Documentation on using ioslides is available here: http://rmarkdown.rstudio.com/ioslides_presentation_format.html Some slides are adopted (or copied) from OpenIntro: https://www.openintro.org/

Announcements

- There will be **NO MEETUP next Thursday**, March 23rd.
- If you need a few extra days for the data project proposal, you can submit the proposals by Friday, March 24th.

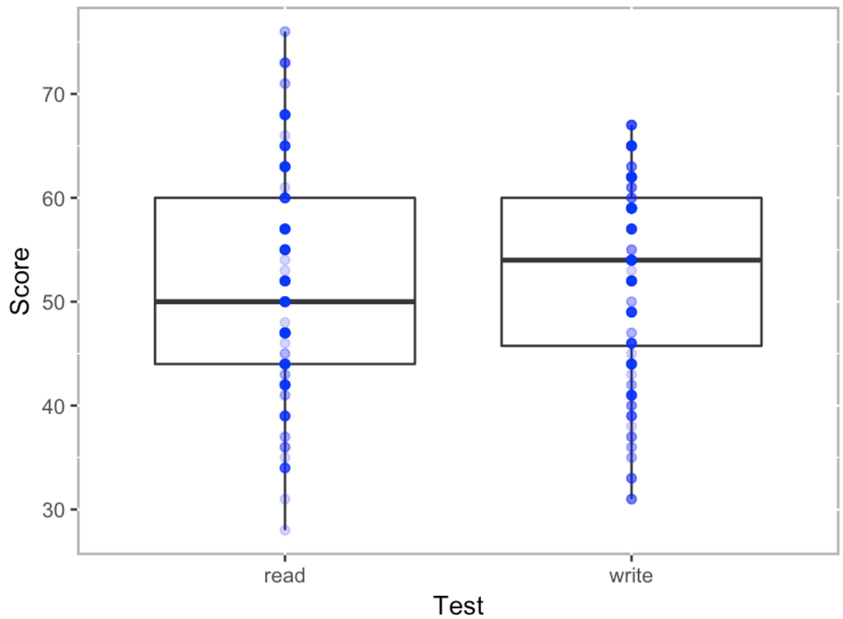
Meetup Presentations

- Daniel Thonn (5.13)

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')
```



4/32

High School & Beyond Survey

```
head(hsb2)
```

```
##    id gender  race    ses schtyp       prog read write math science socst
## 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?

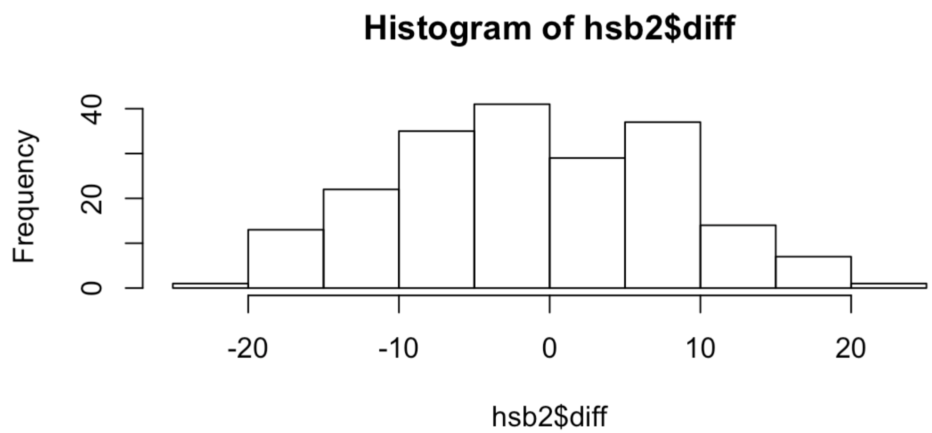
Analyzing Paired Data

- 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.

