5A 8.20

$\bar{X} \sim N(E(\bar{X}), \text{Var}(\bar{X}))$ by CLT

$E(\bar{X}) = \mu = 0$

$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{100}{25} = 4$ so $\bar{X} \sim N(0, 4)$

$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \Rightarrow s^2 = \frac{n}{n-1} \hat{\sigma}^2$

$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

We know, $\frac{(n-1)}{\sigma^2} s^2 \sim \chi^2_{n-1}$

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

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

So $\hat{\sigma}^2 \sim \frac{100}{25} \chi^2_{25-1} = 4\chi^2_{24}$

5B 8.26

There are n animals in the population.

We assume all $\binom{n}{50}$ combinations to capture are equally likely.

The probability that 20 are tagged: $L_n = \dfrac{\binom{100}{20}\binom{n-100}{30}}{\binom{n}{50}}$

We would use the method of maximum likelihood to estimate the population n.

- Ratio of successive terms

$$\frac{L_n}{L_{n-1}} = \frac{(n-100)(n-50)}{n(n-130)} = \frac{n^2-150n+5000}{n^2-130n}$$

- this is >1 if (increasing L_n)

$$n^2-150n+500 > n^2-130n$$
$$-150n+5000 > -130n$$
$$5000 > 20n$$
$$\Downarrow$$
$$n < 250$$

- this is <1 if $n > 250$ (decreasing L_n)

- So $n=250$ is the MLE

5C) 8.30

a) $f(x_1|\lambda) = \lambda e^{-5\lambda}$

$f(x_2|\lambda) = \lambda e^{-3\lambda}$

$P(x_3 > 10) = 1 - P(X \leq 10) = 1 - (1 - e^{-\lambda 10}) = e^{-10\lambda}$

The likelihood function is the product:

$lik(\lambda) = \lambda^2 e^{-5\lambda - 3\lambda - 10\lambda} = \lambda^2 e^{-18\lambda}$

b) Find MLE of λ:

$l(\lambda) = \log(lik(\lambda)) = \log(\lambda^2) + \log(e^{-18\lambda})$

$= 2\log(\lambda) - 18\lambda$

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

$\Rightarrow \frac{2}{\lambda} = 18$

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

5D 8.32

a) done using R

b) $t\left(\frac{.1}{2}\right) = 1.753$ — qt() in R

90% CI for μ: $\bar{x} \pm t\left(\frac{.1}{2}\right)\frac{s}{\sqrt{n}}$ (Done in R)

90% CI for σ^2: $\left(\frac{n\hat{\sigma}^2_{ML}}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{n\hat{\sigma}^2_{ML}}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right)$

c) 90% CI for σ: $\left(\sqrt{\frac{n\hat{\sigma}^2_{ML}}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}}, \sqrt{\frac{n\hat{\sigma}^2_{ML}}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}}\right)$

d) Length of CI is $L = 2t\left(\frac{.1}{2}\right)\frac{s}{\sqrt{n}}$

to half this length: $\frac{1}{\sqrt{kn}} = \frac{1}{2\sqrt{n}} \Rightarrow \sqrt{k} = 2 \Rightarrow k = 4$

Your sample would need to be $\times 4$ the size.

5E 8.48

δ-method gives us: $\text{Var}(g(x)) = g'(E(x))^2 \text{Var}(x)$

$$E(g(x)) = g(E(x)) + \frac{g''(E(x))\,E(x-E(x))^2}{2}$$

$g(y) = -\ln\left(\frac{y}{n}\right)$

$g'(y) = -\frac{n}{y}\cdot\frac{1}{n} = -\frac{1}{y}$

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

$E(Y) = ne^{-\lambda}$

$\text{Var}(Y) = ne^{-\lambda}(1-e^{-\lambda})$

Variance: $\text{Var}\left(-\ln\frac{Y}{n}\right) = \left(-\frac{1}{E(Y)}\right)^2 \cdot \text{Var}(Y) = \left(\frac{1}{ne^{-\lambda}}\right)^2 (ne^{-\lambda})(1-e^{-\lambda})$

$$= \frac{1-e^{-\lambda}}{ne^{-\lambda}} = \boxed{\frac{e^{\lambda}-1}{n}}$$

Expectation: $E\left(-\ln\frac{Y}{n}\right) = -\ln\left(\frac{E(Y)}{n}\right) + \frac{\frac{1}{E(Y)^2}\text{Var}(Y)}{2}$

$$= -\ln\left(\frac{ne^{-\lambda}}{n}\right) + \frac{\frac{1}{(ne^{-\lambda})^2}(ne^{-\lambda})(1-e^{-\lambda})}{2}$$

$$= -\ln(e^{-\lambda}) + \frac{e^{\lambda}-1}{2n} = \lambda + \frac{e^{\lambda}-1}{2n}$$

