Assignment 2 by Senthooran Yogeswaran

```
library(dplyr)
library(ggplot2)
library(lme4)
library(MASS)
library(tinytex)
library(rms)
library(tidyverse)
library(aod)
```

a) Exploring data.

```
cd4 <- read.table("CD4.txt", header = TRUE)
head(cd4)
```

```
##   newpid CD4PCT   visage baseage      time treatmnt
## 1      1     18 3.910000   3.910 0.0000000        1
## 2      1     37 4.468333   3.910 0.5583333        1
## 3      1     13 4.698333   3.910 0.7883333        1
## 4      1     13 5.330833   3.910 1.4208333        1
## 5      1     12 5.848333   3.910 1.9383333        1
## 6      2      1 3.565000   3.565 0.0000000        2
```

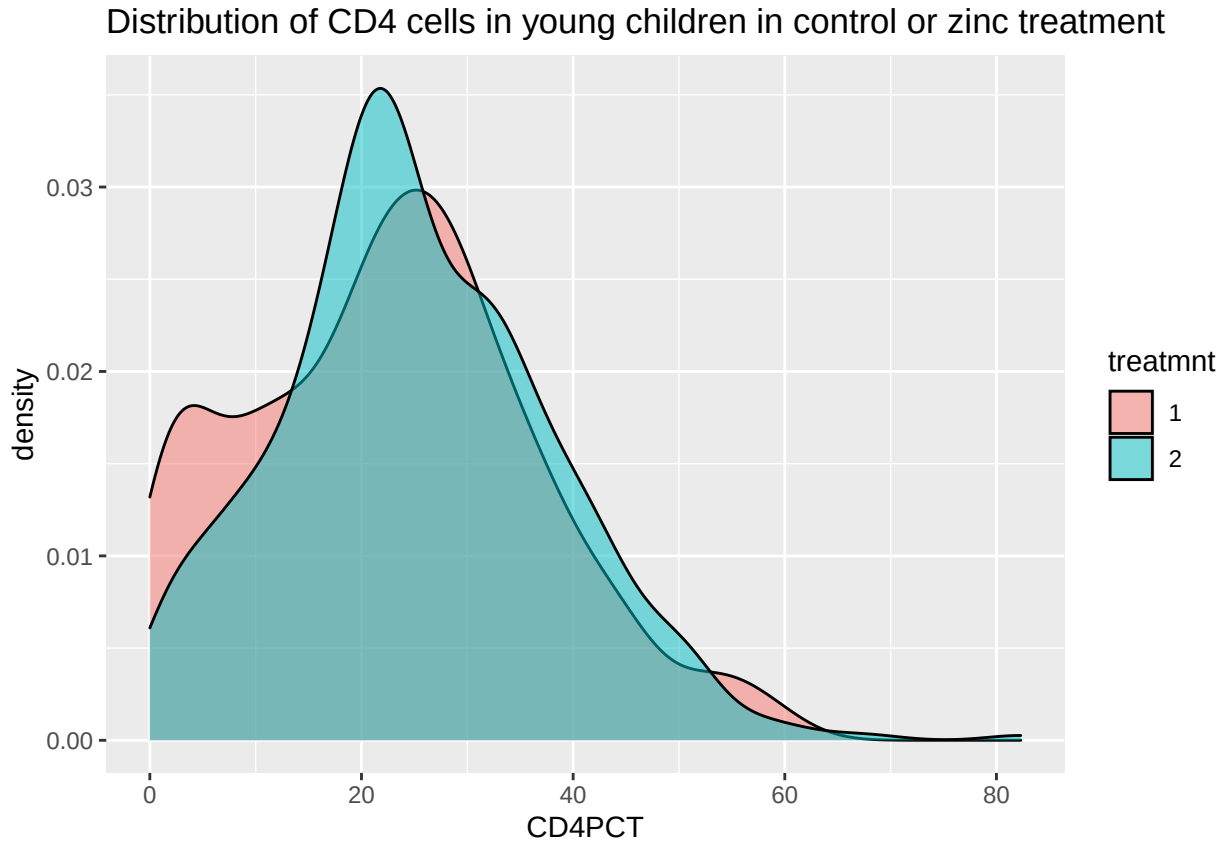
After viewing we convert treatmnt from integer to a factor

```
cd4$treatmnt <- as.factor(cd4$treatmnt)
unique(cd4$treatmnt)
```

```
## [1] 1 2
## Levels: 1 2
```