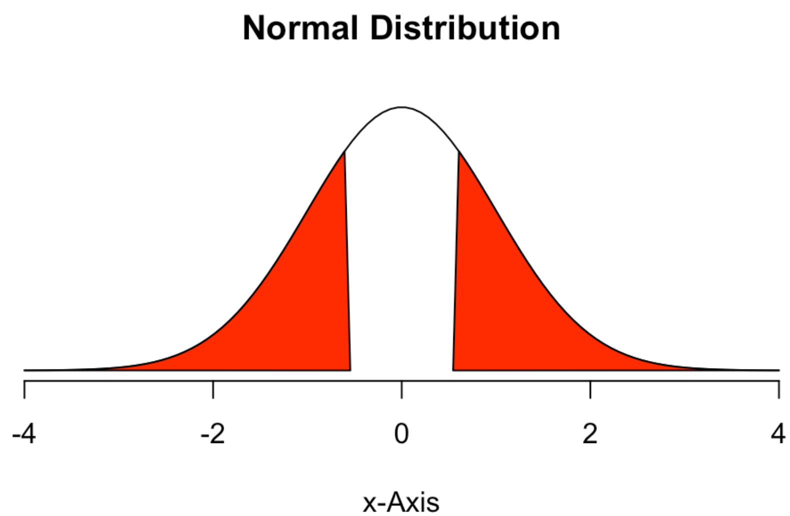
$$\mu_{diff} \neq 0$$

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
```

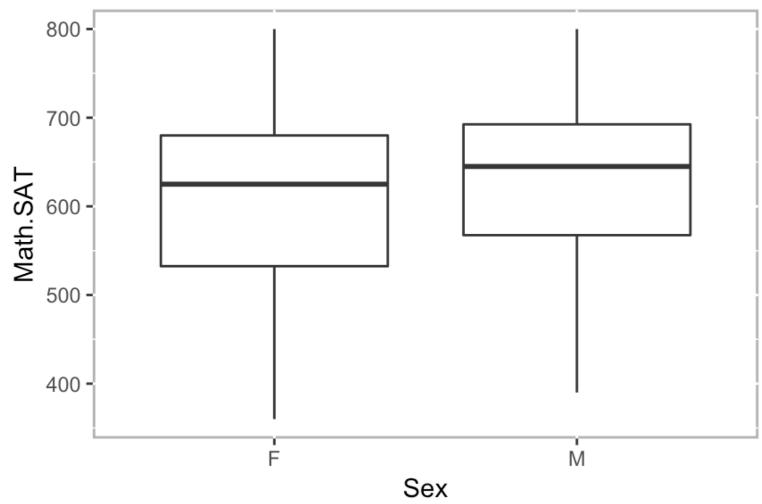
Is there a difference in math scores between males and females?

SAT Scores by Gender

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

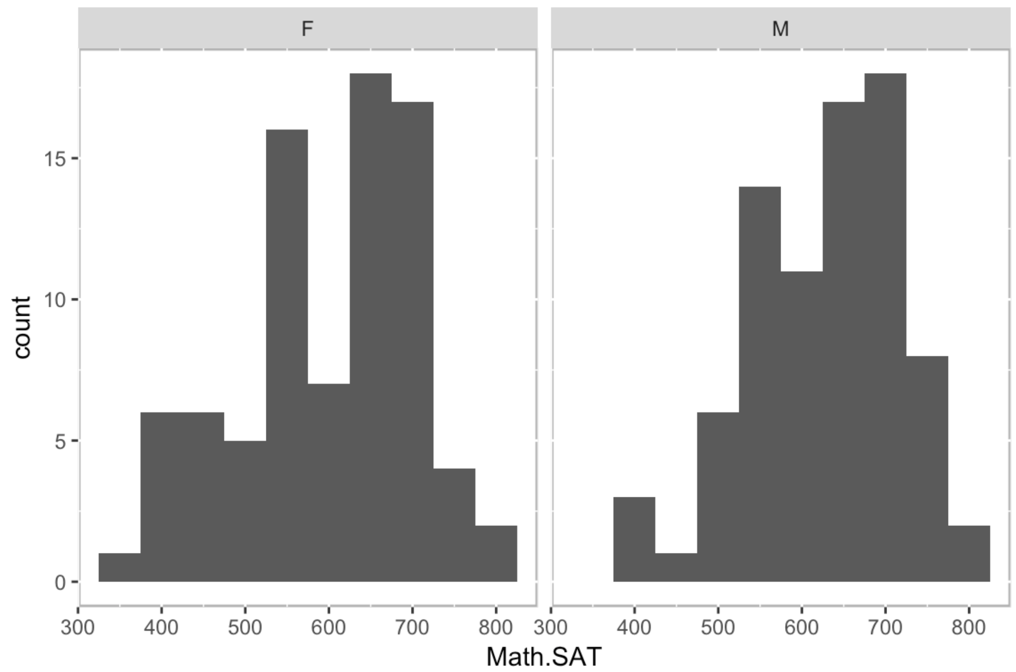
```
##     group1  n     mean        sd min
## X11      F 82 597.6829 103.70065 360
## X12      M 80 626.8750  90.35225 390
```

