Code

HW5

Problem 5A

Suppose that $X_1, X_2, \ldots, X_{25}$ are i.i.d $N(\mu, \sigma^2)$ where $\mu = 0$ and $\sigma = 10$. Plot the sampling distributions of $\bar{X}$ and $\hat{\sigma^2}$.

Since the normal random variables will be normal, $\bar{X}$

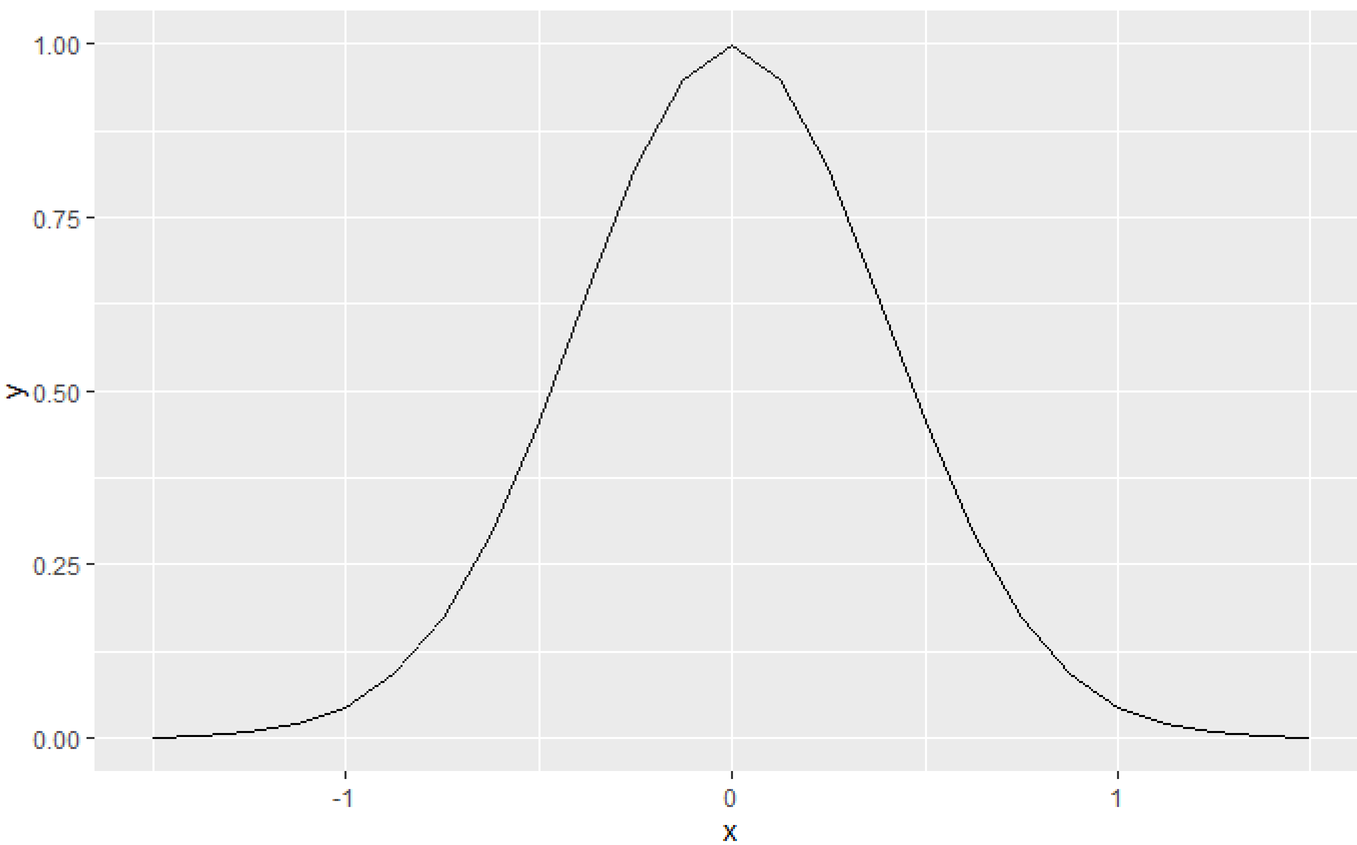
$\mathbb{E}[\bar{X}] = \mathbb{E}[\frac{1}{25}\sum_{i=1}^{25}(X_i)] = [\frac{1}{25}\sum_{i=1}^{25}\mathbb{E}[X_i] = 0$

$Var(\bar{X}) = Var(\frac{1}{25}\sum_{i=1}^{25}X_i) = [\frac{1}{25^2}\sum_{i=1}^{25}Var(X_i)] = \frac{10^2}{25^2} = 0.16$

Therefore, since $\bar{X} \sim N(0, 0.4)$, the plot is given by:

