# HW5: Chapter 3, Section 5.5, Section 6.3, Section 6.4

*Solutions*

The code below just loads some packages and makes it so that enough digits are printed that you won't get confused by rounding errors.

```
library(dplyr) # functions like summarize
library(ggplot2) # for making plots
library(mosaic) # convenient interface to t.test function
library(readr)
library(gmodels)
options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

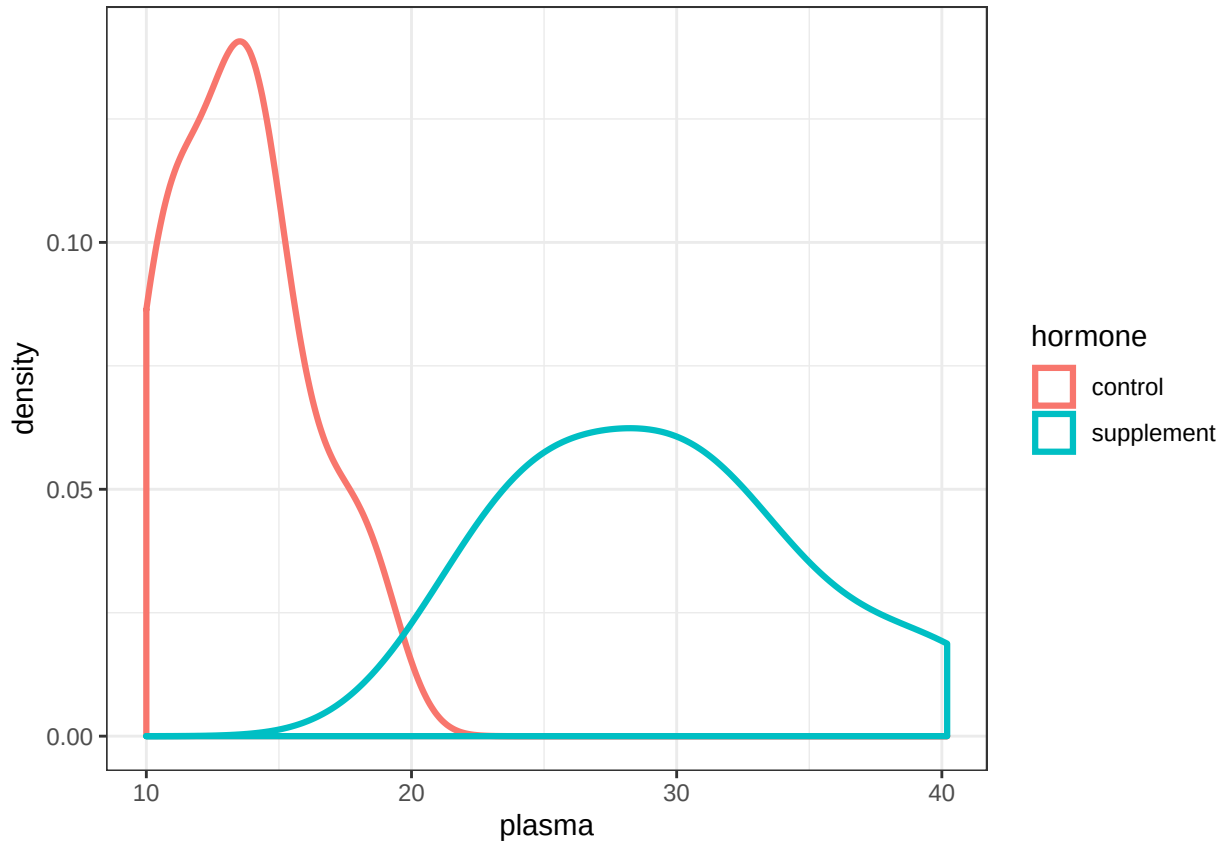
## Problem 1: Bird Calcium

For many animals, the body's ability to use calcium depends on the level of certain sex-related hormones in the blood. The following data set looks at the relationship between hormone supplement (present or absent) and level of calcium in the blood. The subjects were 20 birds. Half the birds got a hormone supplement and the others served as controls. The response is the level of plasma calcium in mg/100 ml.

```
birds <- read.csv("http://www.evanlray.com/data/cobb_doe/bird_calcium_p160.csv") %>%
  transmute(
    hormone = ifelse(hormone == 1, "control", "supplement"),
    plasma = plasma
  )
```

**(a) Check the conditions for conducting an analysis of these data with an ANOVA model. You should write an explicit sentence for each condition explaining why it is or isn't satisfied, with justification; if you need more information to make a determination, explain what else you would need to know. If necessary, find a transformation of the data so that the conditions are as well satisfied as possible.**

```
## Check assumptions graphically (normality, equal variance, outliers)
ggplot(data=birds, aes(x=plasma, color=hormone)) + geom_density(size=1.1) + theme_bw()
```




```
## Check assumption of equal variance by examining standard deviations
birds %>%
  group_by(hormone) %>%
  summarize(
    sd_plasma = sd(plasma)
  )
```

