Lab 5 - Transformations for ANOVA

Solutions

Goals

The goal in this lab is to practice working with transformations for ANOVA.

Loading packages

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

```
library(readr)
library(ggplot2)
library(gridExtra)
library(mosaic)
library(dplyr)
options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

A gas chromatograph is an instrument that measures the amounts of various compounds contained in a sample by separating the various constituents. The total number of counts recorded by the chromatograph is proportional to the amount of the compound present.

A calibration experiment was performed to see how the recorded counts from the chromatograph related to the concentration of a compound in a mixture and the flow rate through the chromatograph. In this lab we will just look at the relationship between the concentration (explanatory variable) and the counts (response variable).

```
chromatography <- read_csv("http://www.evanlray.com/data/sdm3/Chapter_29/Ch29_Chromatography.csv")
```

```
## Parsed with column specification:
## cols(
##   Concentration = col_character(),
##   `Flow Rate` = col_character(),
##   Counts = col_double()
## )
```

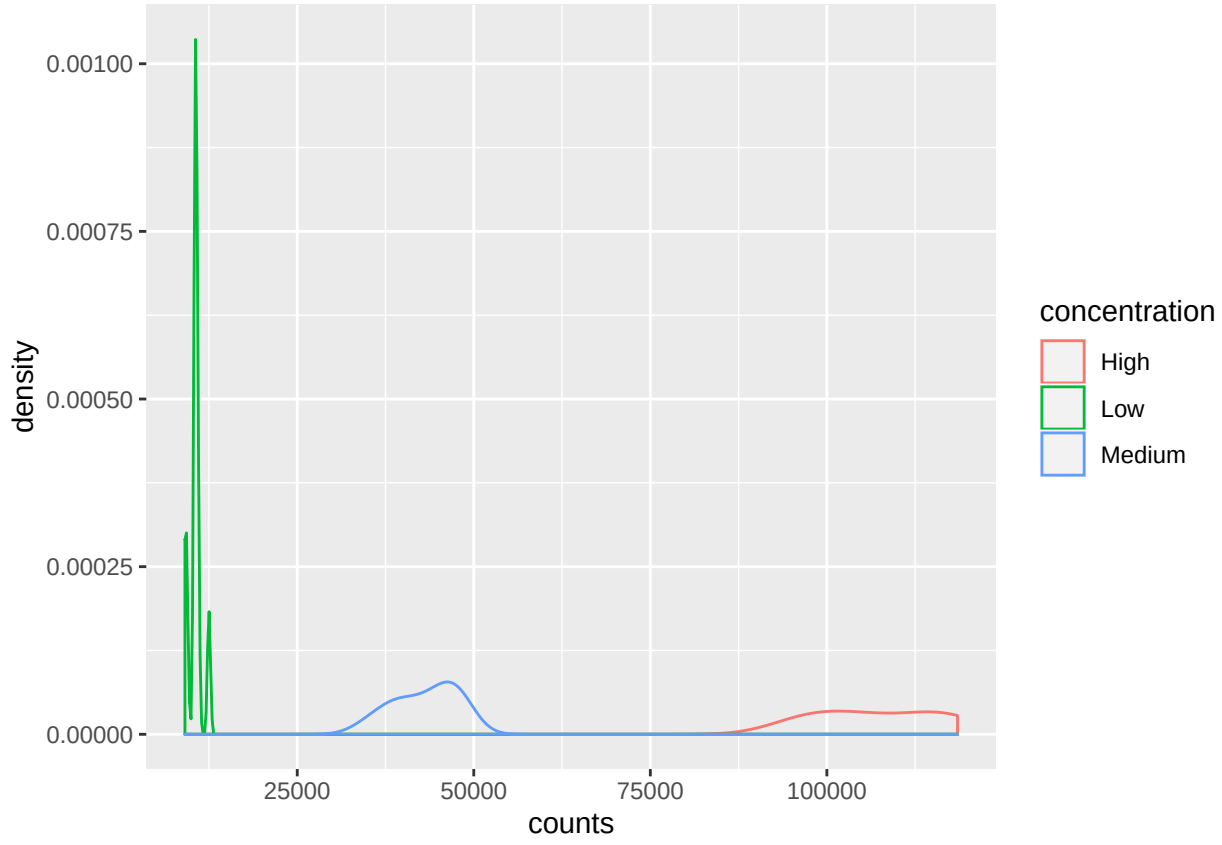
```
names(chromatography) <- c("concentration", "flow_rate", "counts")
chromatography %>%
  count(concentration)
```