Hide

```
library(ggplot2)
p1 <- ggplot(data.frame(x = c(-1.5,1.5)),aes(x))+
  stat_function(fun=dnorm, n=25, args=list(mean=0, sd=0.4))
p1
```



$\hat{\sigma}^2$

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}s^2$

$\Downarrow$

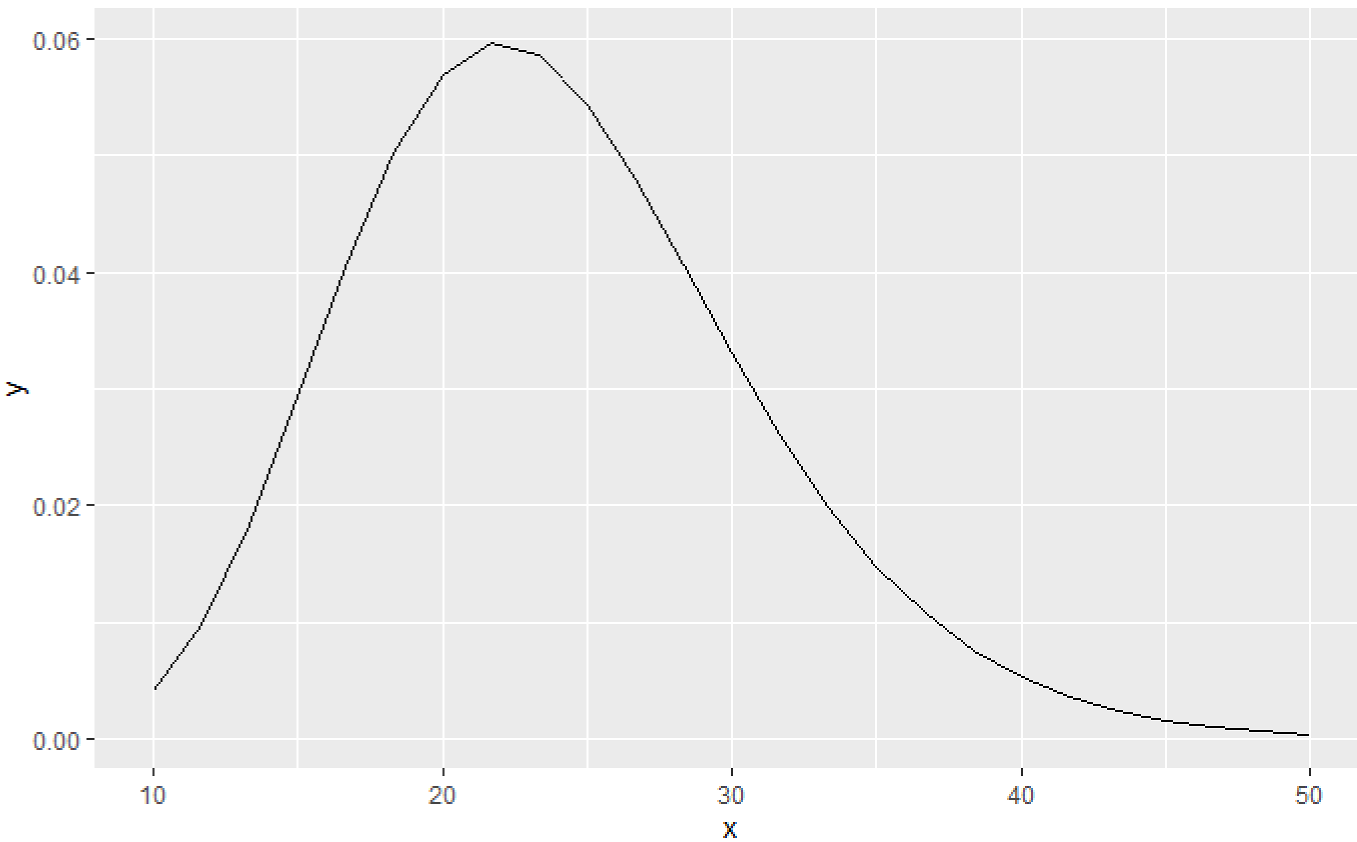
$s^2 = \frac{n}{n-1}\hat{\sigma}^2$

$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \Rightarrow \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$

Therefore, since $\hat{\sigma}^2 \sim \chi^2_{n-1}$, the plot is given by:

Hide

```
library(ggplot2)
p2 <- ggplot(data.frame(x = c(10,50)),aes(x))+
  stat_function(fun=dchisq, n=25, args=list(df=24))
p2
```