Distribution of CD4 cells to figure out how to transform data.

```
cd4 %>%
  ggplot(aes(x= CD4PCT, fill = treatmnt))+
  geom_density(alpha=0.5)+
  labs(title = "Distribution of CD4 cells in young children in control or zinc treatment")
```

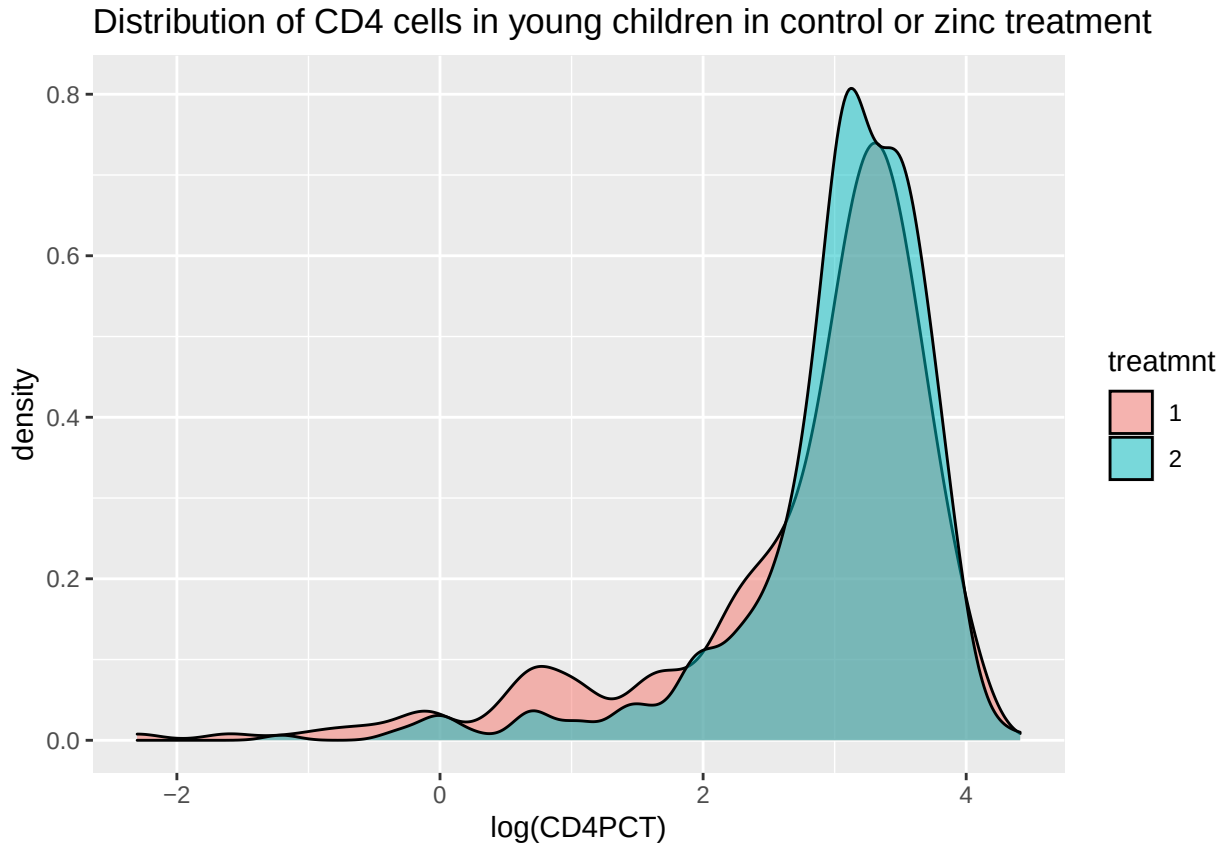


We notice the y-axis is interfering with the distribution data. A log or square root transformation will help pull the data away.

