Lab05 - Transformations for ANOVA

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)
library(gmodels)

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_integer()
## )
```

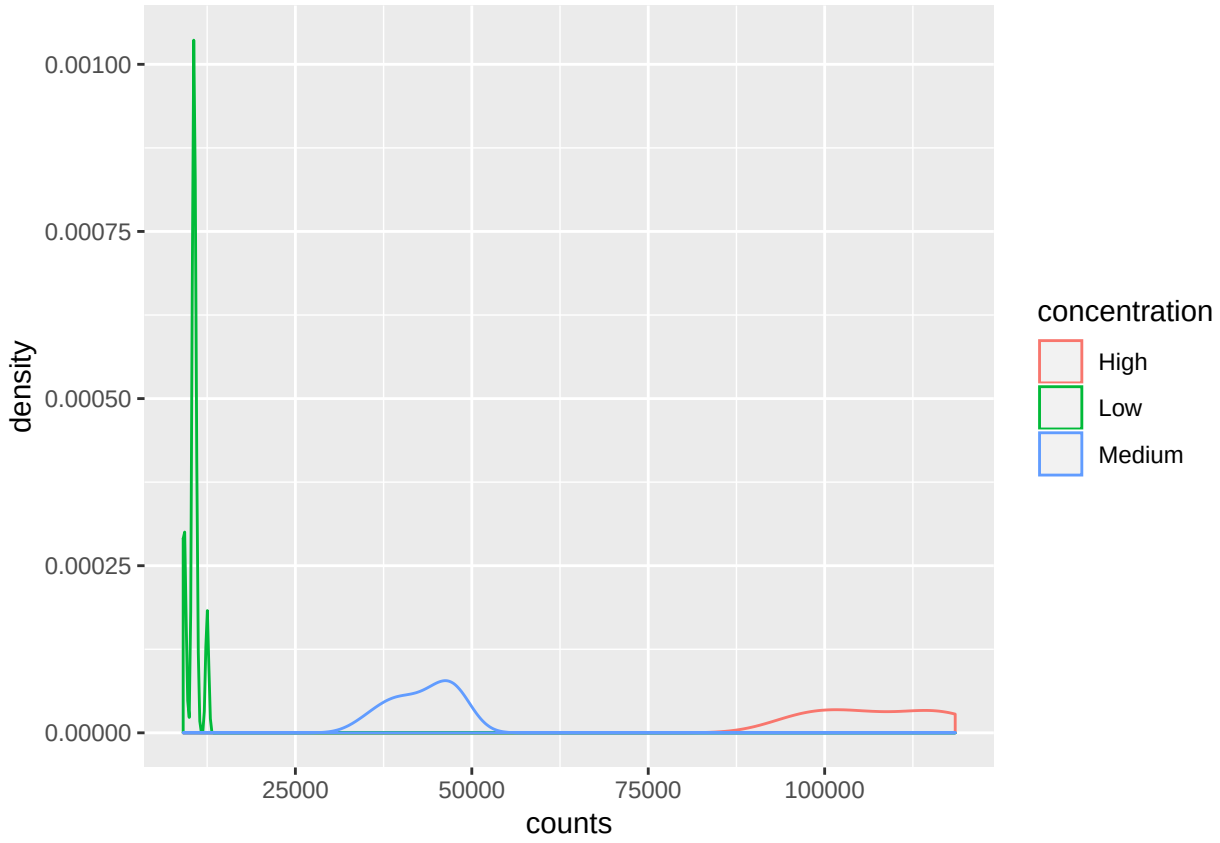
```
names(chromatography) <- c("concentration", "flow_rate", "counts")

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?

```
ggplot(data = chromatography, mapping = aes(x = counts, color = concentration)) +
  geom_density()
```




```
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
```

It would not be appropriate to analyze these data using the standard ANOVA model since the standard deviations of counts is very different for the three groups.