```
## # A tibble: 2 x 2
##   hormone      sd_plasma
##   <chr>            <dbl>
## 1 control    2.620856518
## 2 supplement 5.749792267
```

Conditions:

- Independent observations: Given the available description, we do not have enough information to say definitively whether the observations are independent. For example, we do not know how the birds were chosen for inclusion in the study. If birds were selected from multiple locations, we might expect that birds taken from the same location have more similar calcium blood levels (and thus are not independent). Similarly, male and female birds react to the hormone supplement differently, so knowing one residual is positive (or negative) would give me information about whether another residual is likely to be positive (or negative). **Unknown.**
- Normally distributed: The distributions shown in the density plots are close enough to normally distributed (within each group) for $t$-based inferences to be approximately valid. We are able to proceed without a transformation. **Satisfied.**

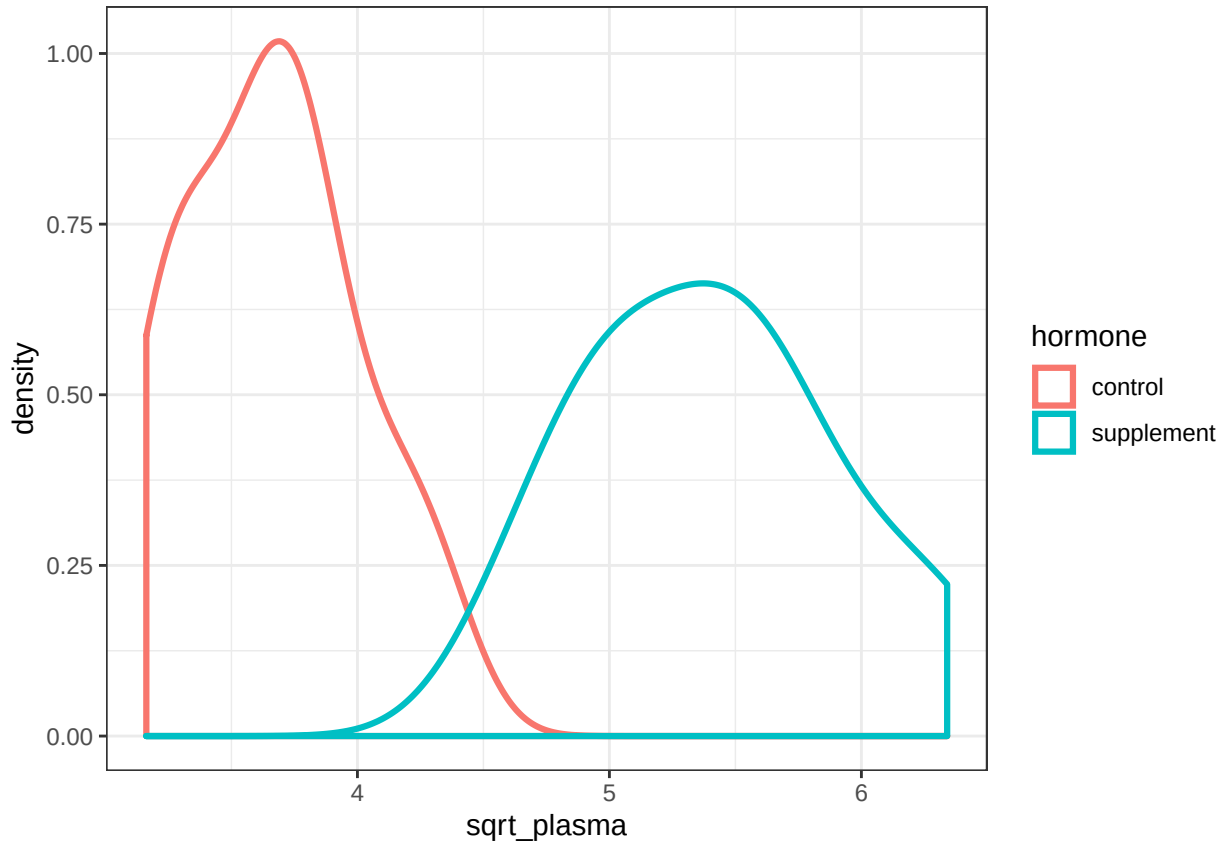
- Equal variances (in each group): This condition is not satisfied; the standard deviation for the group that took the supplement is more than twice as large as the standard deviation for the control group. **Not satisfied - will need a transformation.**
- No outliers: There are no apparent outliers. **Satisfied.**

Now, we need to look for a suitable transformation to address the unequal variances. The group with a larger mean also has a larger standard deviation, so we will try moving down the ladder of powers (this will make all the values smaller and should shrink the variances).

**First attempt - sqrt transformation:**

```
## Add another column: the sqrt transformation
birds <- birds %>%
  mutate(
    sqrt_plasma = sqrt(plasma)
  )

## plot the density curves to assess shape and spread
ggplot(data=birds, aes(x=sqrt_plasma, color=hormone)) + geom_density(size=1.1) + theme_bw()
```