Problem 5G

Hide

5B) 8.26

estimate by MLE

$$P(X=k) = \frac{\binom{100}{k}\binom{N-100}{50-k}}{\binom{N}{50}}$$

$$lk(N) = \frac{\binom{100}{20}\binom{N-100}{30}}{\binom{N}{50}}$$

$$\frac{l(N)}{l(N-1)} = \frac{(N-100)(N-50)}{N(N-130)} \geq 1$$

$$\Rightarrow (N-100)(N-50) \geq N(N-130)$$

$$\Rightarrow N \leq 250$$

$$\boxed{\hat{N}_{MLE} = 250}$$

5c) 8.30

a) $lik(\lambda) = (\lambda e^{-5\lambda})(\lambda e^{-3\lambda})(e^{-10\lambda})$

$$\boxed{= \lambda^2 e^{-18\lambda}}$$

b) $\ell(\lambda) = \log(lik(\lambda)) = 2\log\lambda - 18\lambda$

$$\ell'(\lambda) = \frac{2}{\lambda} - 18 = 0$$

$$\boxed{\frac{1}{9} = \hat{\lambda}_{ML}}$$

5D) 8.32

a) $\mu = \bar{x} \leftarrow \frac{(5.3299 + 4.2537 + \ldots +)}{n}$

$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$

$\boxed{\mu \approx 3.61 \quad \sigma^2 \approx 3.20}$

b) 90% CI

$\hat{\mu} \pm \frac{S}{\sqrt{n}} t_{n-1}(0.05) \doteq \boxed{[2.8, 4.4]}$

$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2} = 1.8488$

$\left[\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(0.5 \pm 0.45)} \right] = \boxed{[2.05, 7.06]}$

c) 90% CI = sqrt(σ^2)

$[\sqrt{2.05}, \sqrt{7.06}] = \boxed{[1.43, 2.66]}$

d) $\frac{1}{2}$ len (CI for μ), would need to 4x it because its a sqrt

5E) 8.48

δ-method

estimate λ

$$\hat{\lambda} = g(Y) = -\log\left(\frac{Y}{n}\right)$$

$$Y \sim Bin(n, e^{-\lambda}), \; g(x) = -\ln\left(\frac{x}{n}\right)$$

$$\mathbb{E}[\hat{\lambda}] \approx g(\mathbb{E}[Y]) + \frac{1}{2} Var(Y) * g''(\mathbb{E}(Y))$$

$$Var(Y) \approx Var(Y) \cdot g''(\mathbb{E}[Y])^2$$

$$\mathbb{E}(Y) = ne^{-\lambda}, \; Var(Y) = ne^{-\lambda}(1-e^{-\lambda})$$

$$g''(x) = \frac{1}{x^2}$$

$$\mathbb{E}(\hat{\lambda}) \approx \lambda + \frac{1}{2} \frac{1-e^{-\lambda}}{ne^{-\lambda}}$$

$$Var(\hat{\lambda}) = \frac{1-e^{-\lambda}}{ne^{-\lambda}}$$

bias of estimate

$$\mathbb{E}(\hat{\lambda}) - \mathbb{E}(\lambda) = \frac{1}{2} \frac{1-e^{-\lambda}}{ne^{-\lambda}}$$

$\cancel{\lambda - \lambda}$

$$\text{efficiency}(\hat{\lambda}_{ML}, \hat{\lambda}) = \frac{\frac{\lambda}{n}}{\frac{1-e^{-\lambda}}{ne^{-\lambda}}} = \frac{\lambda e^{-\lambda}}{1-e^{-\lambda}}$$

λ	1	2	5
efficiency	0.58	0.31	0.03

the $\hat{\lambda}_{ML}$ has a lower variance than $\hat{\lambda}$, meaning that it is a better estimator

5f) Bias - Variance

$$MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2 \quad , \quad \text{prove } \checkmark$$

