(Graphical) Analysis of Variance

DATA 606 - Statistics & Probability for Data Analytics

Jason Bryer, Ph.D.

March 23, 2022

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.

Descriptive Statistics

```
desc <- psych::describeBy(hand$Bacterial.Counts, group = hand$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$Method)) )
```

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

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

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

```
( grand_mean <- mean(hand$Bacterial.Counts) )
```

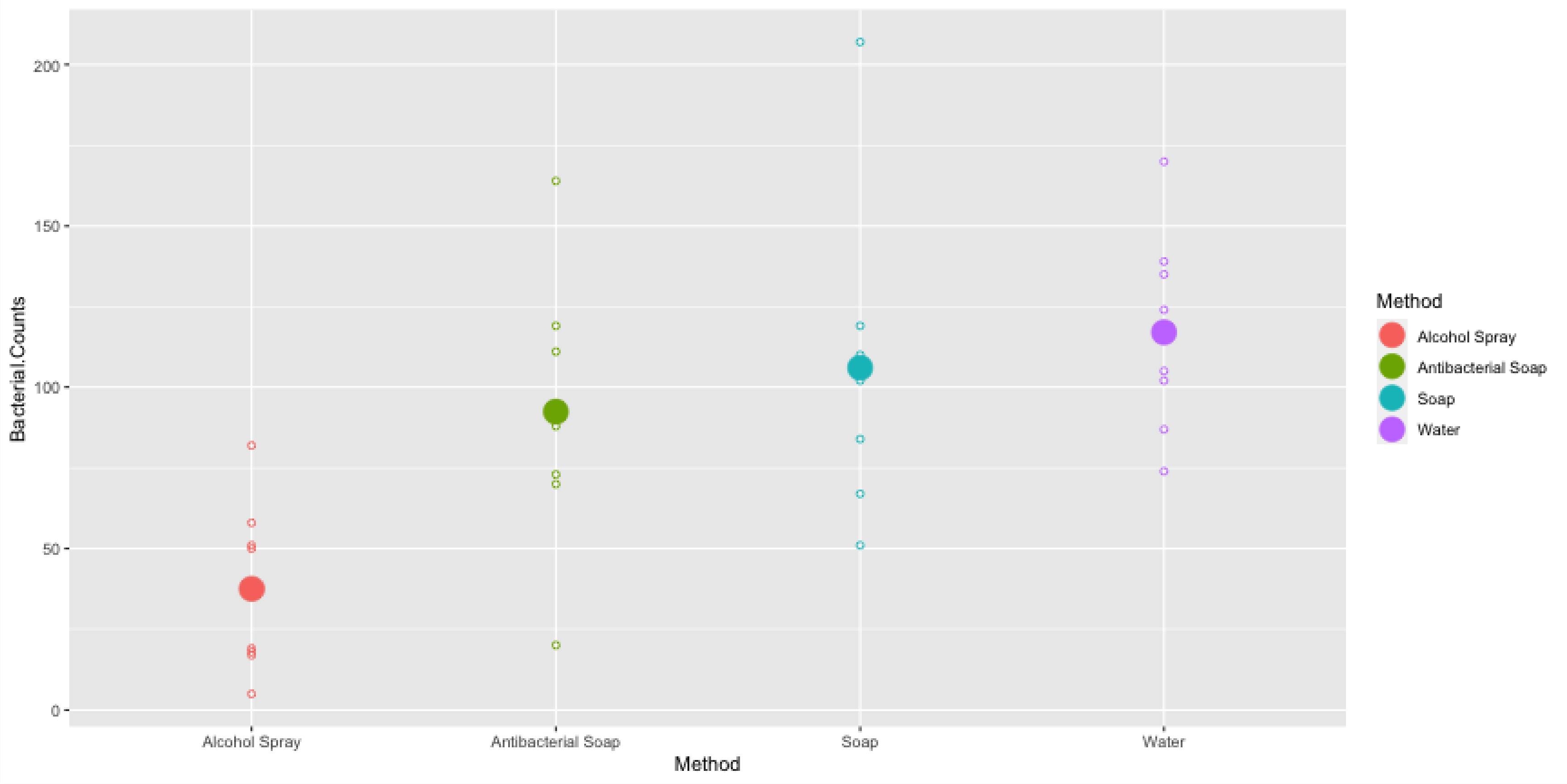
```
## [1] 88.25
```

```
( grand_var <- var(hand$Bacterial.Counts) )
```