```
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

16/32

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 of the difference between two sample means

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

17/32

Confidence Interval for Difference in SAT Scores

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

Analysis of Variance (ANOVA)

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

Example

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

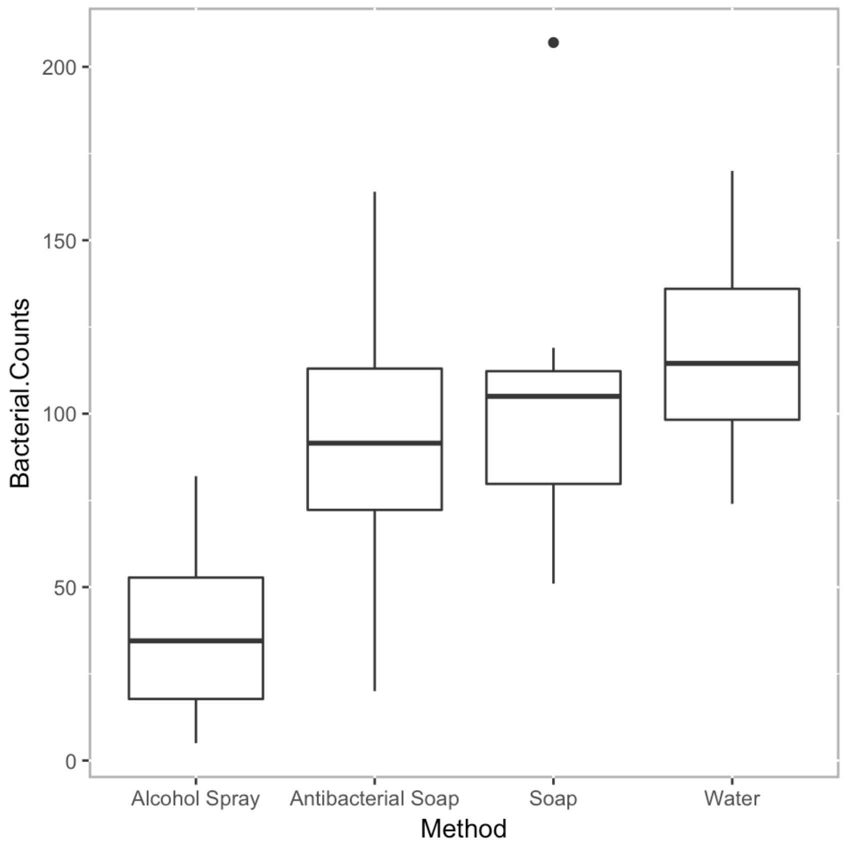
- Each tested with 8 replications
- Treatments randomly assigned