Bias: $E\left(-\ln\frac{Y}{n}\right) - \lambda = \boxed{\frac{e^{\lambda}-1}{2n}}$

MLE for poisson λ, $\hat{\lambda}_{ML} = \bar{X}$

$\text{Var}(\hat{\lambda}_{ML}) = \frac{\lambda}{n}$

Efficiency $\approx \frac{\text{Var}(\hat{\lambda}_n)}{\text{Var}(\tilde{\lambda}_n)} = \frac{\frac{\lambda}{n}}{\frac{e^{\lambda}-1}{n}} = \frac{\lambda}{e^{\lambda}-1} < 1$

so $\hat{\lambda}_{ML}$ always has lower variance except when $\lambda = 0$

5F

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

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

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

($E(\hat{\theta}) - \theta = Bias(\hat{\theta})$)

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

($E(\hat{\theta}) - \theta = Bias(\hat{\theta})$)

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

$$= Var(\hat{\theta}) + 2Bias(\hat{\theta})\underbrace{\left(E(\hat{\theta}) - E(\hat{\theta})\right)}_{=0} + Bias(\hat{\theta})^2$$

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

5K 8.71

$$f(x_1,\dots,x_n|\theta) = \prod_{i=1}^{n} \frac{\theta}{(1+x_i)^{\theta+1}} = \theta^n \prod_{i=1}^{n} \frac{1}{(1+x_i)^{\theta+1}} = \theta^n \left(\prod_{i=1}^{n} \frac{1}{(1+x_i)}\right)^{\theta+1}$$

Let $T_1 = \prod_{i=1}^{n} \frac{1}{(1+x_i)}$, then $f(x_1,\dots,x_n|\theta) = \theta^n T_1^{\theta+1}$

By the factorization theorem, T_1 is a sufficient statistic.

MSS:

$$\frac{\prod_{i=1}^{n} \frac{\theta}{(1+x_i)^{\theta+1}}}{\prod_{i=1}^{n} \frac{\theta}{(1+y_i)^{\theta+1}}} = \frac{\prod_{i=1}^{n} \frac{1}{(1+x_i)^{\theta+1}}}{\prod_{i=1}^{n} \frac{1}{(1+y_i)^{\theta+1}}} = \left(\frac{\prod_{i=1}^{n} \frac{1}{(1+x_i)}}{\prod_{i=1}^{n} \frac{1}{(1+y_i)}}\right)^{\theta+1}$$

This must equal 1 for it not to depend on θ.

So, top = bottom, and MSS is $\prod_{i=1}^{n} \frac{1}{(1+x_i)}$

A non minimal sufficient statistic

$T_2 = \left(\prod_{i=1}^{n} \frac{1}{(1+x_i)}\right)^2$ doesn't add any data

but now $T(x^n) = T(y^n) \not\Leftrightarrow \frac{f(x^n|\theta)}{f(y^n|\theta)}$

5L

T is sufficient because the statistic partition forms a subset of the likelihood partition. In fact SP = LP so T is MSS.

U is sufficient because SP $\subseteq$ LP.

However, U is not MSS because

$u = 91$	$f(x_1, x_2, x_3 \mid \mu) = 1$
$u = 103$	$f(x_1, x_2, x_3 \mid \mu) = 1$

There are more equivalence classes of U than sets in the likelihood partition.