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta}))^2$$

$$MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2$$

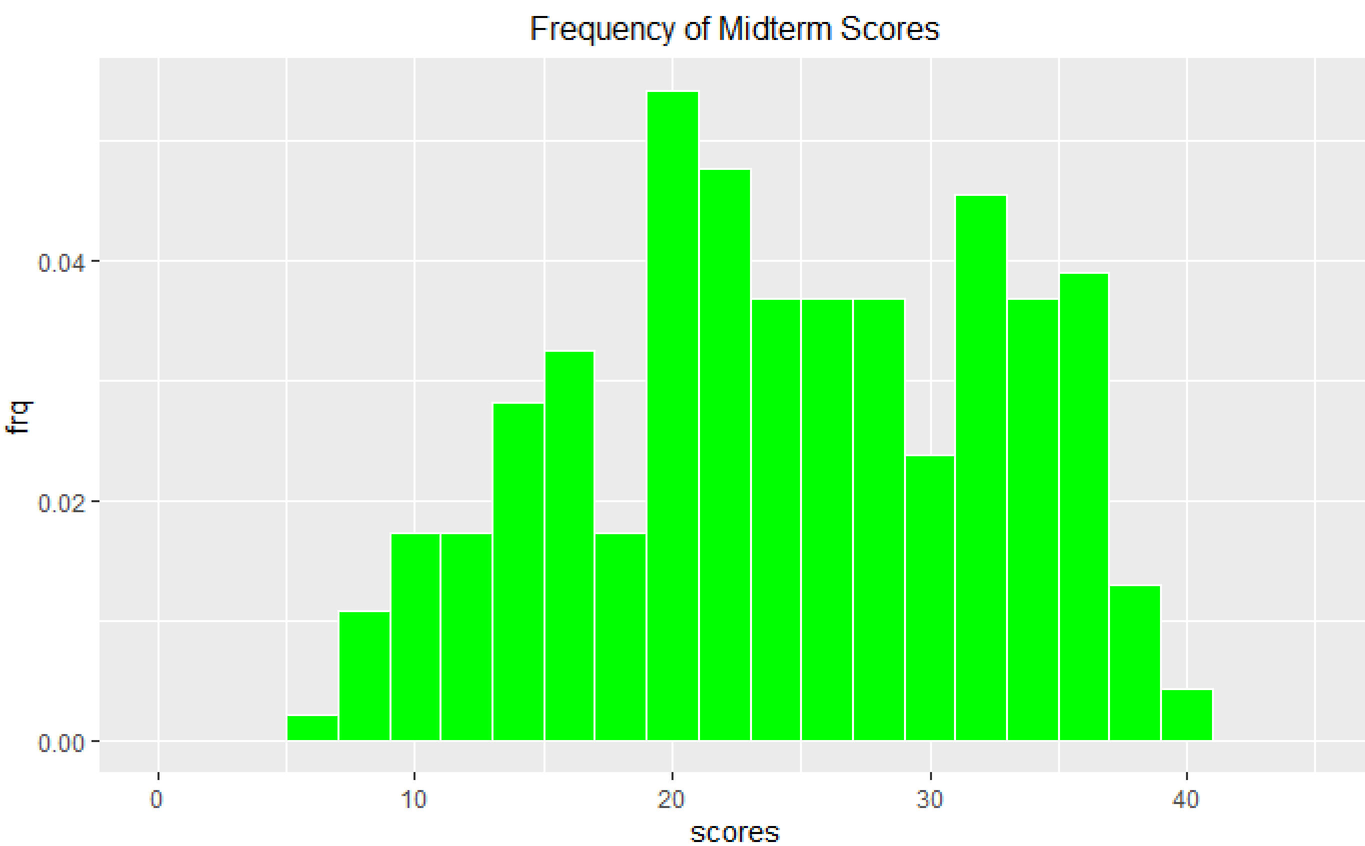
dist

$$= \mathbb{E}\left[\left([\hat{\theta} - \mathbb{E}(\hat{\theta})] + \mathbb{E}(\hat{\theta}) - \theta\right)\right)^2\right]$$

$$= \underbrace{E\left[(\hat{\theta} - E(\hat{\theta}))^2\right]} + \mathbb{E}[E(\hat{\theta}) - \theta)^2] + \cancel{2E\left((\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta} - \theta)\right)}$$