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()
```



20/32

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 {37.5, 92.5, 106.0, 117.0} is 1245.08.
- $\frac{\sigma^2}{8} = 1245.08$
- $\sigma^2 = 9960.64$
- 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?

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: $s_p^2 = 1410.10$
- Other names for s_p^2: Error Mean Square, Within Mean Square, MS_E.

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

24/32

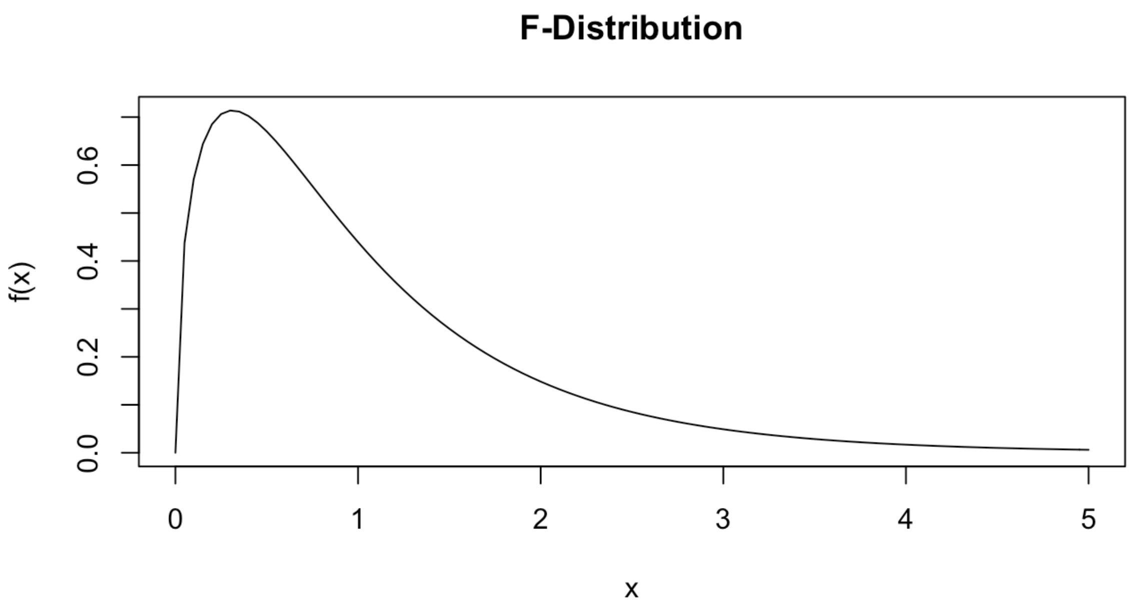
The F-Distribution

- 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: http://shiny.albany.edu/stat/fdist/

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

- $MS_T = 9960.64$
- $MS_E = 1410.14$
- Numerator df = 4 – 1 = 3
- Denominator df = 4(8 – 1) = 28.

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

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

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

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

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

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.

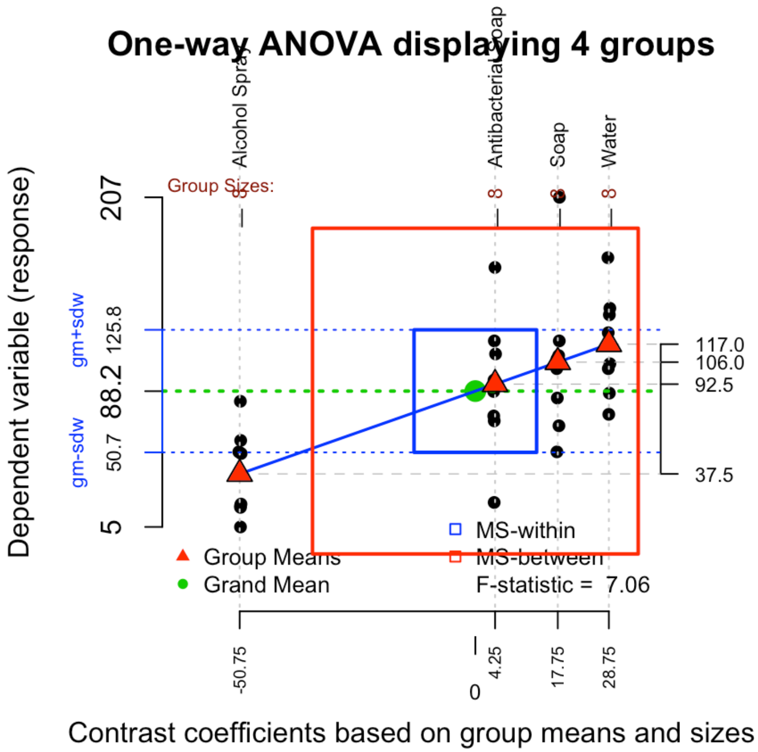
ANOVA in R

```
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

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



30/32

Graphical ANOVA

```
hand.anova
```

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

What Next?

- P-value large $\rightarrow$ Nothing left to say
- P-value small $\rightarrow$ 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) ≈ 0.98333.
  - Use this adjusted value to find t**.