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

Hand Washing Comparison

```
ggplot(hand, aes(x = Method, y = Bacterial.Counts)) +
    geom_point(aes(color = Method), shape = 1) +
    stat_summary(geom = 'point', fun = mean, size = 6, aes(color = Method))
```



Contrasts

A contrast is a linear combination of 2 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
```

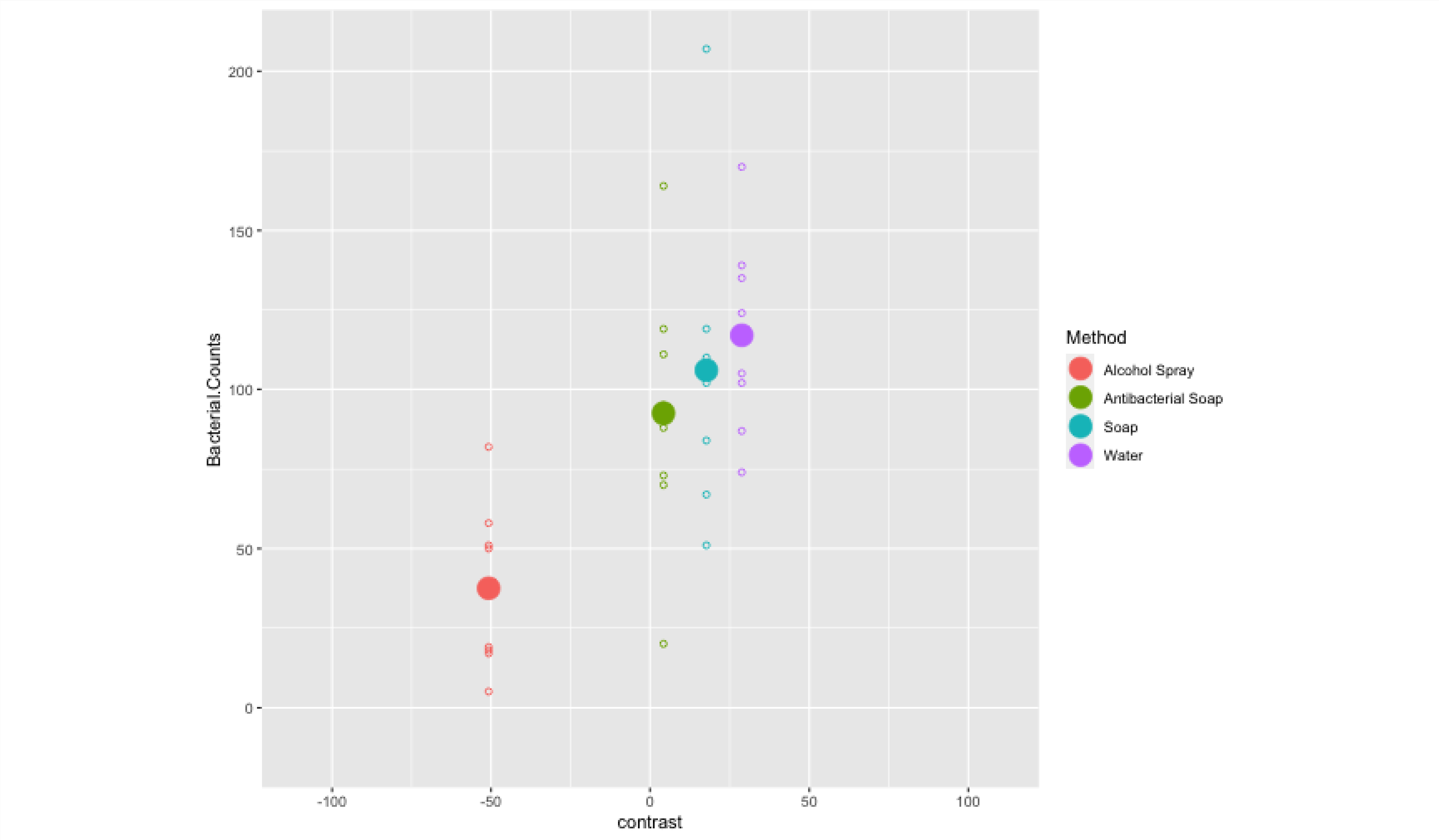
```
##     item             Method vars n  mean       sd min max range        se
## X11    1      Alcohol Spray    1 8  37.5 26.55991   5  82    77  9.390345
## X12    2 Antibacterial Soap    1 8  92.5 41.96257  20 164   144 14.836008
## X13    3               Soap    1 8 106.0 46.95895  51 207   156 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
```

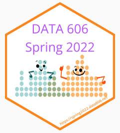
Plotting using contrasts