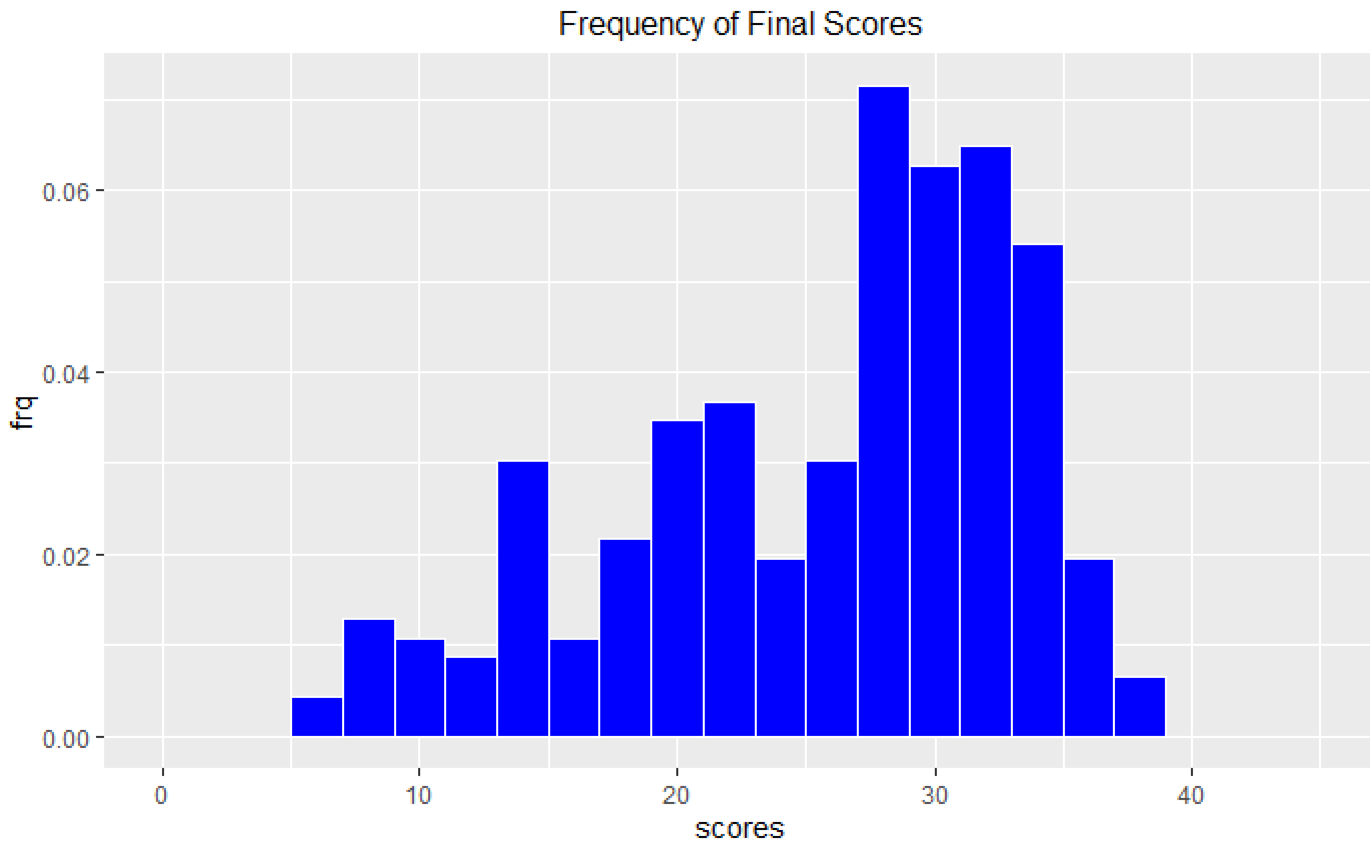
$$\boxed{= var(\theta) + \underbrace{\left[E(\hat{\theta}) - \theta\right]^2}_{[Bias(\hat{\theta})]^2}}$$

```
scoreraw <- read.delim("/Users/sethc/Documents/Berkeley Fall2023/STAT 135/data.scores.txt",sep
='',header=TRUE)
score <- scoreraw[scoreraw$f >0& scoreraw$m >0,]
score$m = score$m *2
score <- score[, c("m","f")]

#mid
ggplot(score, aes(x=m))+ geom_histogram(aes(y=..density..), binwidth=2, colour="white", fill ="g
reen")+
  ggtitle("Frequency of Midterm Scores")+
  theme(plot.title=element_text(size=12, hjust=0.5))+
  xlab("scores")+
  ylab("frq")+
  xlim(c(0,45))
```



Hide

```
#fin
ggplot(score, aes(x=f))+ geom_histogram(aes(y=..density..), binwidth=2, colour="white", fill ="b
lue")+
  ggtitle("Frequency of Final Scores")+
  theme(plot.title=element_text(size=12, hjust=0.5))+
  xlab("scores")+
  ylab("frq")+
  xlim(c(0,45))
```