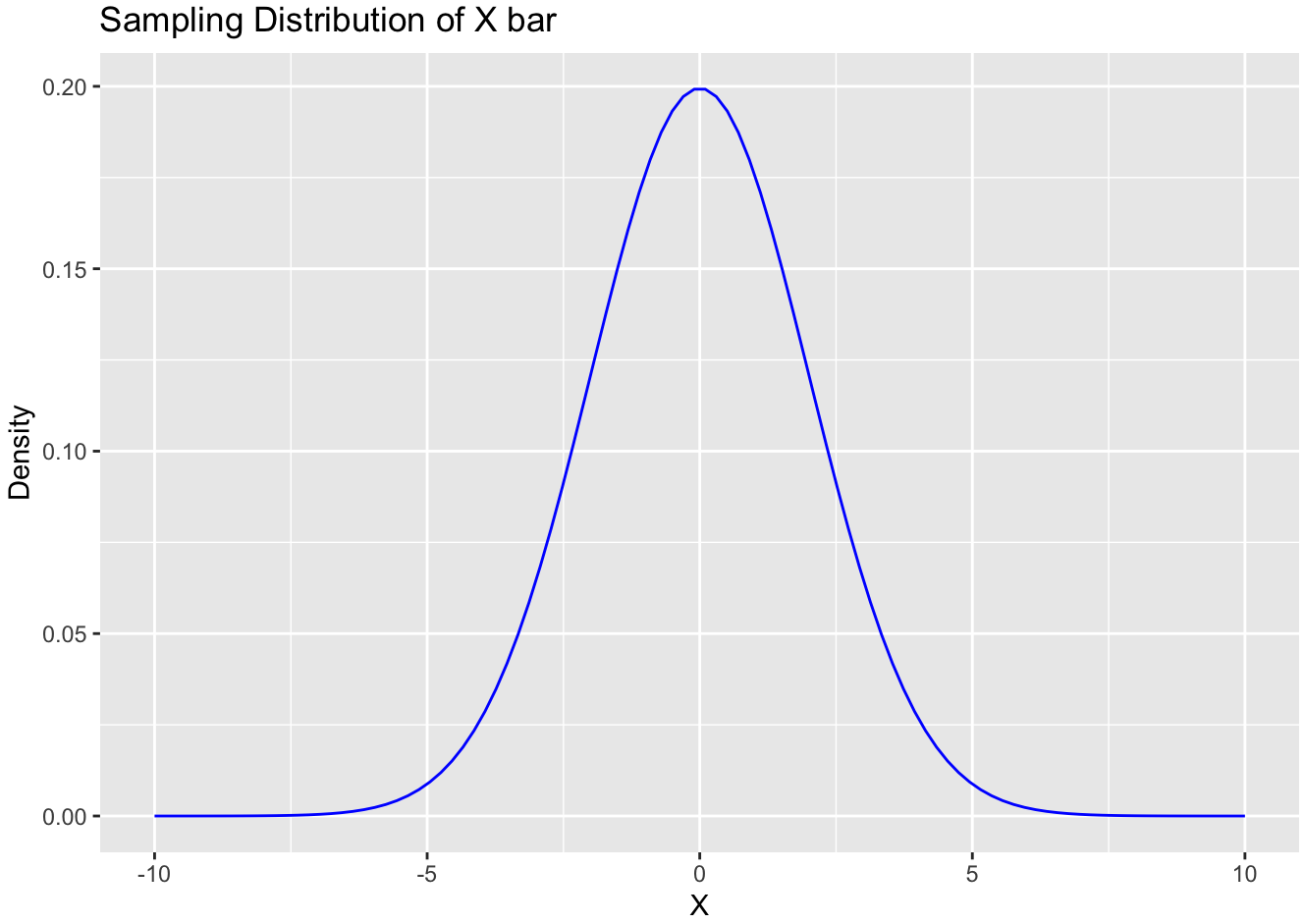
HW5

Natalie Brewer

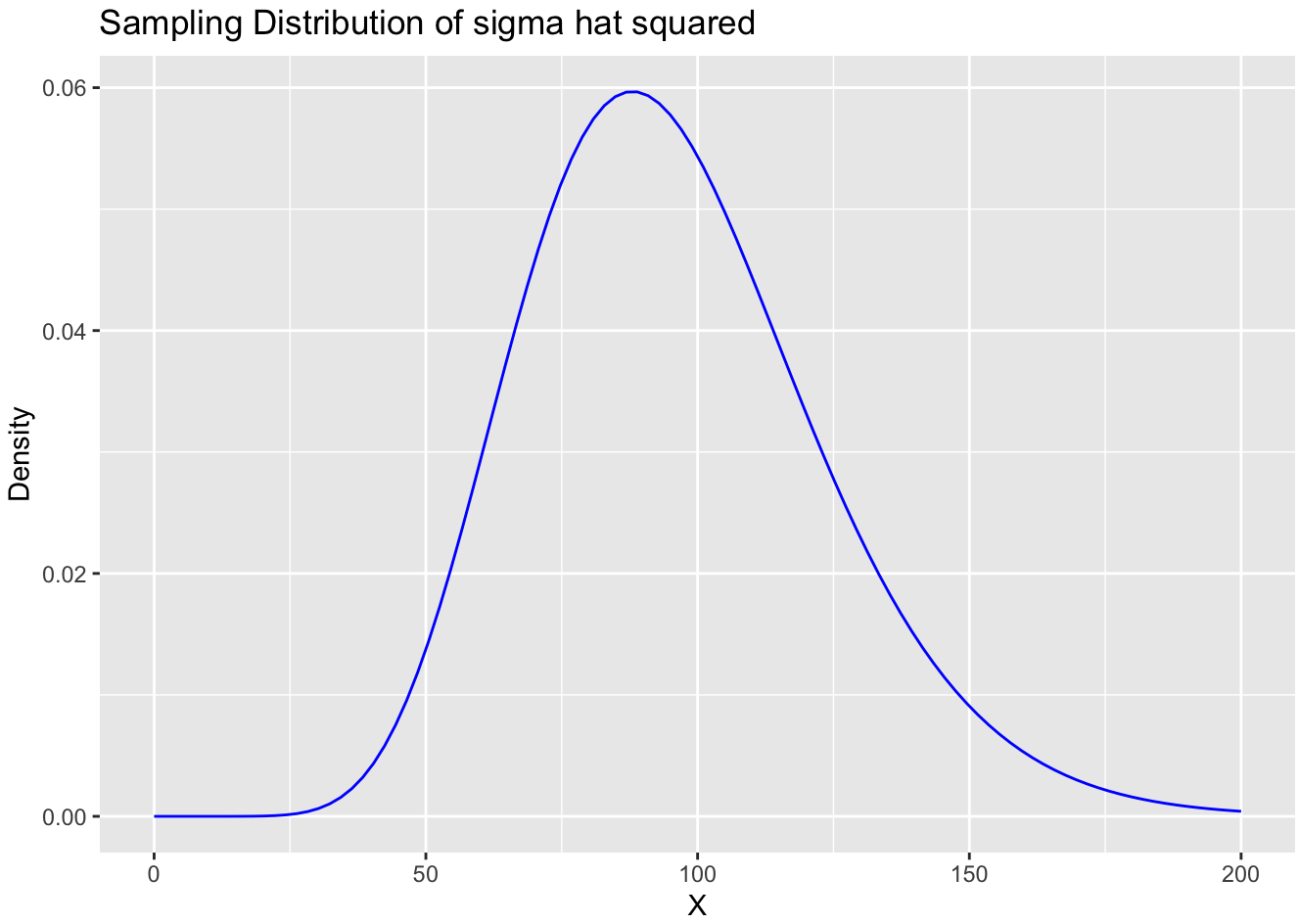
2023-09-28

Problem 5A

```
#sampling distribution for xbar
exp_xbar = 0
exp_sd = sqrt(100/25)

x_values <- seq(-10, 10, length.out = 100)
y_values <- dnorm(x_values, exp_xbar, exp_sd) #calculate density for the x values

df <- data.frame(x = x_values, y = y_values)

ggplot(df, aes(x, y)) +
  geom_line(color = "blue") +
  ggtitle("Sampling Distribution of X bar") +
  xlab("X") +
  ylab("Density")
```



```
#sampling distribution for sigma hat squared
x2_values <- seq(0,200, length.out = 100)
y2_values <- dchisq(x2_values/4, 24)

df2 <- data.frame(x2 = x2_values, y2 = y2_values)

ggplot(df2, aes(x2, y2)) +
  geom_line(color = "blue") +
  ggtitle("Sampling Distribution of sigma hat squared") +
  xlab("X") +
  ylab("Density")
```



Problem 5D

