Lab03 - F tests for ANOVA

Solutions

Goals

The goal in this lab is to practice F tests for ANOVA.

Loading packages

Here are some packages with functionality you may need for this lab. Run this code chunk now.

```r
library(readr)
library(ggplot2)
library(gridExtra)
library(mosaic)
library(dplyr)

options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

A study was conducted to examine the effectiveness of four different hand-washing methods for eliminating bacteria.

```r
soap <- read_csv("http://www.evanlray.com/data/sdm4/Bacterial_Soap.csv")
```

```
## Parsed with column specification:
## cols(
##   `Bacterial Counts` = col_double(),
##   Method = col_character()
## )
```

```r
names(soap) <- c("bacterial_counts", "method")

soap %>%
  count(method)
```
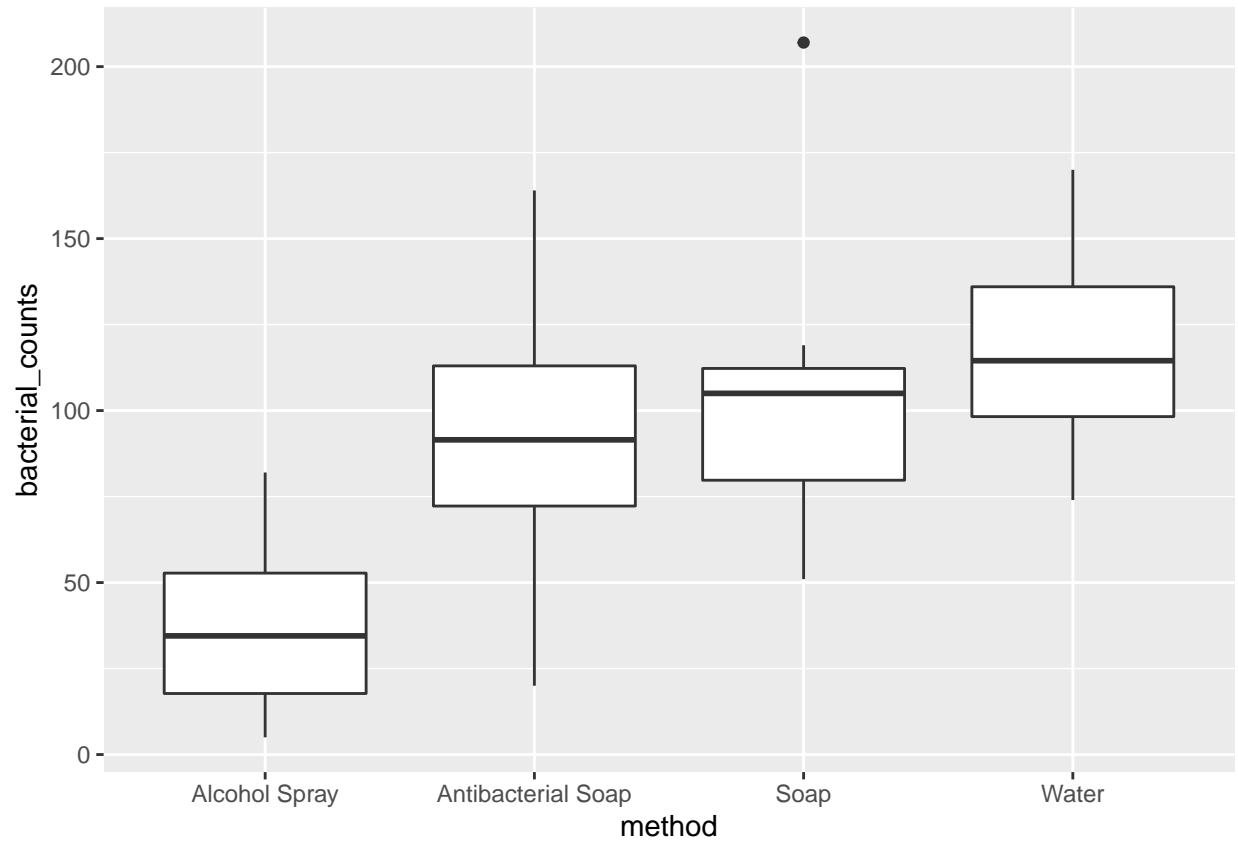
```
## # A tibble: 4 x 2
##   method                 n
##   <chr>              <int>
## 1 Alcohol