```r
xlim <- c(-1.1 * diff(range(hand$Bacterial.Counts)) / 2,
          1.1 * diff(range(hand$Bacterial.Counts)) / 2)
ylim <- c(1.1 * (grand_mean - diff(range(hand$Bacterial.Counts)) / 2),
          1.1 * (grand_mean + diff(range(hand$Bacterial.Counts)) / 2))
p <- ggplot(hand, aes(x = contrast, y = Bacterial.Counts)) +
    geom_point(aes(color = Method), shape = 1) +
    stat_summary(geom = 'point', fun = mean, size = 6, aes(color = Method)) +
    coord_equal() +
    xlim(xlim) + ylim(ylim)
```

DATA 606 Spring 2022

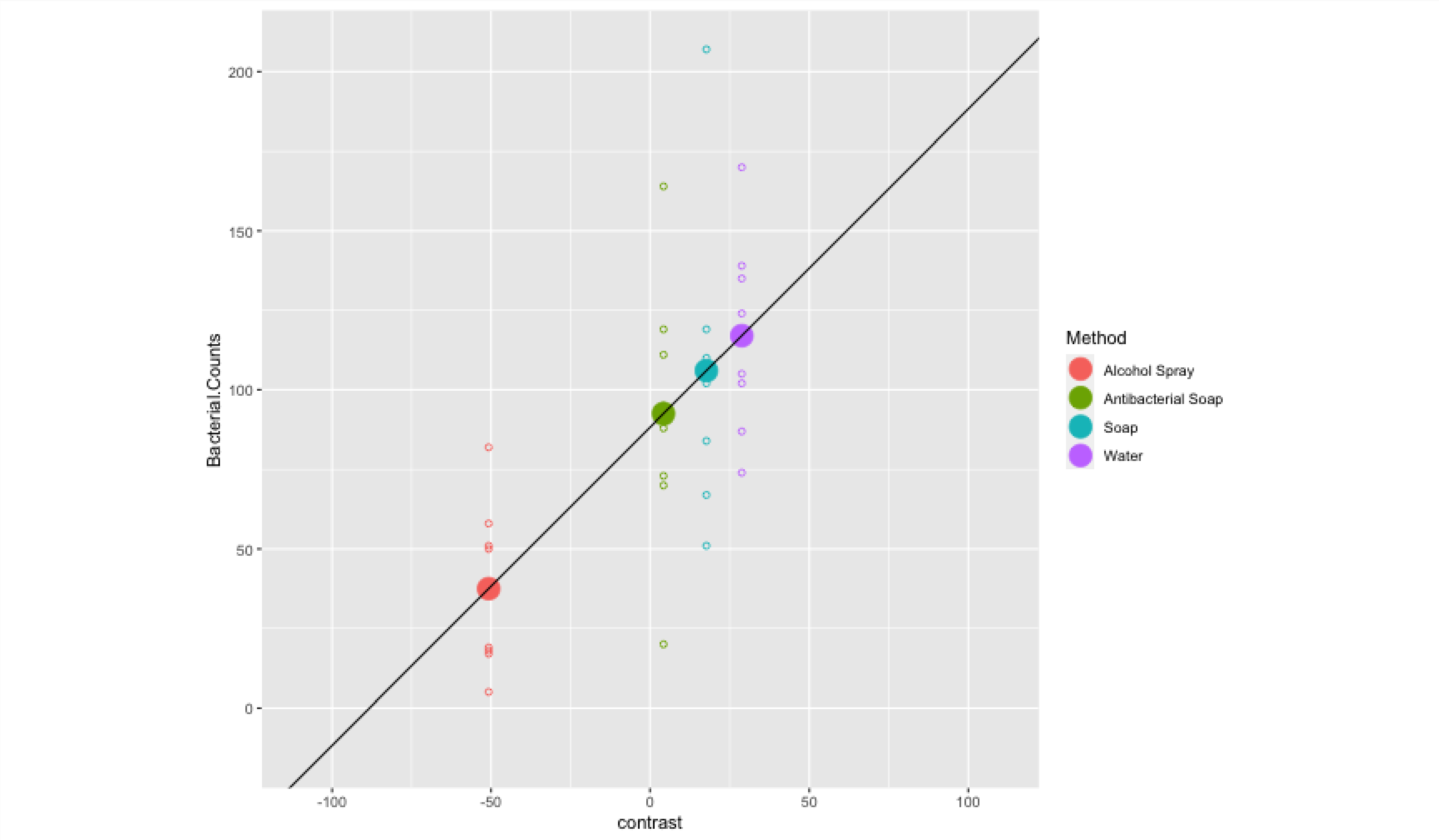
Plotting using contrasts





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