```
## compute the standard deviation for each group
birds %>%
  group_by(hormone) %>%
  summarize(
    sd_plasma = sd(sqrt_plasma)
  )
```

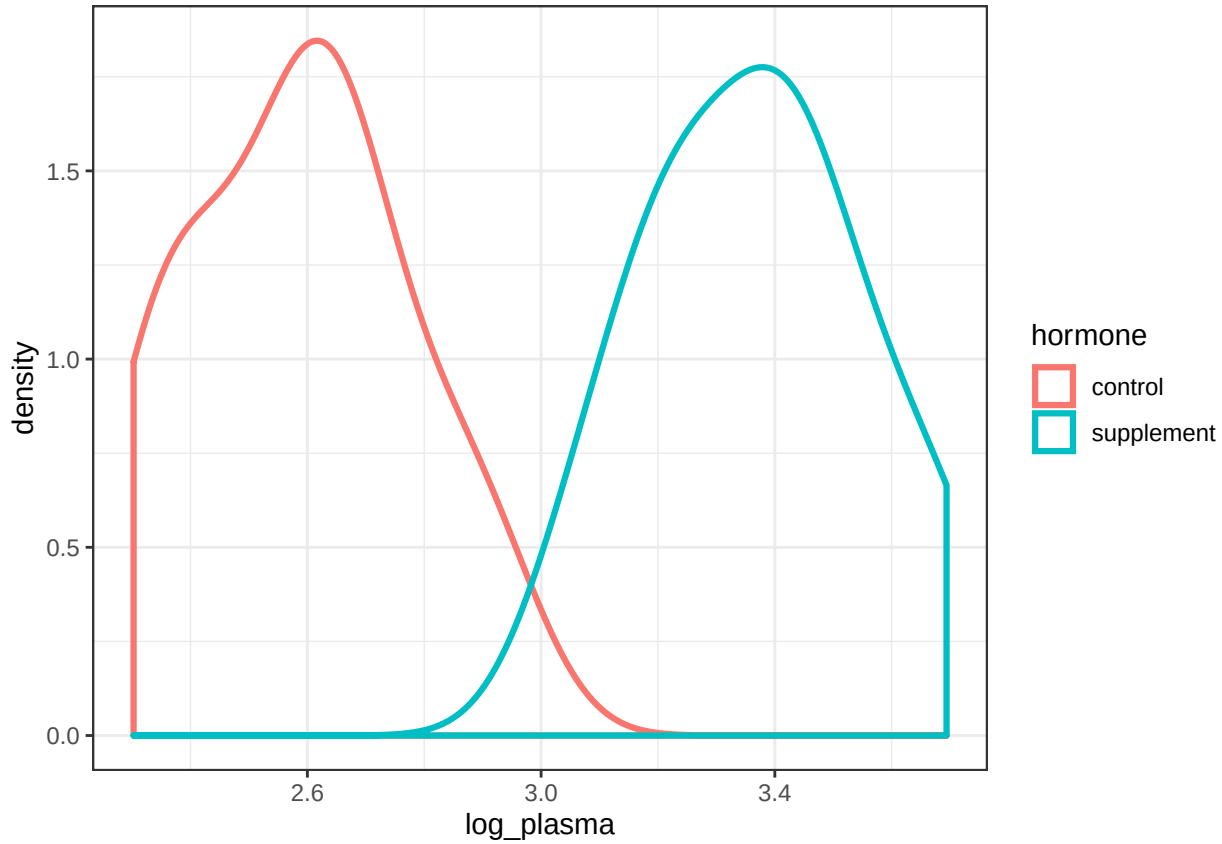
```
## # A tibble: 2 x 2
##   hormone       sd_plasma
##   <chr>             <dbl>
## 1 control    0.3531632690
## 2 supplement 0.5249634181
```

The square root transformation improves things, and now we satisfy the rule of thumb that the ratio of standard deviations should not exceed 2. It is okay to proceed with the analysis on this scale, but we are going to go one more step down the ladder and see if we can do any better.

**Second attempt - log transformation:**

```
## Add another column: the log transformation
birds <- birds %>%
  mutate(
    log_plasma = log(plasma)
  )

## plot the density curves to assess shape and spread
ggplot(data=birds, aes(x=log_plasma, color=hormone)) +
  geom_density(size=1.1) +
  theme_bw()
```




```
## compute the standard deviation for each group
birds %>%
  group_by(hormone) %>%
```

```
  summarize(
    sd_plasma = sd(log_plasma)
  )
```

```
## # A tibble: 2 x 2
##   hormone       sd_plasma
##   <chr>             <dbl>
## 1 control    0.1919166671
## 2 supplement 0.1933674427
```

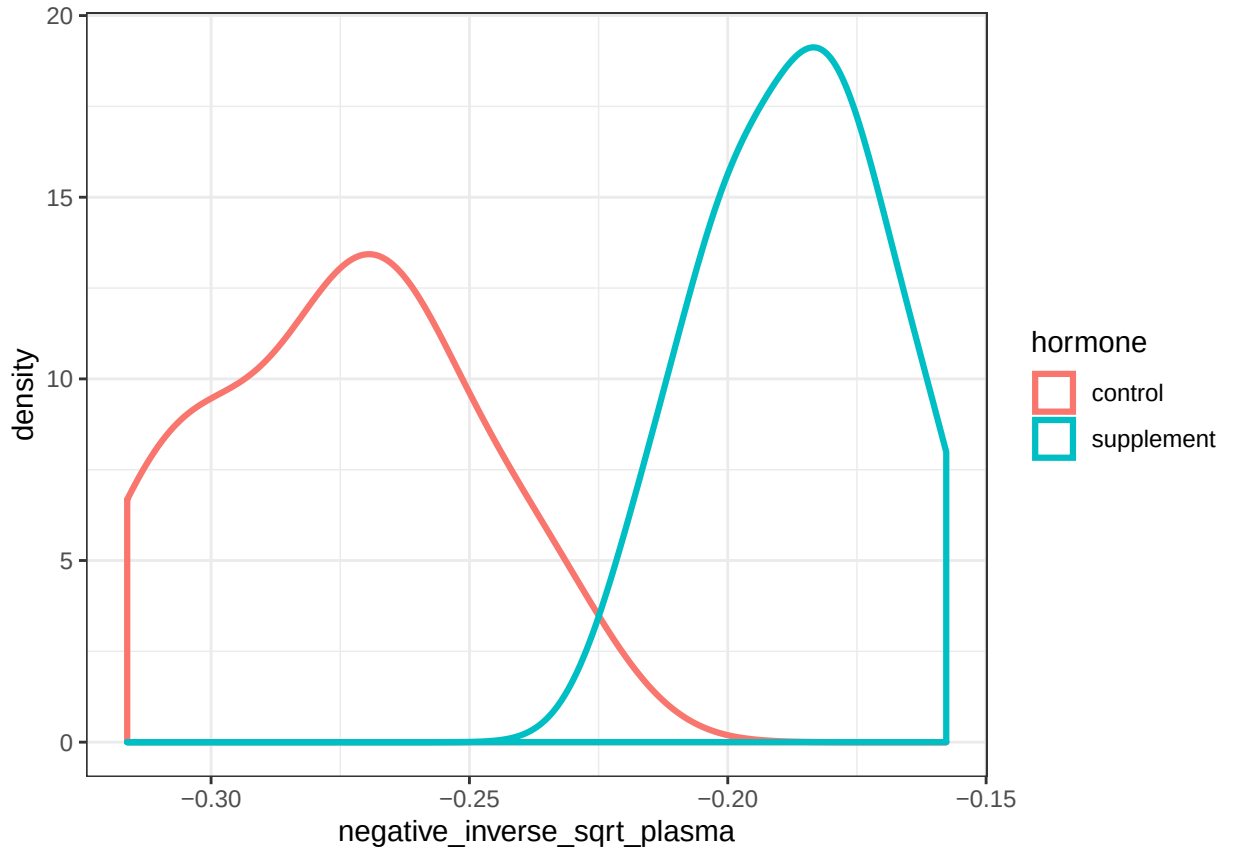
When we examine the assumption of equal variance on the log scale, we see that the standard deviations between the two groups are basically equal. This looks even better than the square root transformation. This is where I would stop.

Note: you could think about going one step further, to see what happens if you continue down the ladder. Do things just get better and better?

**Food for thought - $-y^{-0.5}$ transformation:**

