

# STAT 224 Autumn 2022 HW1

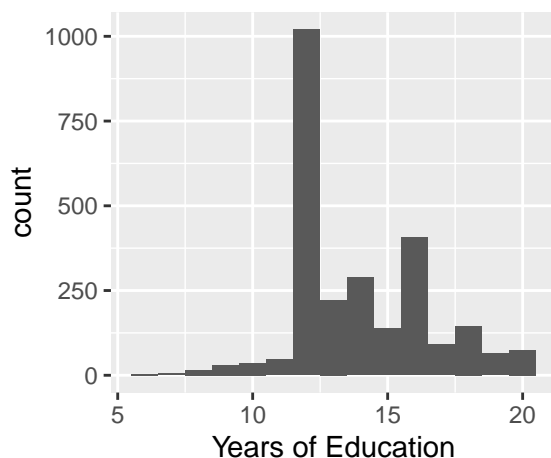
Matthew Zhao

```
NLSY = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/NLSY.txt", header=T)
```

```
library(mosaic)
```

## Q1

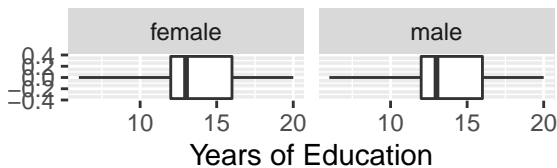
```
library(ggplot2)
ggplot(data = NLSY, aes(x=Edu2006)) +
  geom_histogram(binwidth = 1) +
  xlab('Years of Education')
```



The location of the modes is due to how the education system in the US works, specifically the divisions between 1-12 (elementary, middle, and high school), 13-16 (college), and beyond. Since education before college is provided to all Americans, vast majority can at least complete up to 12 years, hence the large peak at 12. Then, some portion can afford college (but not all complete it), explaining the peak at 16. Finally, there are also many 2-3 year programs e.g. MBA, MA/MS, JD, etc explaining the small peak after. The ideal bin size here is around 1.

## Q2

```
ggplot(data = NLSY, aes(x=Edu2006)) +  
  geom_boxplot() +  
  facet_wrap(~Gender) +  
  xlab('Years of Education')
```



```
summary(NLSY$Edu2006)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
##      6.00  12.00   13.00   13.89  16.00   20.00
```

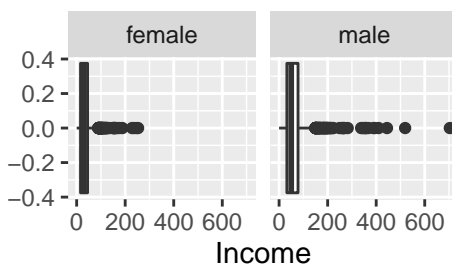
```
print(paste0('Standard Deviation: ', round(sd(NLSY$Edu2006),3)))
```

```
## [1] "Standard Deviation: 2.504"
```

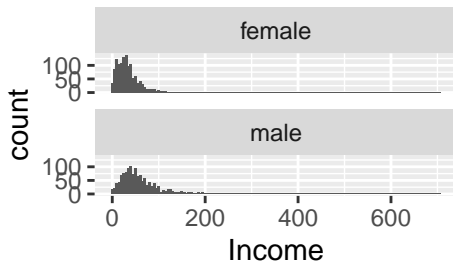
It appears that men and women have similar education levels. The boxplots are identical because the distribution of education is the same for both genders.

## Q3

```
ggplot(data = NLSY, aes(x=Income2005)) +  
  geom_boxplot() +  
  facet_wrap(~Gender) +  
  xlab('Income')
```



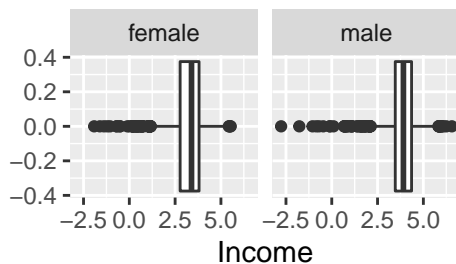
```
ggplot(data = NLSY, aes(x=Income2005)) +
  geom_histogram(binwidth = 5) +
  facet_wrap(~Gender, ncol=1) +
  xlab('Income')
```



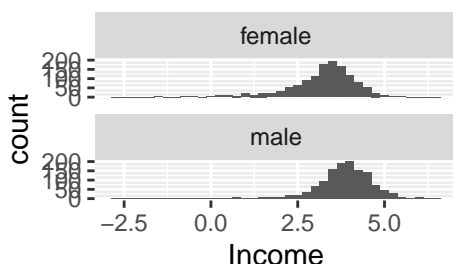
Both histograms are right-skewed and unimodal but the distribution of male income has greater variation (more spread out distribution) and a higher mean and median income as a result.

## Q4

```
ggplot(data = NLSY, aes(x=log(Income2005))) +
  geom_boxplot() +
  facet_wrap(~Gender) +
  xlab('Income')
```



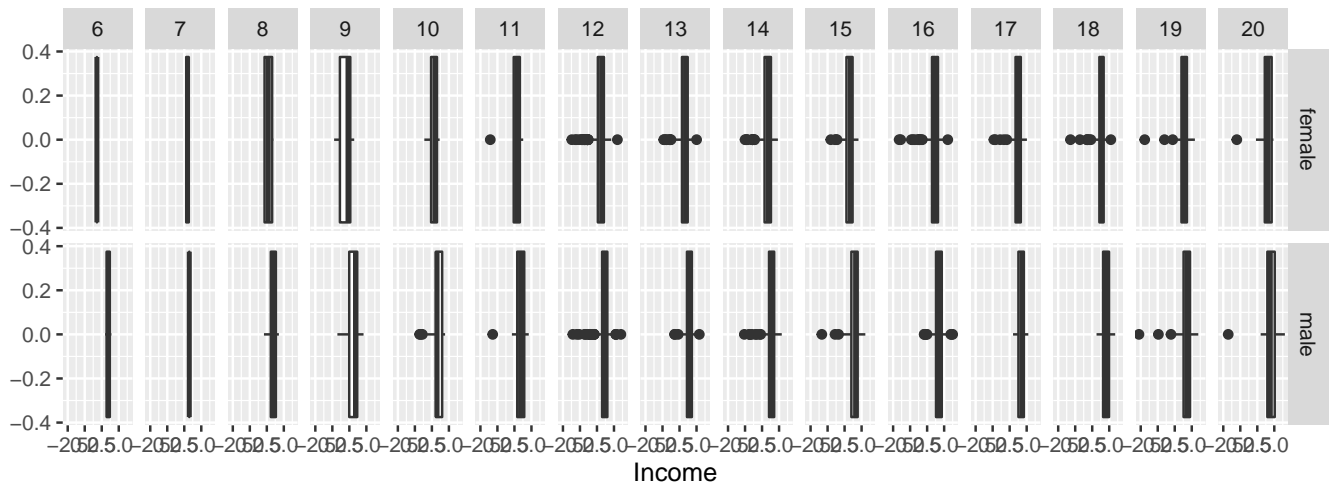
```
ggplot(data = NLSY, aes(x=log(Income2005))) +
  geom_histogram(binwidth = 0.25) +
  facet_wrap(~Gender, ncol=1) +
  xlab('Income')
```



After transformation, the distributions become roughly normal, with a slight left skew.

Q5

```
ggplot(data = NLSY, aes(x=log(Income2005))) +  
  geom_boxplot() +  
  facet_grid(Gender~Edu2006,scale='free_y') +  
  xlab('Income')
```



Q6

Q7

Q8