Hide

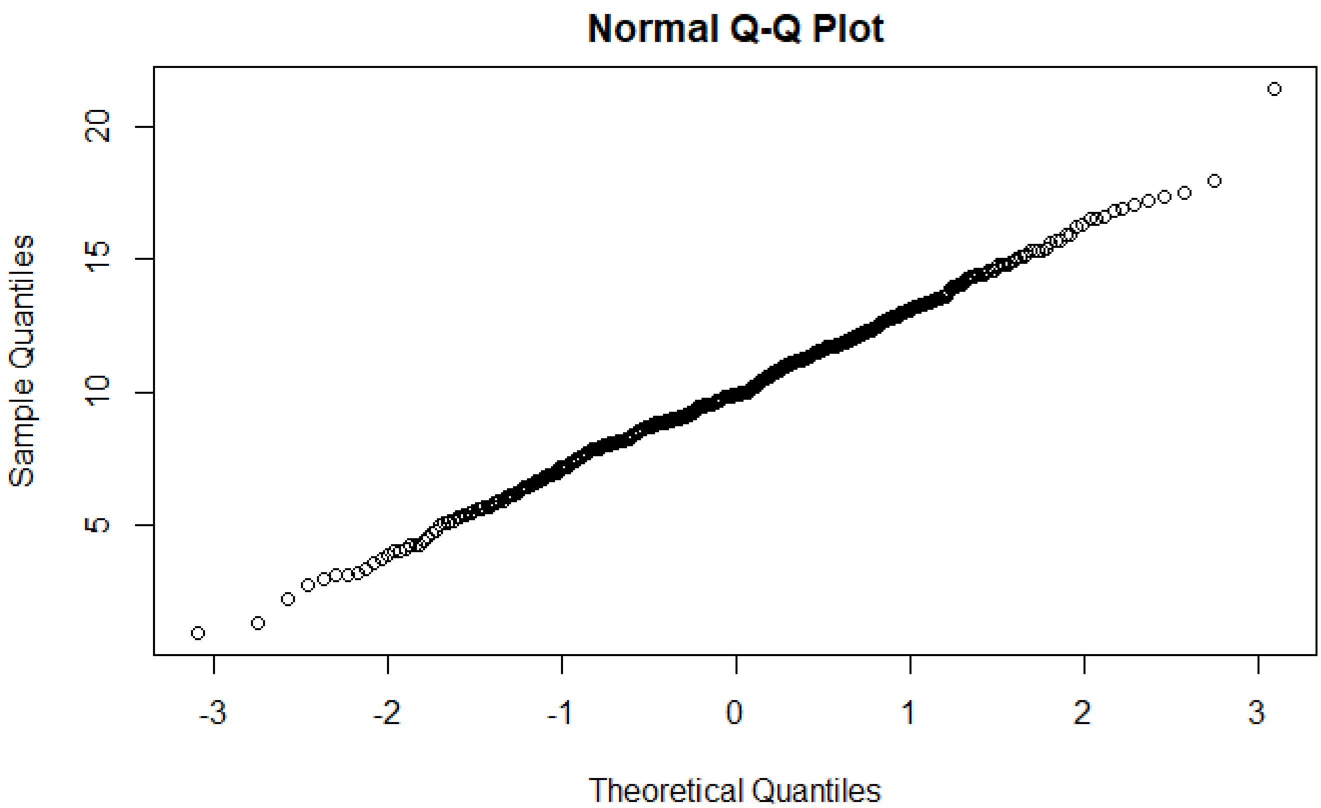
```
#box
# for some reaon it is not letting me use tidy
#library(tidyr)
#score_box <- gather(score,key ="exam",value ="scores", m, f)
#ggplot(score_long)+
  #geom_boxplot(aes(x=exam,y=scores))+
  #ggtitle("Midterm Scores vs Final Scores Distribution")+
  #theme(plot.title=element_text(size=12, hjust =0.5))
```

Observations from box plots: The midterm represents a roughly normal distribution while the final scores are signficantly right-skewed. Because of this we are using the median instead of the mean to describe the data. The median of the midterm is around ~25 but the final median is around 30, which is also seen on the histogram. The spread (percentiles) of the midterm are larger than the ones on the final.