```
## Add another column: the log transformation
birds <- birds %>%
  mutate(
    negative_inverse_sqrt_plasma = -1/sqrt(plasma)
  )

## plot the density curves to assess shape and spread
ggplot(data=birds, aes(x=negative_inverse_sqrt_plasma, color=hormone)) +
  geom_density(size=1.1) +
  theme_bw()
```




```
## compute the standard deviation for each group
birds %>%
  group_by(hormone) %>%
  summarize(
    sd_plasma = sd(negative_inverse_sqrt_plasma)
  )
```

```
## # A tibble: 2 x 2
##   hormone        sd_plasma
##   <chr>              <dbl>
## 1 control    0.02628381583
## 2 supplement 0.01795963899
```

With this transformation, the standard deviations are noe more dissimilar (nearly back to the 2:1 that we had before we transformed the data). Stick to the log transformation!

**(b) For the purpose of this problem, let's assume that the conditions you checked in part (b) were fairly well satisfied (perhaps after suitable transformation). Conduct a test to find out whether there were any differences in the mean level of plasma calcium for birds taking the hormones and the control group (perhaps after suitable transformation). Please define all parameters involved, state your hypotheses in terms of equations involving the parameters and written sentences explaining what the hypotheses mean in context, and interpret the p-value for your test in terms of strength of evidence against the null hypothesis of the test, stated in context.**

Define parameters:

- $\mu_1$ = mean of log plasma calcium in the population of birds not taking a hormone supplement
- $\mu_2$ = mean of log plasma calcium in the population of birds taking a hormone supplement

State hypotheses:

- $H_0 : \mu_1 = \mu_2$ (or $H_0 : \mu_2 - \mu_1 = 0$ for consistency with code). The average log plasma calcium is the same in the population of birds taking hormone supplements and not taking hormone supplements.
- $H_A : \mu_1 \neq \mu_2$ (or $H_0 : \mu_2 - \mu_1 \neq 0$ for consistency with code). The average log plasma calcium is not the same in the population of birds taking hormone supplements and not taking hormone supplements.

```
## Fit the linear model
model_fit <- lm(log_plasma ~ hormone, data=birds)

## Carry out the hypothesis test; C1=-1, C2=1
fit.contrast(model_fit, "hormone", c(-1,1))
```

```
##                     Estimate Std. Error t value     Pr(>|t|)
## hormone c=( -1 1 ) 0.7790897 0.08615276 9.04312 4.101525e-08
## attr(,"class")
## [1] "fit_contrast"
```

The p-value for the test is $4.102 \times 10^{-8}$. The data provide very strong evidence against the null hypothesis that there is no difference between the group mean plasma counts on the log scale. It appears that the hormone supplement increases the calcium levels in the birds' blood.

Note, you will get equivalent results (up to a minus sign) if you used the contrast c(1,-1).

**(c) Find a 95% confidence interval describing the difference in the centers of the distributions of calcium concentrations between birds without the hormone supplement and birds with the hormone supplement. Interpret your confidence interval in context on the original (untransformed) scale of the data.**

```
fit.contrast(model_fit, "hormone", c(-1,1), conf.int=0.95)
```

```
##                     Estimate Std. Error t value     Pr(>|t|)  lower CI
## hormone c=( -1 1 ) 0.7790897 0.08615276 9.04312 4.101525e-08 0.5980895
##                    upper CI
## hormone c=( -1 1 )  0.96009
## attr(,"class")
## [1] "fit_contrast"
```

Since we performed inference after a log transformation, and the distributions of the transformed data were approximately symmetric, we can reverse the transformation in order to make statements about the multiplicative difference in the means on the untransformed scale. Just exponentiate your confidence bounds!

```
c(round(exp(0.5980895),3), round(exp(0.96009),3))
```

```
## [1] 1.819 2.612
```

We are 95% confident that the mean plasma calcium is between 1.8 and 2.6 times higher for birds taking a hormone supplement than for birds not taking a hormone supplement.

## Problem 2: Pesticides in olive oil

Fenthion is a pesticide used against the olive fruit fly in olive groves. It is toxic to humans, so it is important that there be no residue left on the fruit or in olive oil that will be consumed. One theory was that, if there is residue of the pesticide left in the olive oil, it would dissipate over time. Chemists set out to test that theory by taking a random sample of small amounts of olive oil with fenthion residue and measuring the amount of fenthion in the oil at 3 different times over the year: Day 0 (the day the sample was taken), Day 281, and Day 365.

The following R code reads in the data:

```
olives <- read_csv("http://www.evanlray.com/data/stat2/Olives.csv") %>%
  mutate(
    Day = factor(paste0("Day", Day))
  )
```

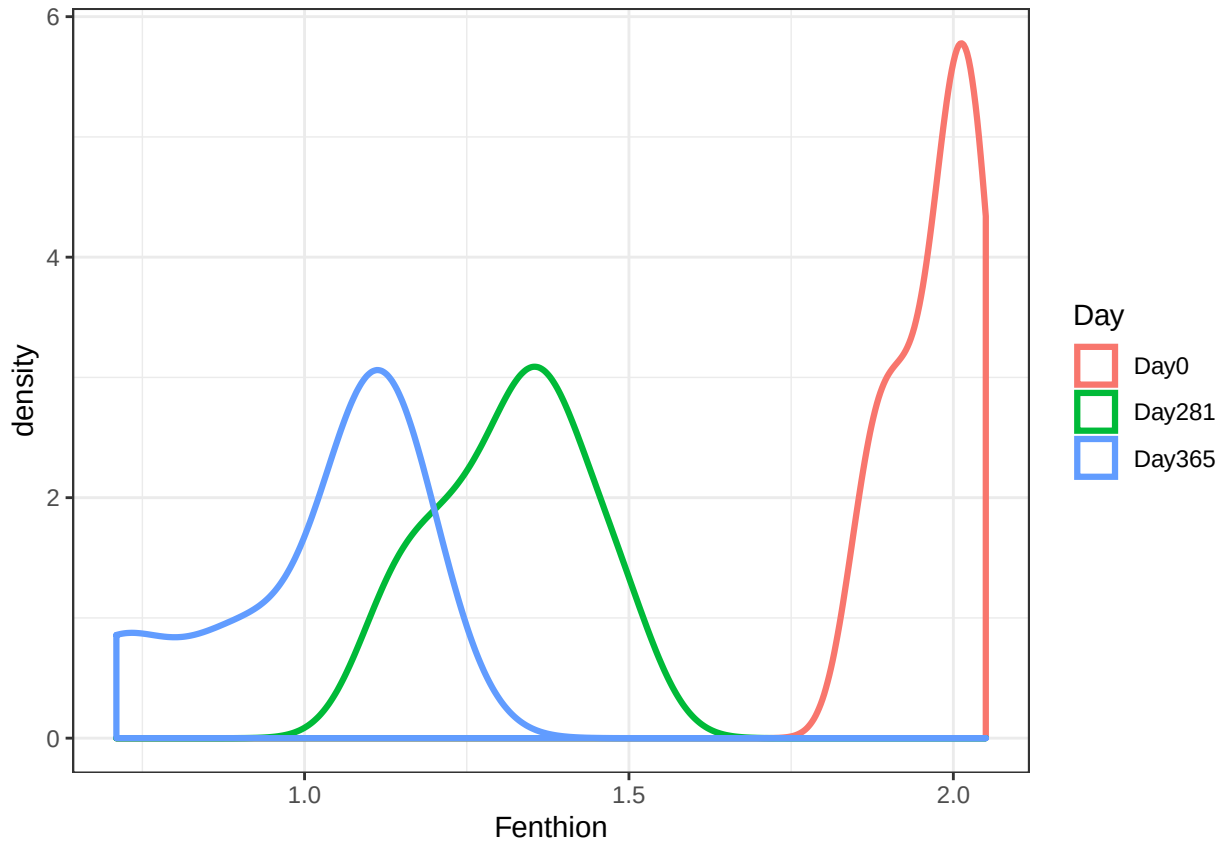
```
## Parsed with column specification:
## cols(
##   SampleNumber = col_double(),
##   Group = col_double(),
##   Day = col_double(),
##   Fenthion = col_double(),
##   FenthionSulphoxide = col_double(),
##   FenthionSulphone = col_double(),
##   Time = col_double()
## )
```

**(a) Two variables in the model are `Fenthion` and `Day`; we will analyze these variables in this problem. Of these variables, which is the explanatory variable and which is the response? Explain.**

Day is the explanatory variable, and Fenthion is the response. We believe that the amount of fenthion in the oil may change over time, or that the day may explain variation in the amount of fenthion in the oil.

**(b) Check the conditions for conducting an analysis of these data with an ANOVA model. You should write an explicit sentence for each condition explaining why it is or isn't satisfied, with justification; if you need more information to make a determination, explain what else you would need to know. If necessary, find a transformation of the data so that the conditions are as well satisfied as possible.**

```
## plot the density curves to assess shape and spread
ggplot(data=olives, aes(x=Fenthion, color=Day)) +
  geom_density(size=1.1) +
  theme_bw()
```




```
## compute the standard deviation for each group
olives %>%
  group_by(Day) %>%
  summarize(
    sd_plasma = sd(Fenthion)
  )
```