Spray          8
## 2 Antibacterial Soap     8
## 3 Soap                   8
## 4 Water                  8
```
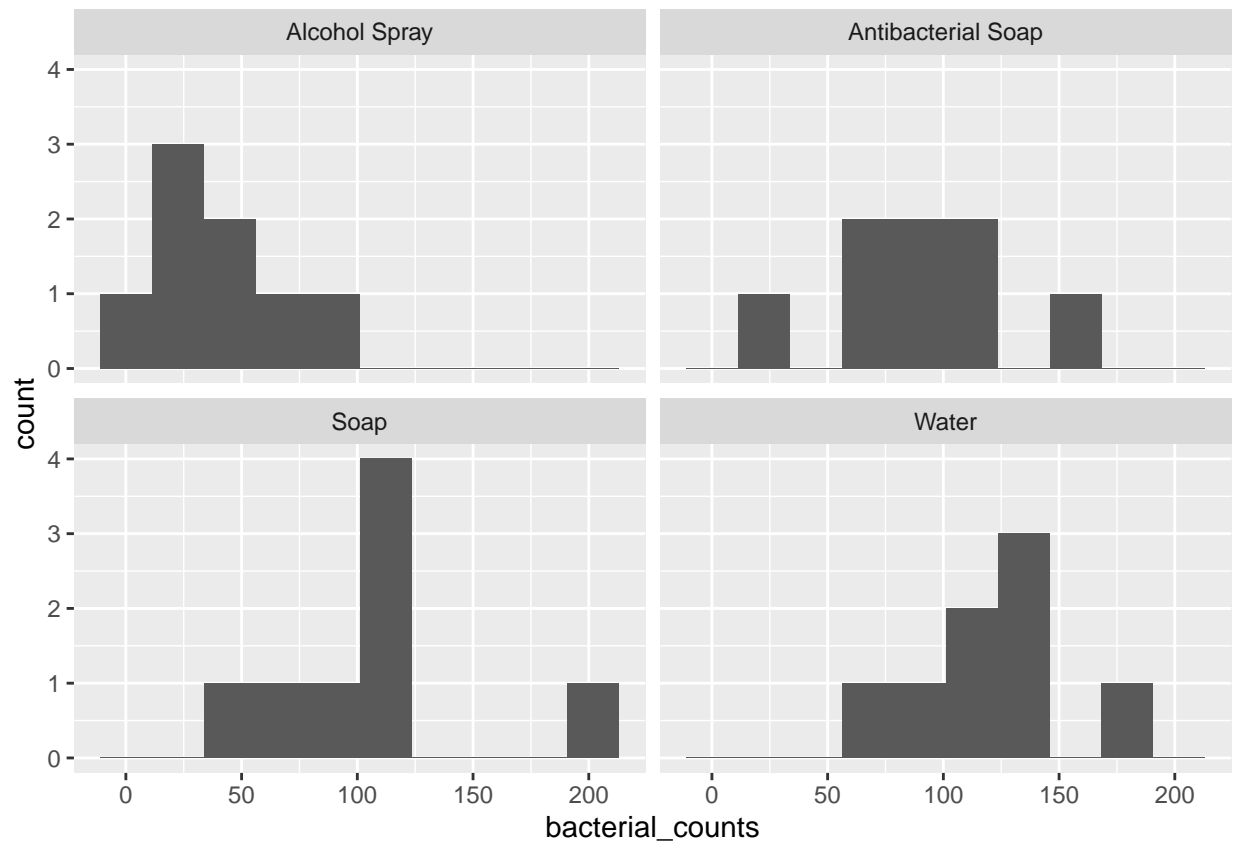
1. Make an appropriate plot of the data.

There are several possibilities - we just want to show the distributions of the four groups in a reasonable way.
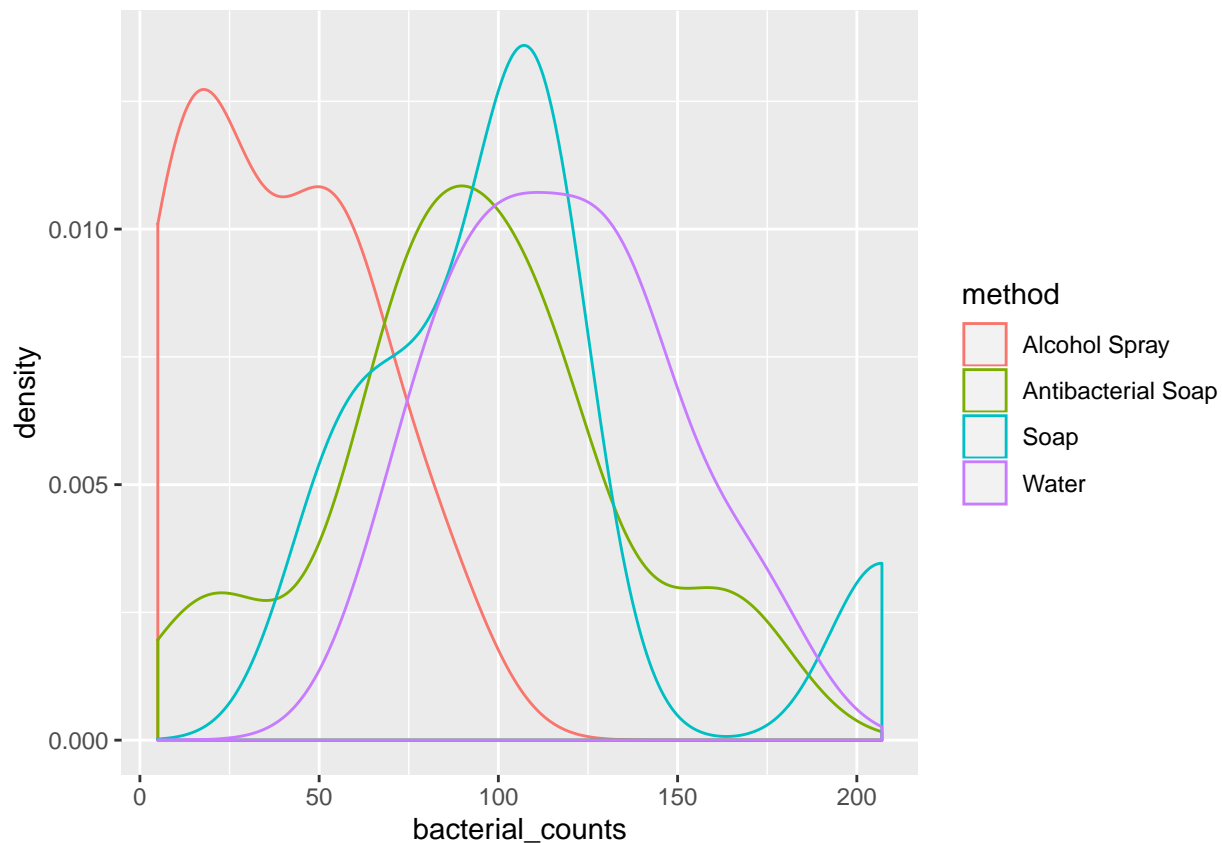
```
## boxplot
p1 <- (ggplot(data=soap, aes(x=method, y=bacterial_counts))
       + geom_boxplot())
p1
```