```
## # A tibble: 3 x 2
##   concentration     n
##   <chr>         <int>
## 1 High             10
## 2 Low              10
## 3 Medium           10
```

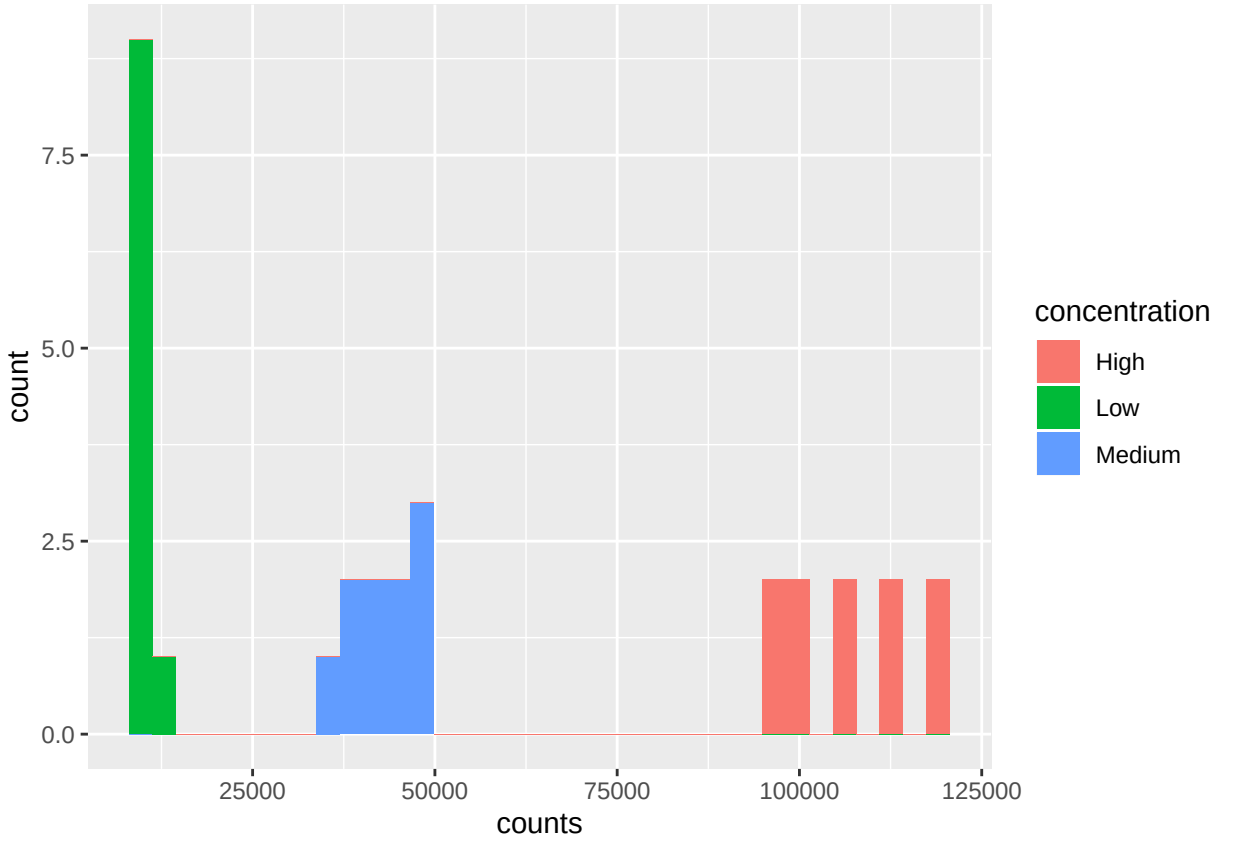
1. Make an appropriate plot of the data: it might be nice to use a histogram or density plot, separately for each value of cylinders. Also calculate the standard deviation for each group. Would it be appropriate to use an ANOVA model for these data?

Here are some possible plots you could consider. Others may be appropriate, too.