Problem 5H

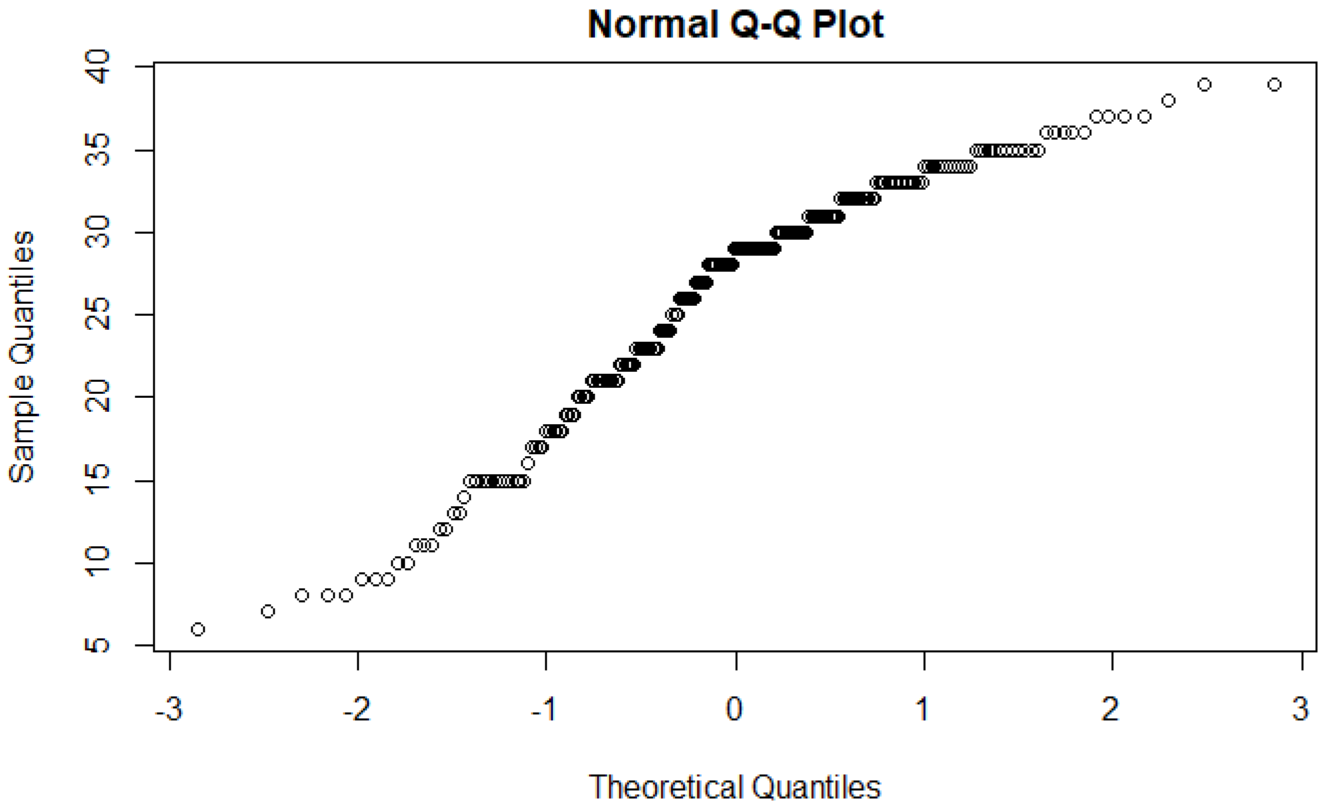
Hide

```
set.seed(1)

# a)
nsample <- rnorm(500,mean=10,sd=3)
qqnorm(nsample)
```



Hide

```
# b)
qqnorm(score$f)
```



This just further proves our observations from before. In the normal sample that we made we can see a linear trend along the graph, but if we take a qqnorm of the scores from the final we can see that the curve is more curved, especially around the middle. This is most likely due to the right-skewed nature of the graph that we saw.

Problem 5l

Hide

```
stem(score$f)
```

```
  The decimal point is at the |

   6 | 00
   8 | 000000
  10 | 00000
  12 | 0000
  14 | 00000000000000
  16 | 00000
  18 | 0000000000
  20 | 0000000000000000
  22 | 00000000000000000
  24 | 000000000
  26 | 00000000000000
  28 | 000000000000000000000000000000000
  30 | 00000000000000000000000000000
  32 | 000000000000000000000000000000
  34 | 0000000000000000000000000
  36 | 000000000
  38 | 000
```

Hide

```
stem(score$f, scale=0.5)
```

```
  The decimal point is 1 digit(s) to the right of the |

  0 | 67888999
  1 | 0011122334
  1 | 5555555555555677778888889999
  2 | 000001111111111122222223333333333444444
  2 | 55566666666777777888888888888999999999999999999999
  3 | 00000000000000011111111111111222222222222223333333333333333444444444
  3 | 555555555555666667777899
```