```
## histogram
p2 <- (ggplot(data=soap, aes(x=bacterial_counts))
       + geom_histogram(bins=10)
       + facet_wrap(~method))
p2
```

```
## density
p3 <- (ggplot(data=soap, aes(x=bacterial_counts))
      + geom_density(aes(col=method)))
p3
```

2. Conduct a test of the claim that all four methods are equally effective.

Define notation:

- $\mu_1 =$ (population) mean bacterial count for alcohol spray
- $\mu_2 =$ (population) mean bacterial count for antibacterial soap
- $\mu_3 =$ (population) mean bacterial count for soap
- $\mu_4 =$ (population) mean bacterial count for water

State hypotheses:

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
- $H_A :$ at least one of the means is different

Conduct test:

```
## Step 1: fit full model
fit_full <- lm(bacterial_counts ~ method, data=soap)

## Step 2: use one of the methods to find the F statistic (here are two)

### Option 1:
summary(fit_full)
```

```
##
## Call:
## lm(formula = bacterial_counts ~ method, data = soap)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72.50 -20.88  -1.00  18.12 101.00
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                37.50      13.28   2.825 0.008629 **
## methodAntibacterial Soap   55.00      18.78   2.929 0.006686 **
## methodSoap                 68.50      18.78   3.648 0.001070 **
## methodWater                79.50      18.78   4.234 0.000224 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.55 on 28 degrees of freedom
## Multiple R-squared:  0.4308, Adjusted R-squared:  0.3698
## F-statistic: 7.064 on 3 and 28 DF,  p-value: 0.001111
```

Option 2:
```
anova(fit_full)
```

```
## Analysis of Variance Table
##
## Response: bacterial_counts
##           Df Sum Sq Mean Sq F value   Pr(>F)
## method     3  29882  9960.7  7.0636 0.001111 **
## Residuals 28  39484  1410.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

State conclusions:

The F test results in an F statistic of 7.064 on 3 and 28 degrees of freedom. The corresponding p-value is 0.001, which gives us very strong evidence that at least one of the means is different. In other words, there is very strong evidence that at least one of the hand washing methods results in a different bacterial count, on average, than the others. It is not appropriate to use the reduced model.

3. Conduct a test of the claim that the "Antibacterial Soap", "Soap", and "Water" methods are equally effective.

Define notation:

- We'll use the same notation we defined in (2).

State hypotheses:

- $H_0 : \mu_2 = \mu_3 = \mu_4$
- $H_A :$ at least one of the means is different

Conduct test:

```
## Step 1: fit full model (I'll skip this step since we did it already)

## Step 2: fit reduced model

### First we need to create a new variable with two groups, one for alcohol
### spray and one for all the other methods
soap <- soap %>% mutate(grouped_methods = ifelse(method %in% c("Antibacterial Soap",
    "Soap", "Water"), "grouped", "Alcohol Spray"))

fit_reduced <- lm(bacterial_counts ~ grouped_methods, data = soap)

## Step 3: anova
anova(fit_reduced, fit_full)
```

```
## Analysis of Variance Table
##
## Model 1: bacterial_counts ~ grouped_methods
## Model 2: bacterial_counts ~ method
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     30 41893
## 2     28 39484  2    2409.3 0.8543 0.4364
```

State conclusions:

The F test results in an F statistic of 0.8543 on 2 and 28 degrees of freedom. The corresponding p-value is 0.4364, which gives us no compelling evidence that at least one of these three hand washing methods results in a different bacterial count, on average, than the others. It is appropriate to use the reduced model with two means, one for alcohol spray and one for the others.

4. If you have extra time and want a refresher, calculate the mean and standard deviation of observations in each group.

```
soap %>%
  group_by(method) %>%
  summarise(
    mean(bacterial_counts),
    sd(bacterial_counts)
  )
```

```
## # A tibble: 4 x 3
##   method             `mean(bacterial_counts)` `sd(bacterial_counts)`
##   <chr>                                 <dbl>                  <dbl>
## 1 Alcohol Spray                          37.5             26.55990534
## 2 Antibacterial Soap                     92.5             41.96256835
## 3 Soap                                   106              46.95894864
## 4 Water                                  117              31.13105936
```