```
## density plot
ggplot(data = chromatography, aes(x = counts, color = concentration)) + geom_density()
```




```
## histogram
ggplot(data = chromatography, aes(x = counts, fill = concentration)) + geom_histogram(bins = 35)
```




```
## Calculate standard deviations
chromatography %>% group_by(concentration) %>% summarize(sd_counts = sd(counts))
```

```
## # A tibble: 3 x 2
##   concentration    sd_counts
##   <chr>                <dbl>
## 1 High          8641.856796
## 2 Low            915.9718579
## 3 Medium        4497.556868
```

No, ANOVA would not be appropriate to use for these data (in their untransformed state). Examination of the density plot calls into question the equal variance assumption. Normality may be okay - the densities look symmetric for the most part. Also, calculation of the standard deviations for each group show that equal variance is an unreasonable assumption.

2. Find a transformation of the data so that the ANOVA model would be appropriate.

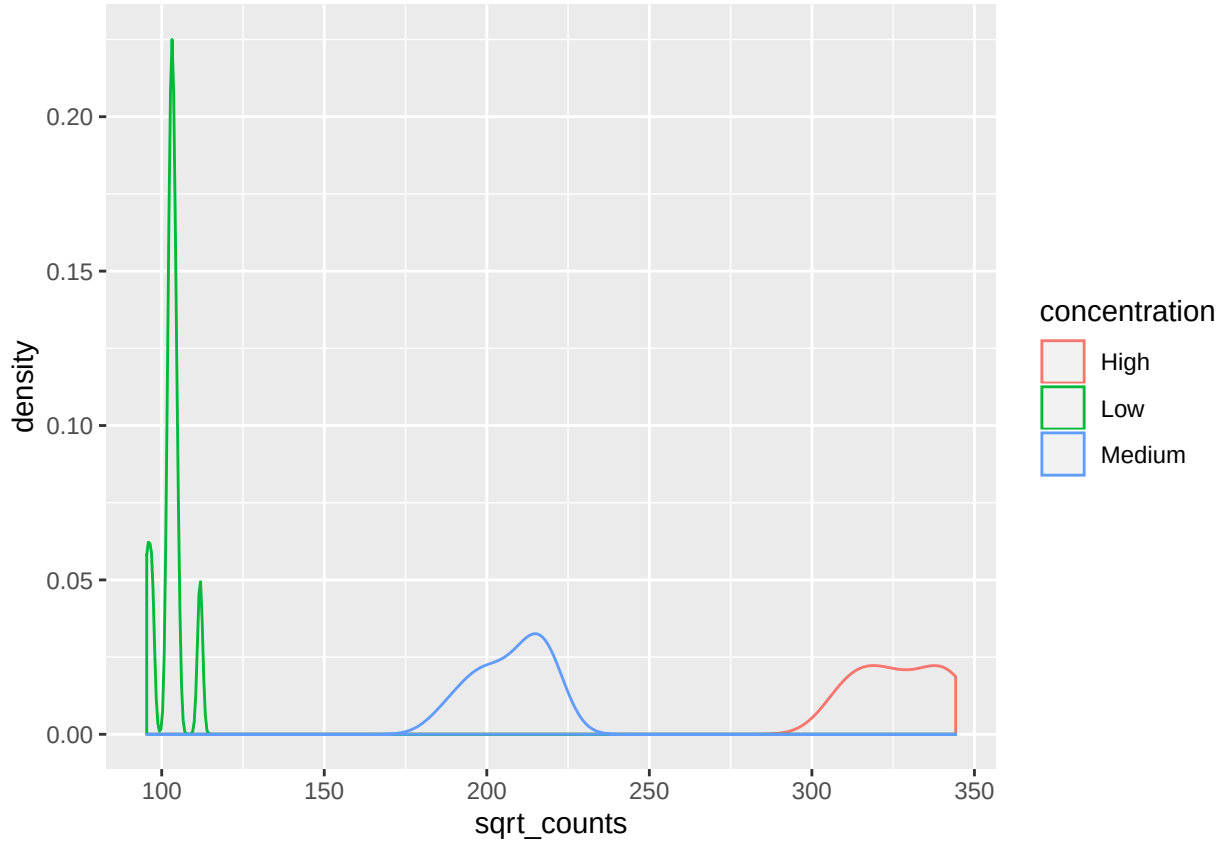
The group with largest mean also has the largest standard deviation, so we can think about moving down the ladder for a transformation.

sqrt transformation:

```
chromatography <- chromatography %>%
  mutate(
    sqrt_counts = sqrt(counts)
```

```
  )

## density plots
ggplot(data=chromatography, aes(x=sqrt_counts, color=concentration)) +
  geom_density()
```




```
## Calculate standard deviations
chromatography %>%
  group_by(concentration) %>%
  summarize(
    sd_sqrt_counts = sd(sqrt_counts)
  )
```

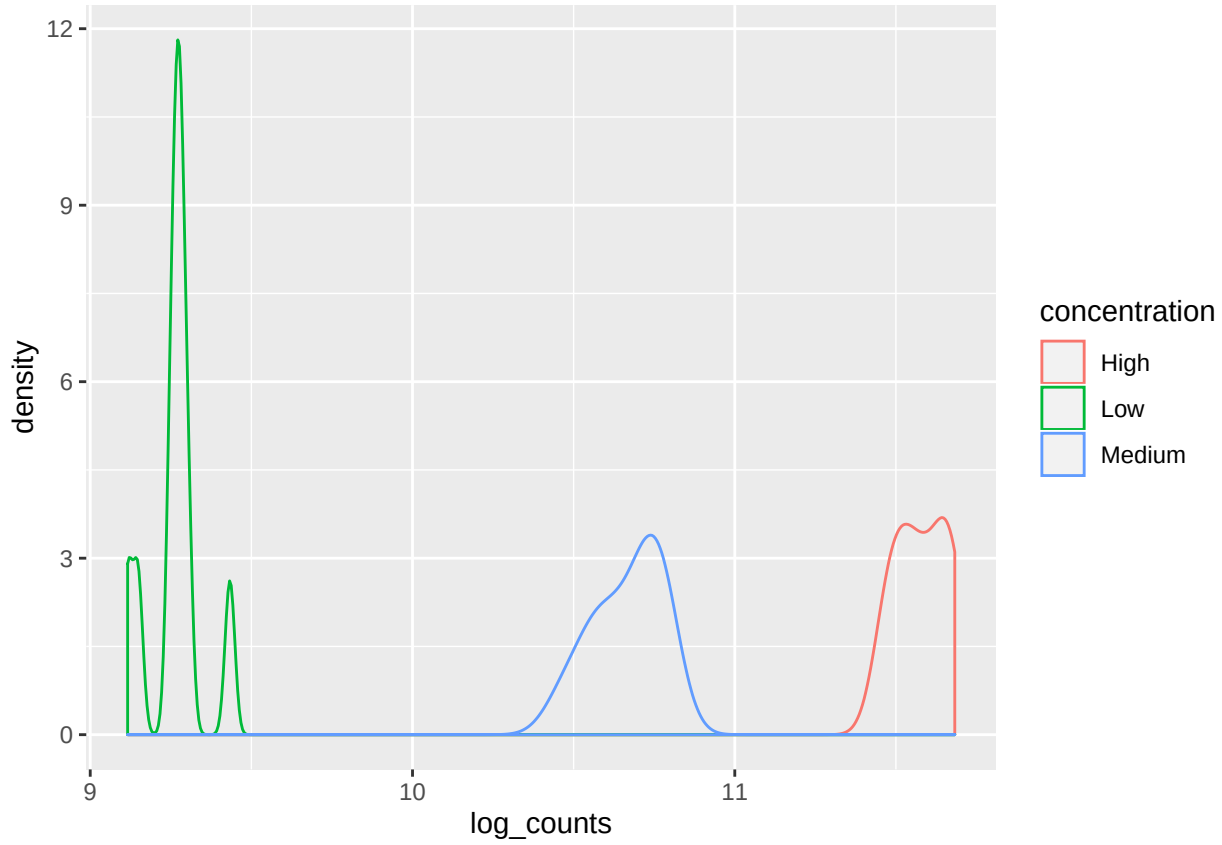
```
## # A tibble: 3 x 2
##   concentration sd_sqrt_counts
##   <chr>                  <dbl>
## 1 High            13.21413071
## 2 Low              4.434431352
## 3 Medium          10.97834749
```

This improves things, but the equal variance assumption is still problematic.

log transformation:

```r
chromatography <- chromatography %>%
  mutate(
    log_counts = log(counts)
  )

## density plots
ggplot(data=chromatography, aes(x=log_counts, color=concentration)) +
  geom_density()
```




```r
## Calculate standard deviations
chromatography %>%
  group_by(concentration) %>%
  summarize(
    sd_log_counts = sd(log_counts)
  )
```

```
## # A tibble: 3 x 2
##   concentration sd_log_counts
##   <chr>                 <dbl>
## 1 High          0.08090121936
## 2 Low           0.08611873813
## 3 Medium        0.1074039081
```

Let's go with the log transformation - this looks good.

3. Conduct a test of the claim that the mean count is the same for all three concentration levels.

Define parameters:

μ_1: mean log count for low concentration μ_2: mean log count for medium concentration μ_3: mean log count for high concentration

State hypotheses:

$H_0 : \mu_1 = \mu_2 = \mu_3$ $H_A :$ at least one of these group means is different

Run code:

```
lm_logcounts <- lm(log_counts ~ concentration, data=chromatography)
anova(lm_logcounts)
```

```
## Analysis of Variance Table
##
## Response: log_counts
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## concentration  2 27.2608 13.6304  1603.8 < 2.2e-16 ***
## Residuals     27  0.2295  0.0085
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation of results:

There is very strong evidence that at least one of the log means for these three groups is different than the others. It is not appropriate to use a model with just one mean.

4. Report and interpret an estimate of the difference in the centers of the distributions of counts for the high concentration and the low concentration, as well as a 95% confidence interval for that difference. You should be able to do this in a few different ways.

```
summary(lm_logcounts)
```

```
##
## Call:
## lm(formula = log_counts ~ concentration, data = chromatography)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19663 -0.05638  0.01090  0.06481  0.17162
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         11.57961    0.02915  397.20   <2e-16 ***
## concentrationLow    -2.31775    0.04123  -56.22   <2e-16 ***
## concentrationMedium -0.91361    0.04123  -22.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09219 on 27 degrees of freedom
## Multiple R-squared:  0.9917, Adjusted R-squared:  0.991
## F-statistic:  1604 on 2 and 27 DF,  p-value: < 2.2e-16
```

```r
confint(lm_logcounts)
```

```
##                           2.5 %     97.5 %
## (Intercept)          11.519790 11.6394240
## concentrationLow     -2.402342 -2.2331533
## concentrationMedium -0.998200 -0.8290115
```

The estimated difference in mean log counts between high and low concentration is -2.318. We are 95% confident that the difference in mean log counts between high and low concentration is between -2.402 and -2.233. If we transform this back to the original scale, then we are 95% confident that the mean count for high concentration is between 0.091 and 0.107 times the mean count for low concentration. (Or, we are 95% confident that the mean count for low concentration is between 9.329 and 11.049 times the mean count for high concentration.)