```
## # A tibble: 3 x 2
##   Day        sd_plasma
##   <fct>          <dbl>
## 1 Day0   0.06774953874
## 2 Day281 0.1205680997
## 3 Day365 0.1726750320
```

Conditions:

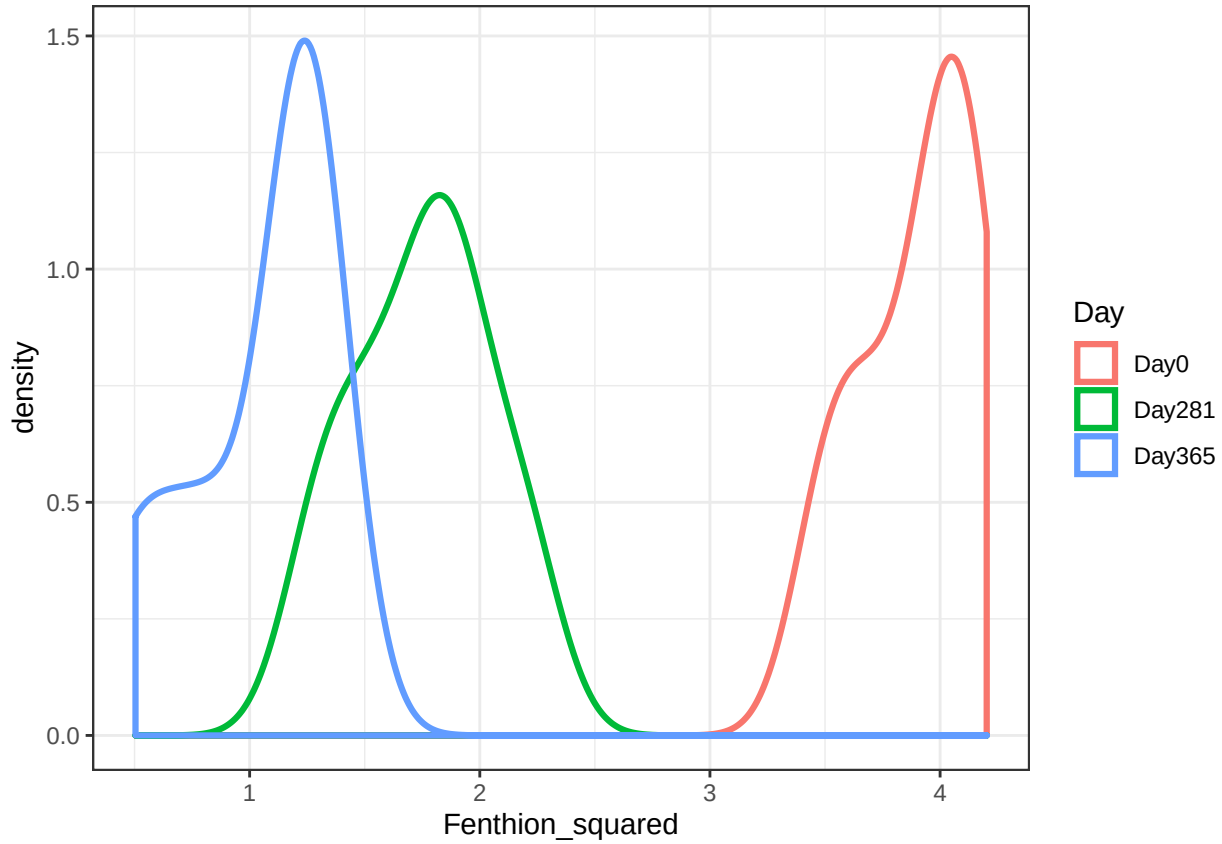
- Independent observations: The observations are not independent. The researchers took a few samples of oil and measured the amount of fenthion in those samples over time. Two different measurements on the same oil sample will not be independent. If an oil sample had a very high fenthion concentration on Day 0, it may still have a high concentration on Day 281 and Day 365. **Not satisfied.**
- Normally distributed: The distributions shown in the density plots are slightly left-skewed, but they are close enough to normally distributed that we can expect for t-based inferences to be approximately valid. **Satisfied.**
- Equal variances (in each group): This condition is not satisfied. The standard deviation is nearly 3 times as large on Day 365 as on Day 0. **Not satisfied.**

- No outliers: There are no apparent outliers. **Satisfied.**

To make the standard deviations more similar to each other, we try a transformation. Since the group with the smaller mean has a larger standard deviation, we will try moving up the ladder of powers. First, we will try a square transformation:

```
## Add another column: the square transformation
olives <- olives %>%
  mutate(
    Fenthion_squared = Fenthion^2
  )

## plot the density curves to assess shape and spread
ggplot(data=olives, aes(x=Fenthion_squared, color=Day)) +
  geom_density(size=1.1) +
  theme_bw()
```




```
## compute the standard deviation for each group
olives %>%
  group_by(Day) %>%
  summarize(
    sd_Fenthion = sd(Fenthion_squared)
  )
```

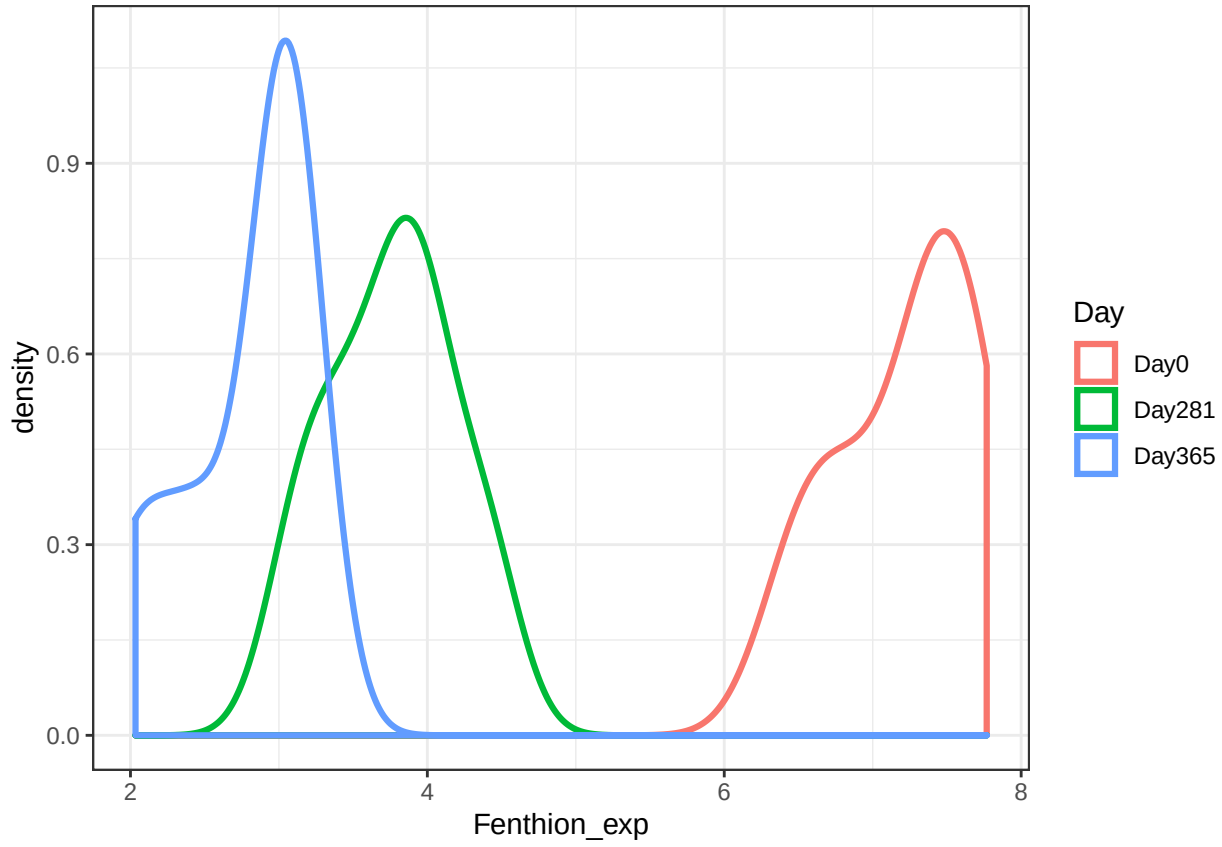
```
## # A tibble: 3 x 2
```