```
sample <- read.table("/Users/nataliebrewer/Desktop/Stat 135/HW5/data.8.32.txt", sep=",")
$V1
sample
```

```
##  [1] 5.3299 4.2537 3.1502 3.7032 1.6070 6.3923 3.1181 6.5941 3.5281 4.7433
## [11] 0.1077 1.5977 5.4920 1.7220 4.1547 2.2799
```

```
#MLEs
xbar <- mean(sample)
xbar
```

```
## [1] 3.610869
```

```
s2 <- sum((sample - xbar)^2) / 15
var_MLE <- (15/16) * s2
var_MLE
```

```
## [1] 3.204461
```

```
#CI for mu
t <- qt(1 - .1/2, 15)

CI_mu <-c(xbar - (t*sqrt(s2 / 16)), xbar + (t*sqrt(s2 / 16)))
CI_mu
```

```
## [1] 2.800605 4.421132
```

```
#CI for sigma squared
chi_right <- qchisq(.1/2, 15)
chi_left <- qchisq(1-.1/2, 15)

CI_var <- c((16*var_MLE)/chi_left, (16*var_MLE)/chi_right)
CI_var
```

```
## [1] 2.051201 7.061256
```

```
#CI for sigma
CI_sigma <- c(sqrt((16*var_MLE)/chi_left), sqrt((16*var_MLE)/chi_right))
CI_sigma
```