I didn't ask us to formally check all conditions, but to use an ANOVA model we would also have to check the following conditions:

Independence: unclear from the information given. We would need to assume that the people conducting the experiment ensured measurements for the different samples were independent, for example by thoroughly cleaning the apparatus between each run. Although we aren't given much information, this seems like a reasonable assumption.

Normally distributed errors: The distribution within each group is not exactly normal, but doesn't appear to be so skewed that this would be a problem.

Outliers. There are not any serious outliers to be concerned about.

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

Since the overall trend in the data was that the groups with larger means had larger variances, we need to step down on the ladder. Put another way, overall across the three groups, the data are skewed right.

Below are the group standard deviations for the next three steps down on the ladder of powers.

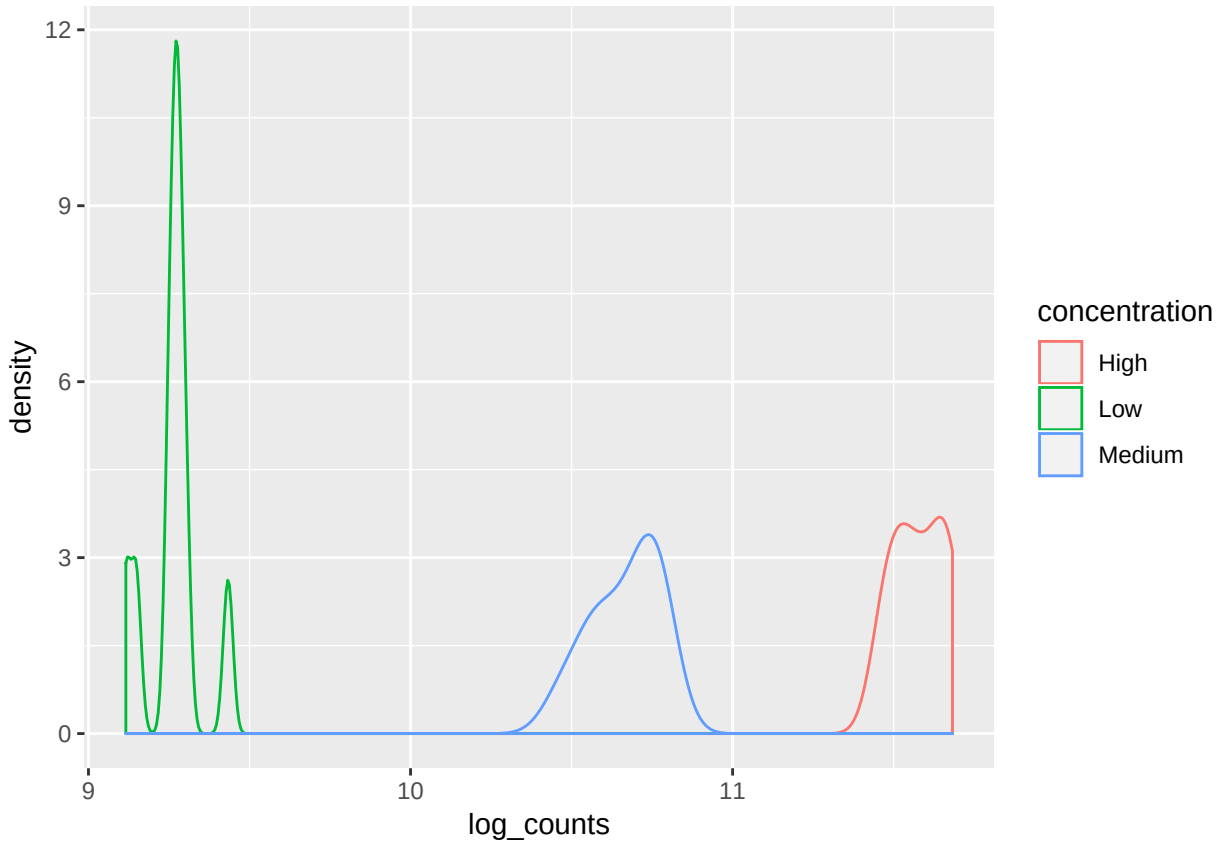
```
chromatography <- chromatography %>%
  mutate(
    sqrt_counts = sqrt(counts),
    log_counts = log(counts),
    neg_inv_sqrt_counts = -1/sqrt(counts)
  )

chromatography %>%
  group_by(concentration) %>%
  summarize(
    sd_sqrt = sd(sqrt_counts),
    sd_log = sd(log_counts),
    sd_neg_inv_sqrt = sd(neg_inv_sqrt_counts)
  )
```

```
## # A tibble: 3 x 4
##   concentration      sd_sqrt        sd_log sd_neg_inv_sqrt
##   <chr>                <dbl>         <dbl>           <dbl>
## 1 High          13.21413071  0.08090121936 0.0001239476822
## 2 Low            4.434431352 0.08611873813 0.0004192992433
## 3 Medium        10.97834749  0.1074039081  0.0002632217346
```