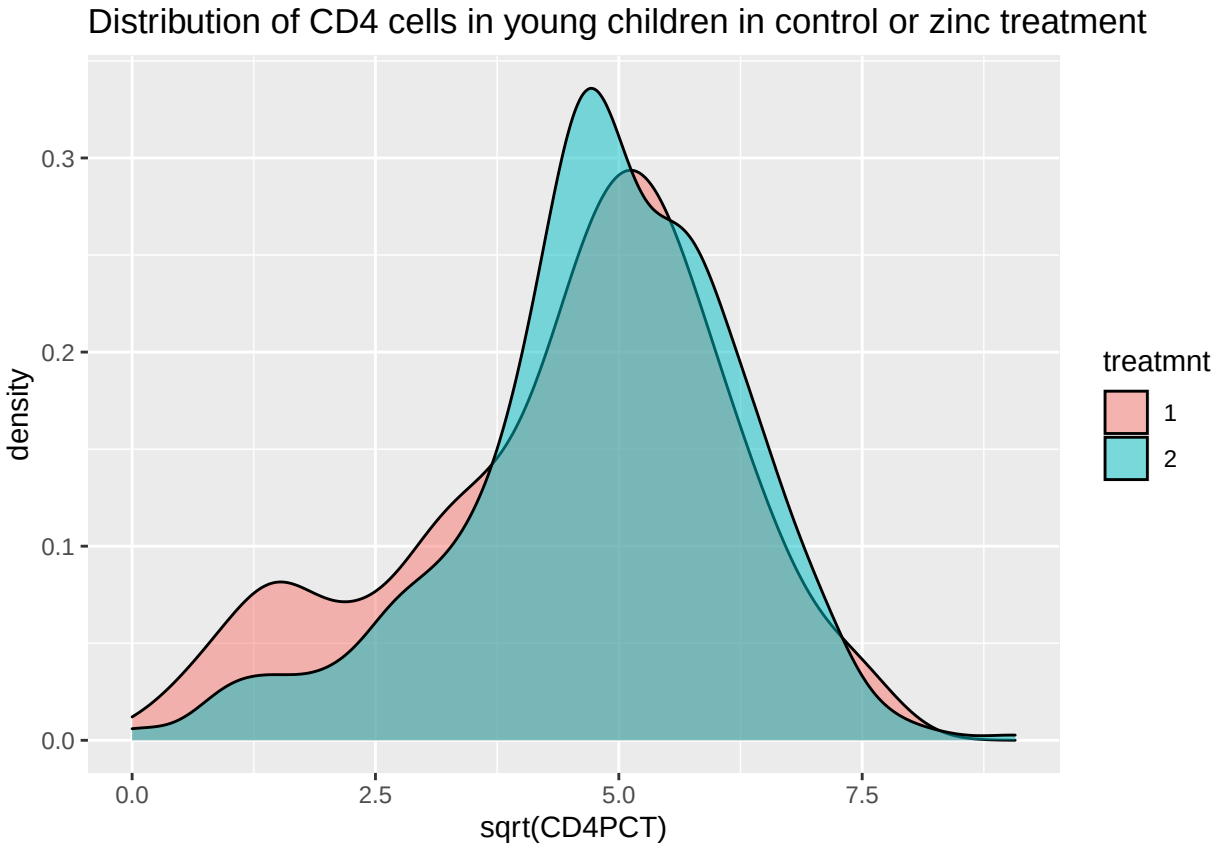
The log transformation adds skewness which may be problematic. Fortunately, the square root transformation makes our data look more symmetrical and normal- like. Therefore, we will select this transformation.

```
cd4 %>%
  ggplot(aes(x= log(CD4PCT), fill = treatmnt))+
  geom_density(alpha=0.5)+
  labs(title = "Distribution of CD4 cells in young children in control or zinc treatment")
```

```
## Warning: Removed 4 rows containing non-finite values (stat_density).
```