```
( p <- p  + geom_abline(slope = 1, intercept = mean(hand$Bacterial.Counts)) )
```



DATA 606 Spring 2022

Grand Mean

```
( p <- p + geom_hline(yintercept = grand_mean, linetype = 2) +
           geom_vline(xintercept = 0, linetype = 2) )
```

Overall

Grand Variance

```
( p + geom_rect(xmin = 0 - sqrt(grand_var), ymin = grand_mean - sqrt(grand_var),
                xmax = 0 + sqrt(grand_var), ymax = grand_mean + sqrt(grand_var),
                color = 'blue', linetype = 2, fill = 'blue', alpha = 0.005) )
```

Group Variances

```
df_rect_within <- hand %>%
    mutate(square = (Bacterial.Counts - mean)^2) %>%
    group_by(Method) %>%
    summarize(contrast = mean(Bacterial.Counts) - grand_mean,
              mean = mean(Bacterial.Counts),
              MS = sum(square) / (n() - 1)) %>%
    mutate(xmin = contrast - sqrt(MS),
           xmax = contrast + sqrt(MS),
           ymin = mean - sqrt(MS),
           ymax = mean + sqrt(MS))
df_rect_within
```

```
## # A tibble: 4 × 8
##   Method              contrast  mean    MS   xmin  xmax  ymin  ymax
##   <chr>                  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Alcohol Spray         -50.8   37.5  705. -77.3  -24.2  10.9  64.1
## 2 Antibacterial Soap      4.25  92.5 1761. -37.7   46.2  50.5 134.
## 3 Soap                   17.8  106   2205. -29.2   64.7  59.0 153.
## 4 Water                  28.8  117    969.  -2.38  59.9  85.9 148.
```

Group Variances