```
##   Day     sd_Fenthion
##   <fct>         <dbl>
## 1 Day0   0.2653640273
## 2 Day281 0.3153596132
## 3 Day365 0.3226386921
```

The standard deviations are much closer to being equal with the squared transformation. These standard deviations are close enough that this is a suitable transformation to use. For the sake of completeness, we can explore a little more to see if we can do better. Let's move another step up the ladder to an exponential transformation:

```
## Add another column: the exp transformation
olives <- olives %>%
  mutate(
    Fenthion_exp = exp(Fenthion)
  )

## plot the density curves to assess shape and spread
ggplot(data=olives, aes(x=Fenthion_exp, color=Day)) +
  geom_density(size=1.1) +
  theme_bw()
```




```
## compute the standard deviation for each group
olives %>%
  group_by(Day) %>%
```

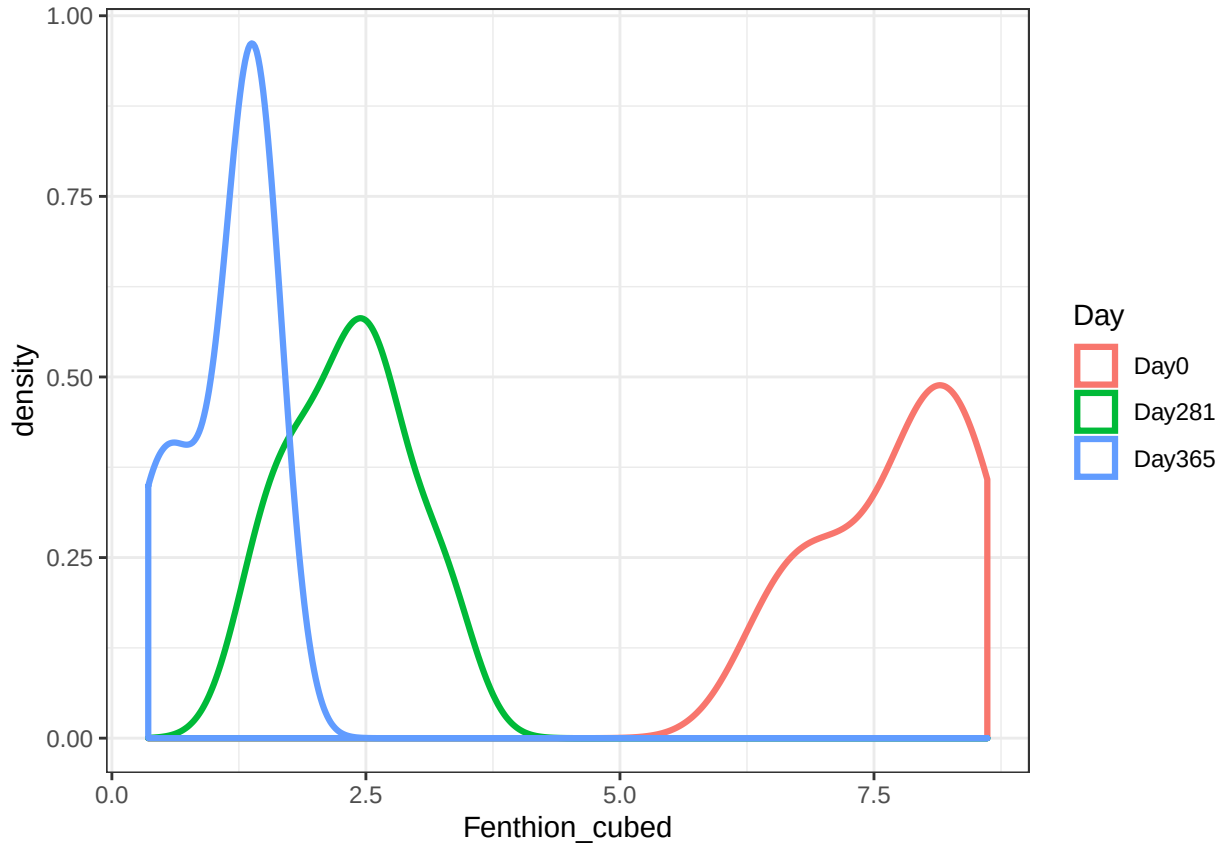
```
  summarize(
    sd_Fenthion = sd(Fenthion_exp)
  )
```

```
## # A tibble: 3 x 2
##   Day     sd_Fenthion
##   <fct>         <dbl>
## 1 Day0   0.4808057544
## 2 Day281 0.4478307349
## 3 Day365 0.4425330832
```

After an exponential transformation, the standard deviations for the three groups are even closer to being equal than they were for the squared transformation. This transformation is slightly preferable for thies reason. We will try one more transformation before settling on a transformation for the analysis in (c):

```
## Add another column: the cubed transformation
olives <- olives %>%
  mutate(
    Fenthion_cubed = Fenthion^3
  )

## plot the density curves to assess shape and spread
ggplot(data=olives, aes(x=Fenthion_cubed, color=Day)) +
  geom_density(size=1.1) +
  theme_bw()
```