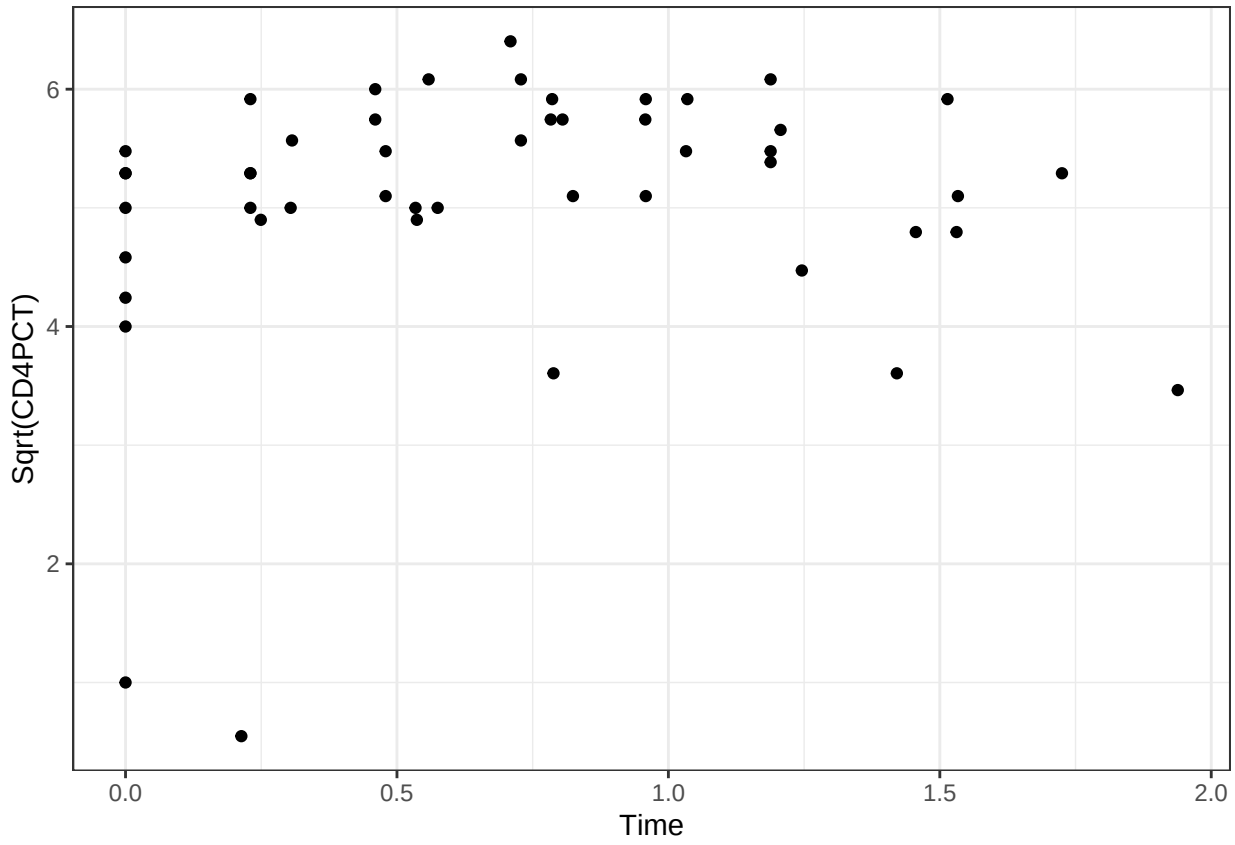



```
cd4 %>%
  ggplot(aes(x= sqrt(CD4PCT), fill = treatmnt))+
  geom_density(alpha=0.5)+
  labs(title = "Distribution of CD4 cells in young children in control or zinc treatment")
```