```
p + geom_segment(data = desc, aes(x = contrast, xend = contrast, y = mean - sd, yend = mean + sd), alpha = 0.6) +
    geom_rect(data = df_rect_within, aes(xmin = xmin, ymin = ymin, xmax = xmax, ymax = ymax, y = 0, color = Method),
              alpha = 0.005, linetype = 2) + scale_color_brewer(type = 'qual', palette = 7)
```

DATA 606
Spring 2022

Between Group Variance (treatment)

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

```
( 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
```

Between Group Variance (treatment)

```r
p + geom_rect(xmin = 0 - sqrt(MS_between), ymin = grand_mean - sqrt(MS_between),
              xmax = 0 + sqrt(MS_between), ymax = grand_mean + sqrt(MS_between),
              color = 'darkgreen', linetype = 2, fill = 'green', alpha = 0.005)
```

DATA 606
Spring 2022

Group Variances (cont.)

Within Group Variance (error)

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

```r
( df_total <- n - 1)
```

```
## [1] 31
```

```r
( ss_total <- sum((hand$Bacterial.Counts-grand_mean)^2)
```

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

```r
( MS_total <- ss_total / df_total )
```

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

```r
( df_within <- n - k )
```

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

```r
( ss_within <- ss_total - ss_between )
```

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

```r
( MS_within <- ss_within / df_within )
```

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

Within Group Variance (error)

```
p + geom_rect(xmin = 0 - sqrt(MS_within), ymin = grand_mean - sqrt(MS_within),
              xmax = 0 + sqrt(MS_within), ymax = grand_mean + sqrt(MS_within),
              color = 'red', linetype = 2, fill = 'red', alpha = 0.005)
```

Within group variance is the average varance across

```r
mean( (df_rect_within$ymax - df_rect_within$ymin) * (df_rect_within$xmax - df_rect_within$xmin) ) / k
```

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

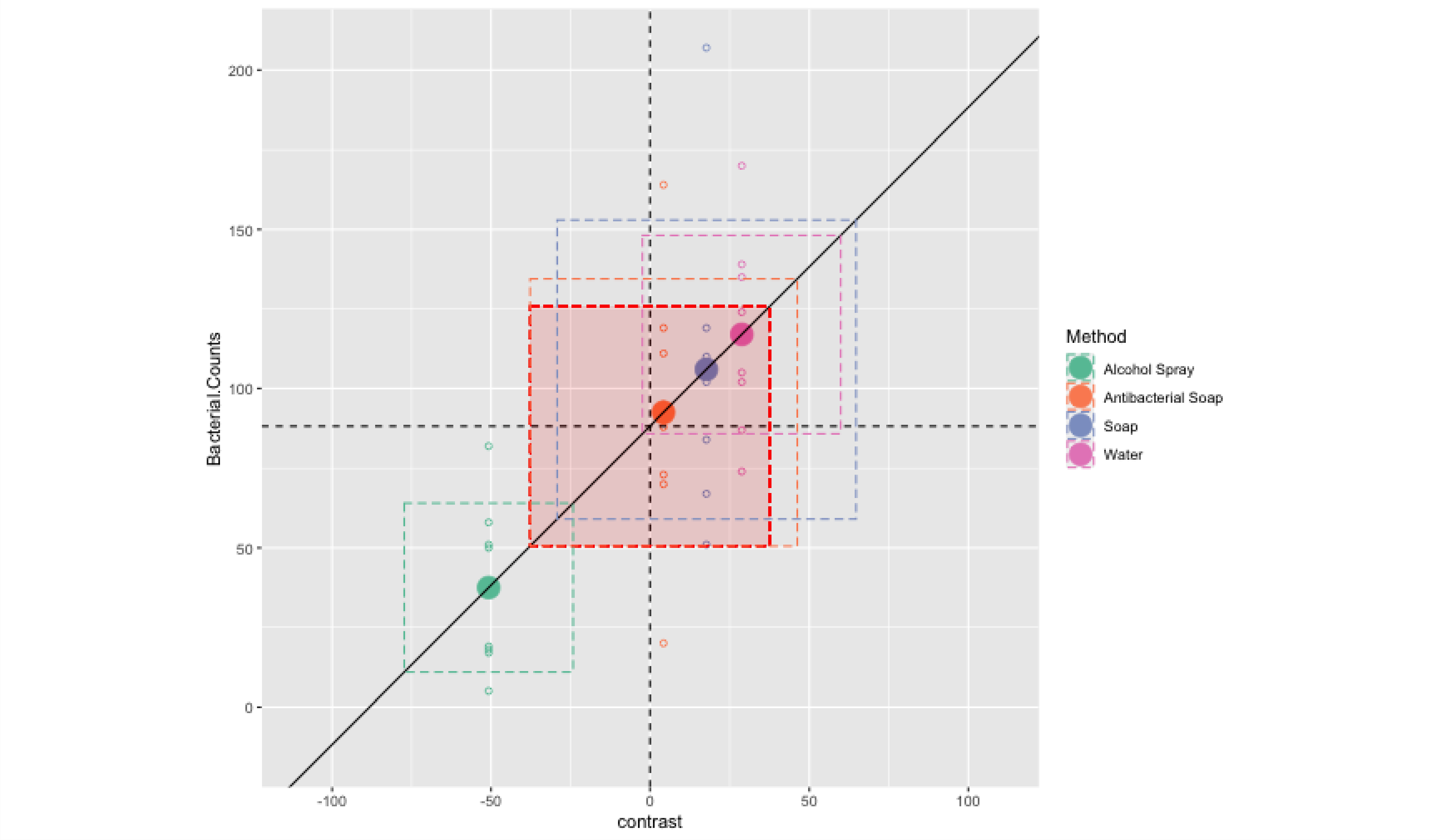
```r
MS_within
```

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