Hide

```
stem(score$f, scale=2)
```

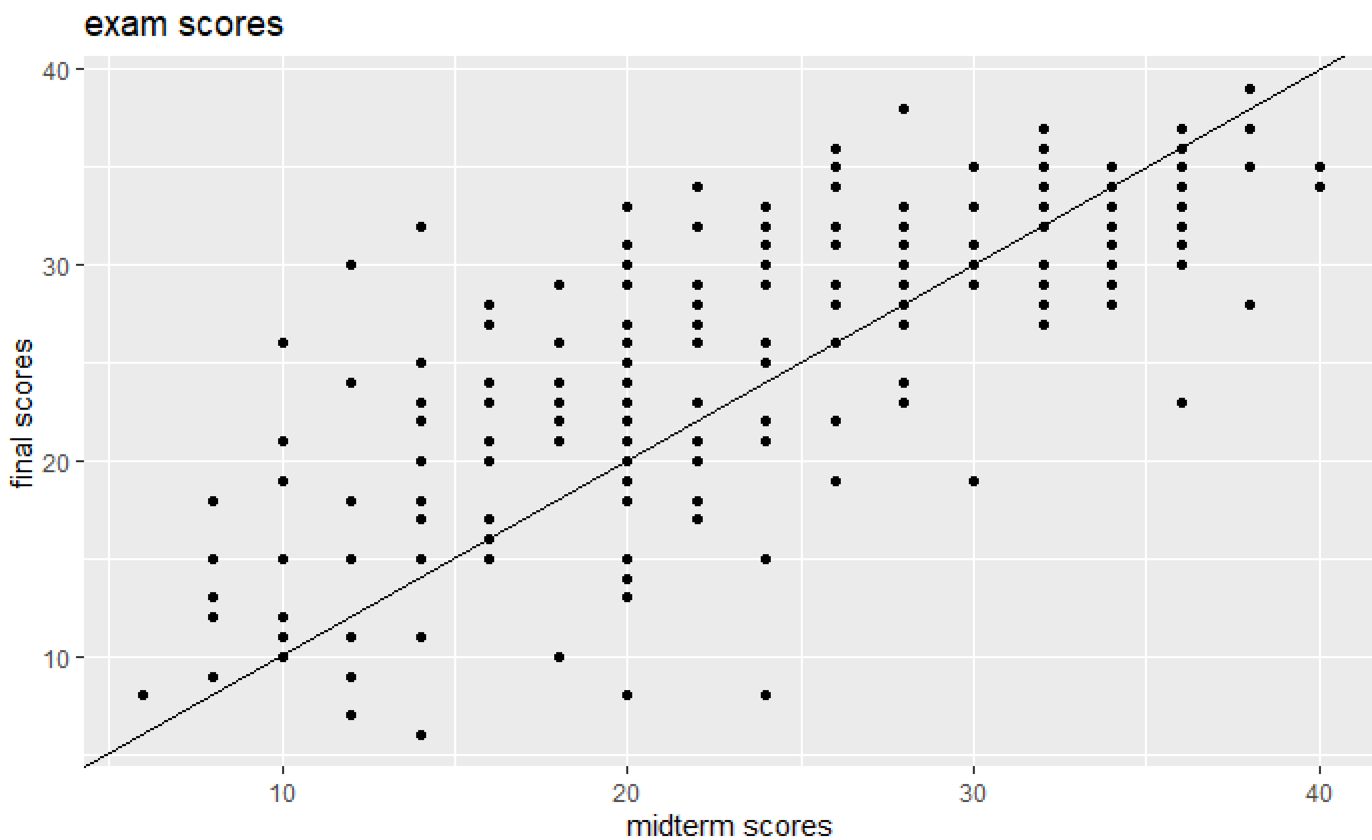
```
The decimal point is at the |

 6 | 0
 7 | 0
 8 | 000
 9 | 000
10 | 00
11 | 000
12 | 00
13 | 00
14 | 0
15 | 0000000000000
16 | 0
17 | 0000
18 | 000000
19 | 0000
20 | 00000
21 | 00000000000
22 | 0000000
23 | 0000000000
24 | 000000
25 | 000
26 | 00000000
27 | 000000
28 | 000000000000
29 | 000000000000000000000
30 | 000000000000000
31 | 00000000000000
32 | 00000000000000
33 | 0000000000000000
34 | 0000000000000
35 | 000000000000
36 | 00000
37 | 0000
38 | 0
39 | 00
```

It seems that the default stem and leaf plot seems to be the easiest to visually use, however I think that the last stem and leaf plot can also be used if exact grading distribution would like to be seen. Overall, I think the first one and the last one are significantly better than the middle one because it does not allow you to see the distribution of the scores for the exam and is too complicated visually while looking at it.

Problem 5J

Hide

```
ggplot(score, aes(x=m,y=f))+
  geom_point()+
  ggtitle("exam scores")+
  xlab("midterm scores")+
  ylab("final scores")+
  geom_abline()
```



Hide

```
c <- sum(score$m<score$f)
c
```

```
[1] 137
```

Hide

```
# [1] 137
```

The amount of people that did better due to the grading curve was 137 students.

A line that is more in line with the actual distribution of the grades would fit better. Estimation using m=1 for the slope of the line is not a good fit because the line is left with over 50% of the data lying above the line, a smaller slope or moving the line above the 0 intercept would be a better choice.

5k)

$$\frac{f(x_1, x_2, \dots x_n|\theta)}{f(y_1, y_2, \dots y_n|\theta)} = \prod_{i=1}^{n} \frac{\theta e^{-(\theta+1)\ln(1+x_i)}}{\theta e^{-(\theta+1)\ln(1+y_i)}}$$

$$= e^{-(\theta+1)\sum_{i=1}^{n}[\ln(1+x_i)-\ln(1+y_i)]}$$

independant iff $\Longleftrightarrow$

$$\sum_{i=1}^{n} \ln(1+x_i) = \sum_{i=1}^{n} \ln(1+y_i)$$

$$T = \sum_{i=1}^{n} \ln(1+x_i)$$

is sufficient statistic

5L)

T and U are both sufficient

$$SP_U \subseteq LP_U, SP_T \subseteq LP_T$$

While T is minimally sufficient, U is not minimally sufficient because

$$SP_T = LP_T \quad SP_U \neq LP_U$$