B) Selecting 10 children and showing the transformed CD4 percents at each time point. From the data rows 1 to 49 have 10 subjects with repeated measurements over time.

```
cd4_subset <- cd4[1:49, ]
cd4_subset %>%
  ggplot(aes(x= time, y = sqrt(CD4PCT)))+
  geom_point()+
  theme_bw()+
  labs(x="Time", y = "Sqrt(CD4PCT)")
```



c)

We write a model with CD4 percents as a function of time with intercepts varying by child.

```
lmer1 <- lmer(data=cd4, formula =  sqrt(CD4PCT) ~
                time + (1|newpid) )
```

α_j $N(4.76, 1.40^2)$

We see our estimated model parameters for μ_a

to be about 4.76 and the β

to be about -0.37

So as time increases that the percentage of CD4 decreases by 0.37% for each child.

```
##               Estimate Std. Error   t value
## (Intercept)  4.7634086 0.09647959 49.372189
## time        -0.3660932 0.05398921 -6.780858
```

```
summary(lmer1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: sqrt(CD4PCT) ~ time + (1 | newpid)
##    Data: cd4
```

```
##
## REML criterion at convergence: 3140.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7379 -0.4379  0.0024  0.4324  5.0017
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  newpid   (Intercept) 1.9569   1.3989
##  Residual             0.5968   0.7725
## Number of obs: 1072, groups:  newpid, 250
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  4.76341    0.09648  49.372
## time        -0.36609    0.05399  -6.781
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.278
```

d)

We notice as time increases that the percentage of CD4 decreases changes from 0.366% to 0.362% which means the treatment or the base age had a very small correlation with the time students were measured The mu hat alpha intercept has grown from 4.76 to 5.08. Compared to the radon example we were only concerned about the basement effect and had a fixed slope model with partial, and complete pooling. Here, we have more fixed effects. There are three to be exact.

```
lmer2 <- lmer(data=cd4, formula =  sqrt(CD4PCT) ~
               time + treatmnt + baseage + (1|newpid) )
summary(lmer2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: sqrt(CD4PCT) ~ time + treatmnt + baseage + (1 | newpid)
##    Data: cd4
##
## REML criterion at convergence: 3137.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7490 -0.4392  0.0097  0.4282  5.0141
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  newpid   (Intercept) 1.8897   1.3747
##  Residual             0.5969   0.7726
## Number of obs: 1072, groups:  newpid, 250
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  5.08614    0.18793  27.064
```

```
## time        -0.36216    0.05399  -6.708
## treatmnt2    0.18008    0.18262   0.986
## baseage     -0.11945    0.04000  -2.986
##
## Correlation of Fixed Effects:
##           (Intr) time   trtmn2
## time      -0.135
## treatmnt2 -0.462  0.010
## baseage   -0.727 -0.017 -0.003
```

e)

Keeping our transformation in mind, treatment seems to be effective. In our previous model it showed that children that were treated have about 0.18% more CD4 than those who didn't. Children who were older at the base age have 0.11% less CD4 than children who were younger at the base age.

f)

We want to simulate data from a child with newpid #9 so we take the coefficient [9,1]. We fix beta at 1.4 since our time is asked at 1.4 years. Our prediction interval from simulating from this data is: (6.18, 9.21) This means from simulating from our data their CD4 percentage at 1.4 years should be between 6.18 and 9.21.

```
lmer2 <- lmer(data=cd4, formula =  sqrt(CD4PCT) ~
                time + treatmnt + baseage + (1|newpid) )
alpha1 <- coef(lmer2)$newpid[9,1]
#fixing time at 1.4 years
beta <- 1.4
sigma <- summary(lmer2)$sigma

cd4_sims <- rnorm(10000, mean = alpha1 + beta, sigma)

cd4_prediction_interval <- quantile(cd4_sims, probs = c(0.025,0.5,0.975))
cd4_prediction_interval
```

```
##     2.5%      50%    97.5%
## 6.173406 7.719741 9.243352
```

g)