```r
p2 <- p + geom_rect(xmin = 0 - sqrt(MS_within), ymin = grand_mean - sqrt(MS_within),
              xmax = 0 + sqrt(MS_within), ymax = grand_mean + sqrt(MS_within),
              color = 'red', linetype = 2, fill = 'red', alpha = 0.005) +
    geom_rect(data = df_rect_within, aes(xmin = xmin, ymin = ymin, xmax = xmax, ymax = ymax, y = 0, color = Method),
              alpha = 0.005, linetype = 2) + scale_color_brewer(type = 'qual', palette = 7)
```

DATA 606
Spring 2022

Within group variance is the average varance across



DATA 606 Spring 2022

$MS_{Between}/MS_{Within}$ = 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.

Hand Washing Comparison (cont.)

```
desc <- describeBy(hand$Bacterial.Counts, hand$Method, mat=TRUE, skew = FALSE)
desc$Var <- desc$sd^2
print(desc, row.names=FALSE)
```

```
##  item              group1 vars n  mean       sd min max range        se
##     1       Alcohol Spray    1 8  37.5 26.55991   5  82    77  9.390345
##     2 Antibacterial Soap    1 8  92.5 41.96257  20 164   144 14.836008
##     3                Soap    1 8 106.0 46.95895  51 207   156 16.602496
##     4               Water    1 8 117.0 31.13106  74 170    96 11.006492
##        Var
##   705.4286
##  1760.8571
##  2205.1429
##   969.1429
```

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

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

DATA 606
Spring 2022

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

DATA 606 Spring 2022

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/

DATA 606 Spring 2022

The F-Distribution (cont.)

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

DATA 606
Spring 2022

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			

DATA 606
Spring 2022

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
```

DATA 606
Spring 2022

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
```

DATA 606
Spring 2022

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 $F_{3,28} \approx 0.0011$

DATA 606
Spring 2022

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.

DATA 606
Spring 2022

One Minute Paper

Complete the one minute paper:

https://forms.gle/qxRnsCyydx1nf8sXA

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