```r
## compute the standard deviation for each group
olives %>%
  group_by(Day) %>%
  summarize(
    sd_Fenthion = sd(Fenthion_cubed)
  )
```

```
## # A tibble: 3 x 2
##   Day     sd_Fenthion
##   <fct>         <dbl>
## 1 Day0   0.7801013605
## 2 Day281 0.6226429449
## 3 Day365 0.4607894387
```

This transformation goes too far - see how the standard deviations are now more different? We will go with the exponential transformation for (c), although the squared transformation would also be okay.

Note that we are going to proceed with the analysis in (c), but the violation of independence could be a big problem in practice, and other methods should be explored.

**(c) For the purpose of this problem, let's assume that the conditions you checked in part (b) were fairly well satisfied (perhaps after suitable transformation). Conduct a test to find out whether there were any differences in the mean amount of fenthion at the three different times of year (if necessary, conduct a test about means on the transformed scale). Please define all parameters involved, state your hypotheses in terms of equations involving the parameters and written sentences explaining what the hypotheses mean in context, and interpret the p-value for your test in terms of strength of evidence against the null hypothesis of the test, stated in context.**

Define parameters:

- $\mu_1$ = mean of the exponentiated fenthiol levels in the population of oils at "Day 0"
- $\mu_2$ = mean of the exponentiated fenthiol levels in the population of oils at "Day 281"
- $\mu_3$ = mean of the exponentiated fenthiol levels in the population of oils at "Day 365"

State hypotheses:

- $H_0 : \mu_1 = \mu_2 = \mu_3$. The average exponentiated fenthiol levels are the same at all three days.
- At least one of $\mu_1$, $\mu_2$, $\mu_3$ is not equal to the others. The average exponentiated fenthiol levels are different at different lengths of time after the oil was collected.

Conduct an F test:

```r
model_fit <- lm(Fenthion_exp ~ Day, data=olives)
anova(model_fit)
```

```
## Analysis of Variance Table
##
## Response: Fenthion_exp
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Day        2 65.244  32.622  155.95 9.176e-11 ***
## Residuals 15  3.138   0.209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the test is $9.176 \times 10^{-11}$. The data provide extremely strong evidence against the null hypothesis of no difference between the group mean fenthion levels on the exponentiated scale. It appears that the amount of fenthion in the oil changes over time (we don't know if it increases or decreases because the test statistic for an F test is always positive).

**(d) Find three confidence intervals with a familywise confidence level of 95%: one for the difference between the mean amount of fenthion present at day 0 and the mean amount present at day 281; a second for the difference between the mean amount of fenthion present at day 0 and the mean amount present at day 365; and a third for the difference between the mean amount of fenthion present at day 281 and the mean amount present at day 365. Find the confidence intervals using the Bonferroni adjustment for the familywise confidence level. Interpret your confidence intervals in context. For which pairs of days do the data provide statistically significant evidence of a difference in means? All of your inferences can be on the transformed scale, if you selected a transformation in part (b).**

Since we are finding three confidence intervals, in the Bonferroni calculations the value of $k$ is 3. The quantile of the $t$-distribution to use for the multiplier is $1 - \frac{0.05}{2\times 3} \approx 0.99167$. Each individual confidence interval will be an approximate $(1 - 0.05/3) \times 100\% = 98.3\%$. Our sample size is 18, and there are 3 groups, so the degrees of freedom are 18-3=15. The multiplier (critical value) is:

```
qt(1-0.05/(2*3), df=18-3)
```

```
## [1] 2.693739
```

To find the intervals, we can use fit.contrast:

```
## Day 0 versus Day 281
fit.contrast(model_fit, "Day", c(1,-1,0), conf.int=1-0.05/3)
```

```
##                  Estimate Std. Error  t value     Pr(>|t|) lower CI
## Day c=( 1 -1 0 ) 3.460563  0.2640627 13.10508 1.285247e-09 2.749247
##                  upper CI
## Day c=( 1 -1 0 )  4.17188
## attr(,"class")
## [1] "fit_contrast"
```

```
## Day 0 versus Day 365
fit.contrast(model_fit, "Day", c(1,0,-1), conf.int=1-0.05/3)
```

```
##                  Estimate Std. Error  t value     Pr(>|t|) lower CI
## Day c=( 1 0 -1 ) 4.437542  0.2640627 16.80488 3.864877e-11 3.726226
##                  upper CI
## Day c=( 1 0 -1 ) 5.148858
## attr(,"class")
## [1] "fit_contrast"
```

```
## Day 281 versus Day 365
fit.contrast(model_fit, "Day", c(0,1,-1), conf.int=1-0.05/3)
```

```
##                   Estimate Std. Error  t value    Pr(>|t|)  lower CI
## Day c=( 0 1 -1 ) 0.9769785  0.2640627 3.699797 0.002139976 0.2656624
```

```
##                   upper CI
## Day c=( 0 1 -1 ) 1.688295
## attr(,"class")
## [1] "fit_contrast"
```

We are 95% confident that the difference in means of the exponentiated fenthiol levels in the population of oils at Day 0 and at Day 281 is between 2.749 and 4.172, the difference in means of the exponentiated fenthiol levels in the population of oils at Day 0 and at Day 365 is between 3.726 and 5.149, and the difference in means of the exponentiated fenthiol levels in the population of oils at Day 281 and at Day 365 is between 0.266 and 1.688. For 95% of samples, a set of three confidence intervals calculated in this way would simultaneously contain the respective differences in means they are estimating.

Note, you can do this interpretation on the transformed scale if you wish. As an example, the bounds on the interval for Day 0 and Day 281 would be $\log(2.749) = 1.011$ and $\log(4.172) = 1.428$; the mean fenthiol levels in the population of oils for Day 0 would be between 1.011 and 1.428 times the mean fenthiol levels in the population of oils for Day 281.