The new prediction interval for predicting CD4 percentage after 1 year is: (2.12, 10.05). This is larger as we don't really have any new information pertaining the children. Fortunately, this interval is only slightly wider. If you had used the model from C I would anticipate the interval to be even wider as we removed information pertaining the children.

```
#beta here is fixed at 1 year
new_beta = 1
mu_alpha1 <- summary(lmer2)$coefficients[1,1]
sigma_alpha1 <- summary(lmer2)$varcor[1]$newpid[1]
alpha_new1 <- rnorm(10000, mu_alpha1, sigma_alpha1 )

new_cd4_sims <- rnorm(10000, alpha_new1 + new_beta, sigma)
cd4_prediction_interval_new <- quantile(new_cd4_sims, probs = c(0.025,0.5,0.975))
cd4_prediction_interval_new
```

```
##      2.5%       50%     97.5%
##  2.076568  6.116605 10.153723
```

~ **Question 2** ~ Introduction ~

It's always been important to excel at your studies and education as it leads to a brighter future. Typically, being labelled as smart brings pride to your family as well as social recognition. There's always been a stereotype that females are more studious than males, while males might be recognized for their athletic feats in sports. My research question will look at a scholarship data set and verify if females are more likely to win academic award than males. Additionally, I will be looking if socioeconomic status

~ Data Analysis ~

```
scholarships <- read.table("scholarships.txt", header = TRUE)
head(scholarships)
```

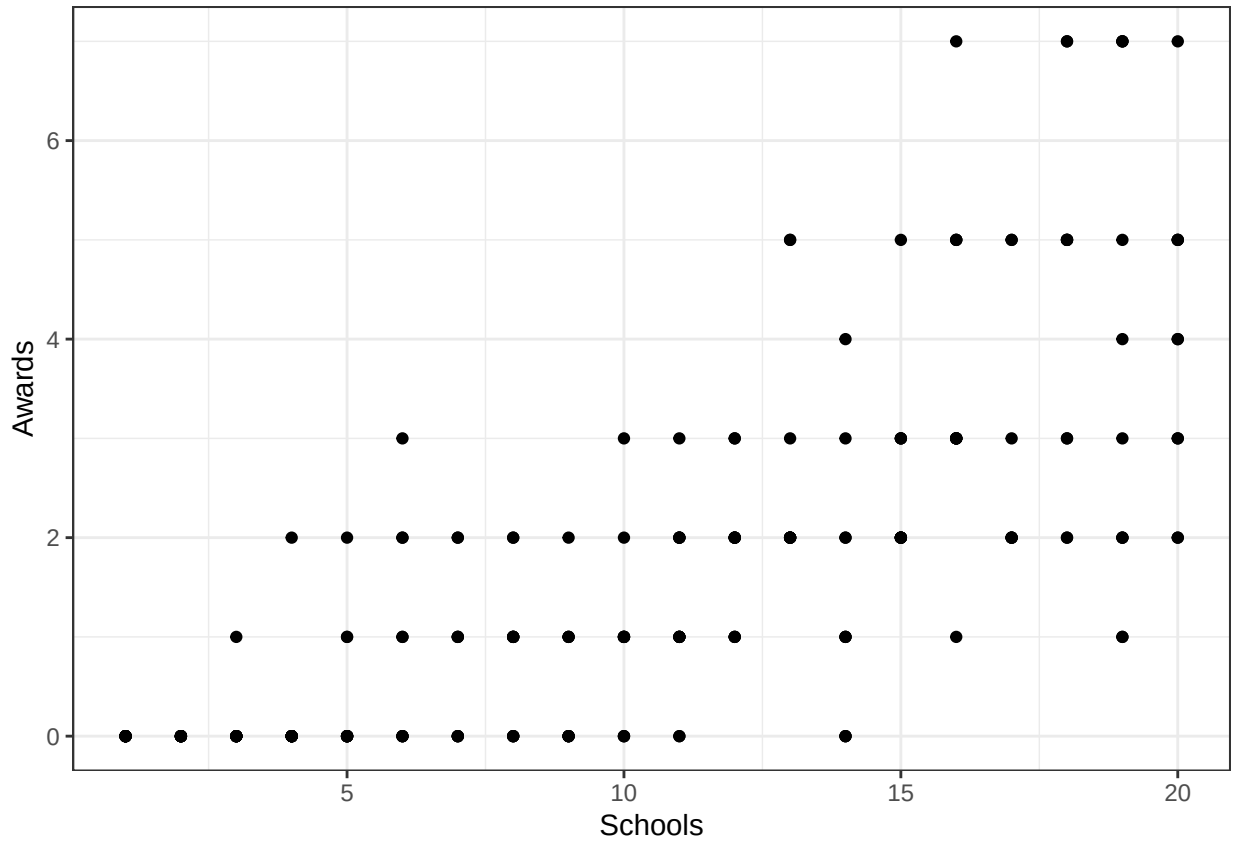
```
##   female ses schtyp prog awards cid
## 1      1   1      1    3      0   1
## 2      0   2      1    1      0   1
## 3      0   3      1    3      0   1
## 4      0   1      1    3      0   1
## 5      0   2      1    3      0   1
## 6      1   3      1    1      0   1
```

