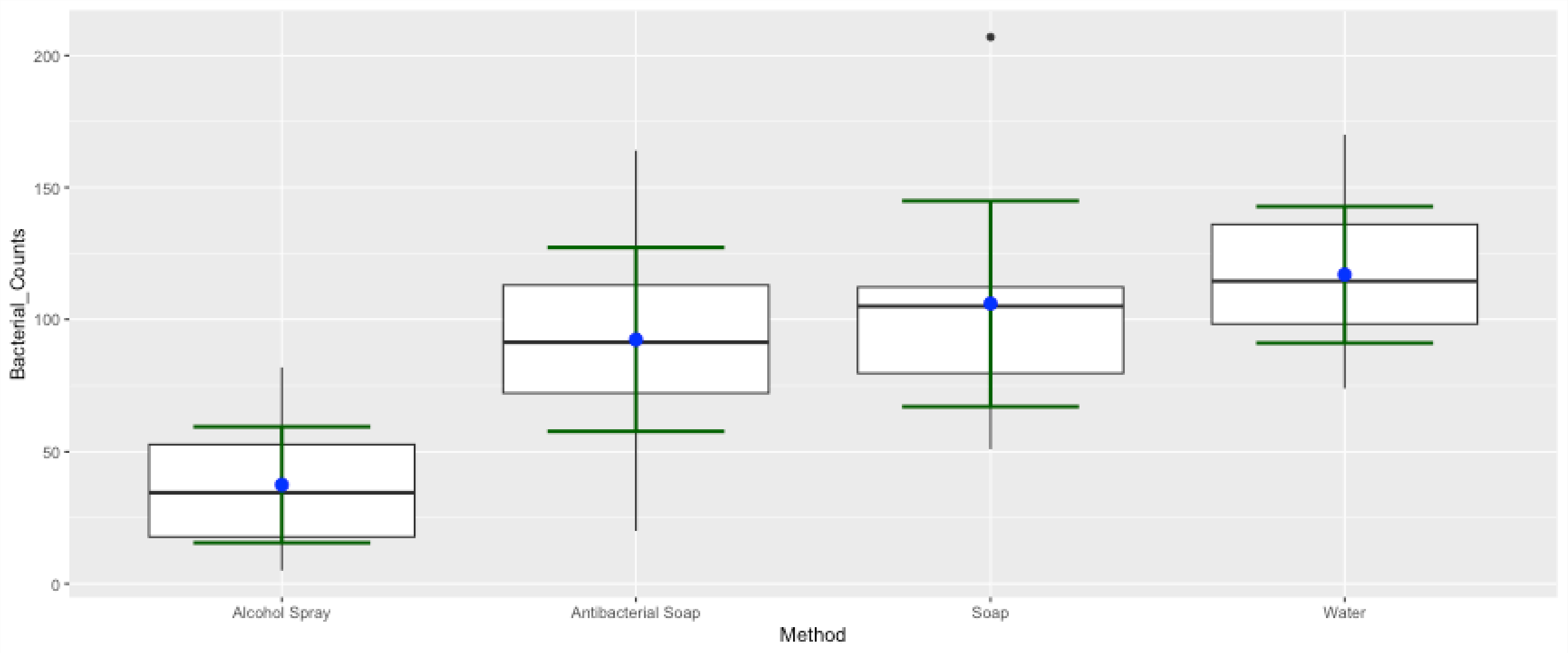
Multiple Comparisons (no Bonferroni adjustment)

```
cv <- qt(0.05, df = 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)
```



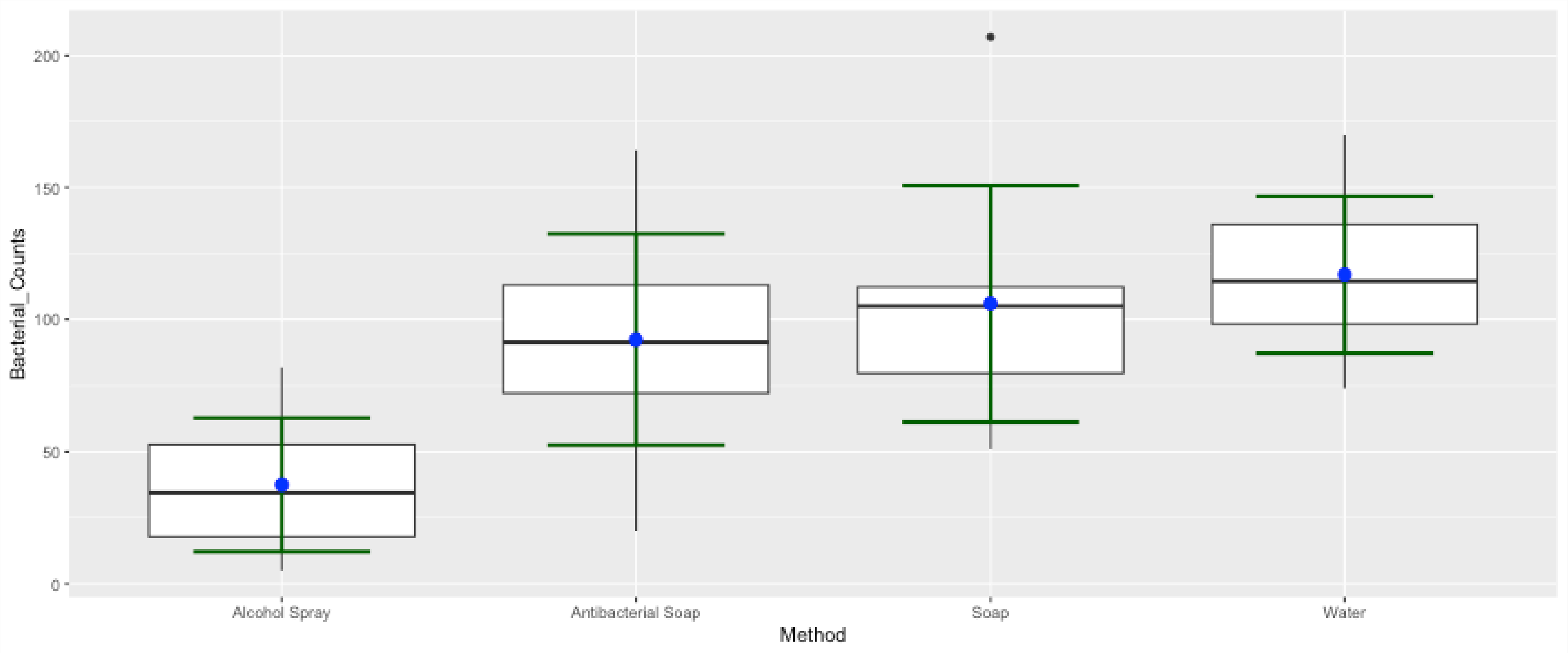
Multiple Comparisons (3 paired tests)

```
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

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

https://forms.gle/N8WjTAysfKbGLptLA