The group standard deviations are quite different for the square root transformation and the negative inverse square root transformation ($-1/\sqrt{y}$). The square root transformation is not strong enough, and the negative inverse square root transformation goes too far. We will use the log transformation, where the group standard deviations are roughly similar. To confirm, here's a plot:

```
ggplot(data = chromatography, mapping = aes(x = log_counts, color = concentration)) +
  geom_density()
```



2. Conduct a test of the claim that the mean count is the same for all three concentration levels, on the transformed scale.

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

```
## 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
```

Define μ_1 to be the mean log count recorded by the chromatograph at the high concentration (in a population of similarly prepared samples), μ_2 the mean log count at the low concentration, and μ_3 the mean log count at the medium concentration. (These are in alphabetic order.)

$H_0 : \mu_1 = \mu_2 = \mu_3$. The mean log counts are the same for all three concentration levels.

H_A : At least one of μ_1, μ_2, and μ_3 is not equal to the others. The mean log counts are not the same for all three concentration levels.

The p-value for the test is less than 2.2×10^{-16}. The data provide extremely strong evidence that the average of the log counts is not equal for all three concentration levels.

3. 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.

We are interested in $\gamma = \mu_1 - \mu_2$, so we will use the constants $(1, -1, 0)$ in defining the linear combination of means to study.

```
library(gmodels)
fit.contrast(anova_fit, "concentration", c(1, -1, 0), conf.int = 0.95)
```

```
##                             Estimate Std. Error  t value     Pr(>|t|)
## concentration c=( 1 -1 0 ) 2.317748 0.04122867 56.21689 1.625873e-29
##                             lower CI upper CI
## concentration c=( 1 -1 0 ) 2.233153 2.402342
```

Interpretation on the log scale:

We estimate that the difference in mean log counts between the high concentration and the low concentration is about 2.32 log counts, with a confidence interval of 2.23 to 2.40 log counts.

Interpretation on the original scale:

```
exp(2.32)
```

```
## [1] 10.17567
```

```
exp(2.23)
```

```
## [1] 9.299866
```

```
exp(2.40)
```

```
## [1] 11.02318
```

We estimate that the median count for the high concentration is about 10.18 times higher than the median count for the low concentration; a 95% confidence interval for this multiplicative difference between median counts for those groups is between 9.30 and 11.02.

I didn't ask you to do this, but it's interesting to note that this corresponds fairly well to the ratio of the medians for those groups in our sample, which could be viewed as a different estimate of the ratio of population medians.

```
chromatography %>%
  group_by(concentration) %>%
  summarize(
    median_counts = median(counts)
  )
```

```
## # A tibble: 3 x 2
##   concentration median_counts
##   <chr>                 <dbl>
## 1 High                 106765
## 2 Low                   10650
## 3 Medium                43910
```

```
106765/10650
```

```
## [1] 10.02488
```