Let's clean up some of the data. Specifically, changing the female, SES, schtype and progtype in factors since they are marked as the wrong variables.

```
scholarships$female <- as.factor(scholarships$female)
scholarships$ses <- as.factor(scholarships$ses)
scholarships$schtyp <- as.factor(scholarships$schtyp)
scholarships$prog <- as.factor(scholarships$prog)
```

We want to explore the data to understand what type of model to fit, therefore plotting the data should help us. We understand the variables female, awards, and and ses are the utmost importance and will most likely be used in future models.

```
scholarships %>%
  ggplot(aes(x=cid, y=awards))+
  geom_point()+
  theme_bw()+
  labs(x="Schools", y = "Awards")
```

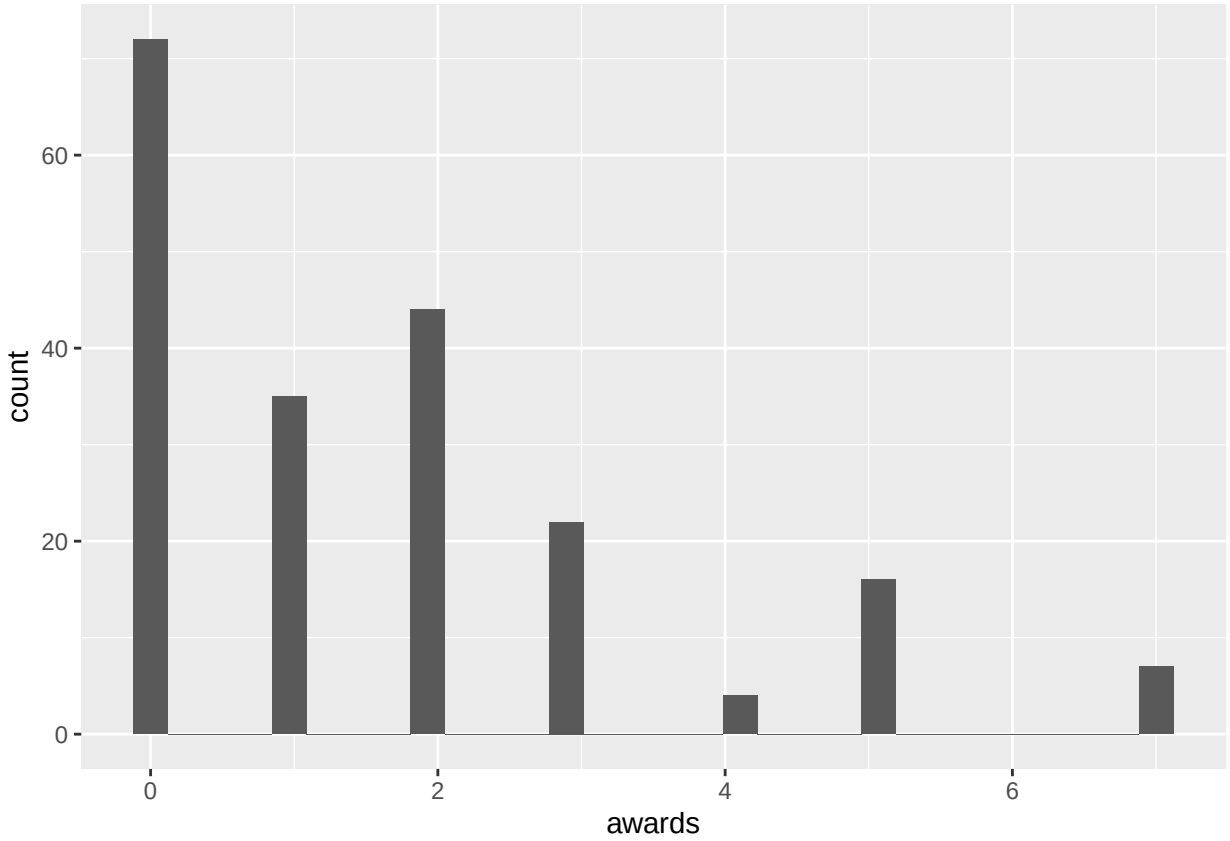


We created a plot to understand the relationship of schools and the amount of awards. We notice some schools have a low amount of awards. Due to this, its important to use random effects, particularly fixing the effect over a normal distribution.

Also, remember our outcome variable awards is an unrestricted count. we could use a normal model but a LMM seems more reasonable.

```
scholarships %>% ggplot(aes(x=awards))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We notice the school id's vary from 1 to 20 but many of them have repeated measurements. This is exactly why the current model works alongside the low amount of awards for some of them described. This implies, that our intercepts vary by school id and this is our random intercept.

```
scholar_lmer <- lmer(data=scholarships, formula = awards ~ 1 + (1|cid) + female + ses + prog)
summary(scholar_lmer)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: awards ~ 1 + (1 | cid) + female + ses + prog
##    Data: scholarships
##
## REML criterion at convergence: 675
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6236 -0.5104 -0.1682  0.3263  3.1202
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  cid      (Intercept) 1.836    1.355
##  Residual             1.297    1.139
## Number of obs: 200, groups:  cid, 20
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.19667    0.39722   3.013
```

```
## female1      0.60254    0.17288   3.485
## ses2         0.03015    0.22526   0.134
## ses3         0.26519    0.25076   1.058
## prog2        0.12182    0.22351   0.545
## prog3       -0.01058    0.25112  -0.042
##
## Correlation of Fixed Effects:
##         (Intr) femal1 ses2   ses3   prog2
## female1 -0.305
## ses2    -0.380  0.165
## ses3    -0.331  0.134  0.609
## prog2   -0.333 -0.011 -0.070 -0.140
## prog3   -0.269 -0.002 -0.125 -0.054  0.528
```

Our model used awards as the outcome with school id as our random intercept, and whether they were a female or not, their socioeconomic status and progtype as indicators.

$yi\ N(1.20, 1.14^2)$

$\mu_i = \alpha_j[i] + \beta_k[i] + \beta_s[i] + \beta_n[i]$ where alpha is the school effect and beta_k is the effect of being a female or not. Beta_s is socioeconomic status and beta_n is the progtype

Our α_j has a distribution of $\alpha_j\ N(1.20, 1.14^2)$

~ Discussion ~

Based off our summary output we can make an inference on the data analyzed. Firstly, we notice females have a 60% chance of winning an award compared to males. Those with a higher socioeconomic status have about 26% of winning an award compared to a lower status. Those who had a progtype of academic had a 12% chance.

~ Results and Conclusion ~

The data analysis conducted shows that females are more likely to win an academic award compared to males. Socioeconomic status is associated with winning an award which may make sense since coming from a family with a higher income and stability could make studying easier and ensuring you have the optimal nutrition to enhance your brain's function.