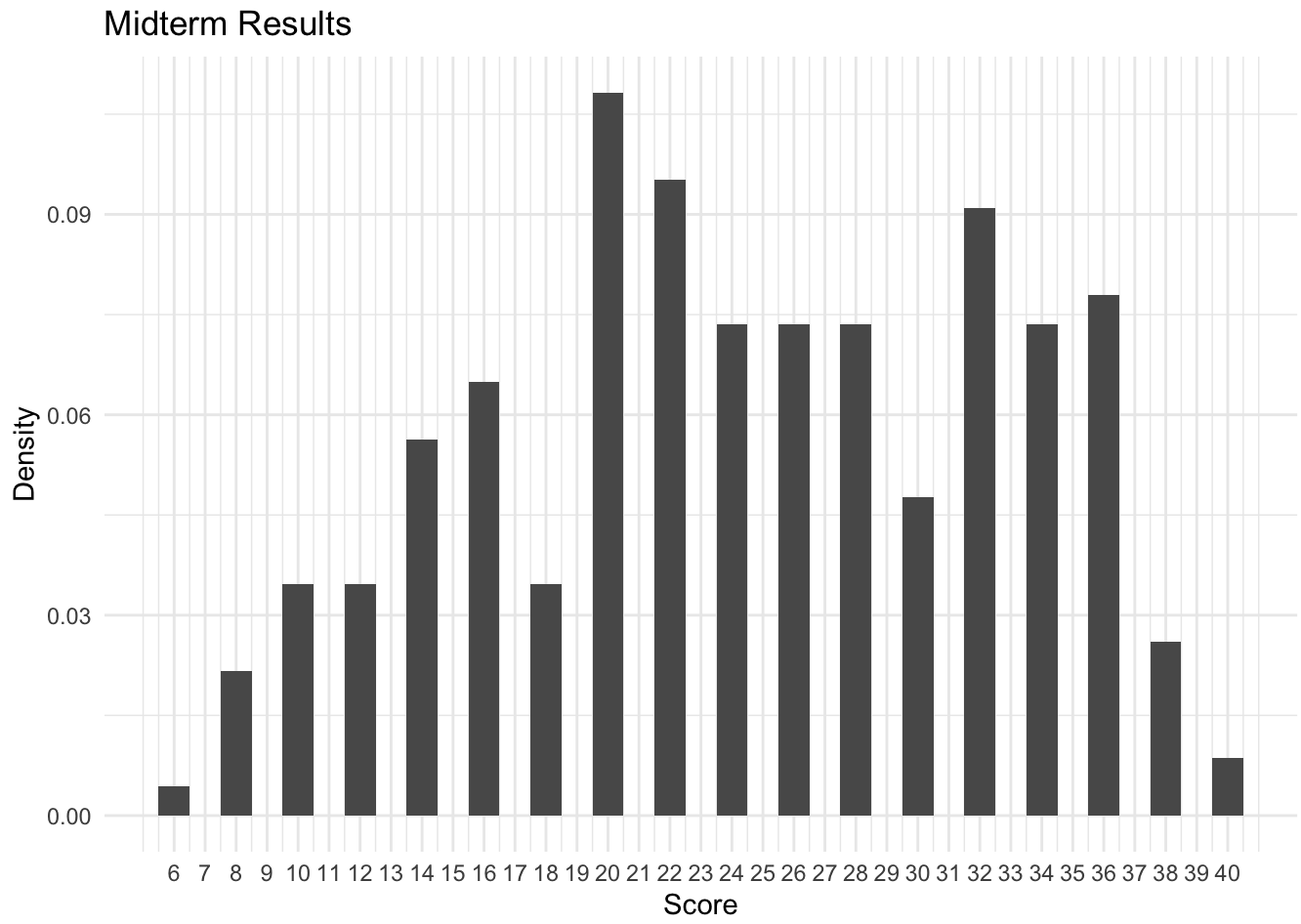
```
## [1] 1.432201 2.657302
```

Problem 5G

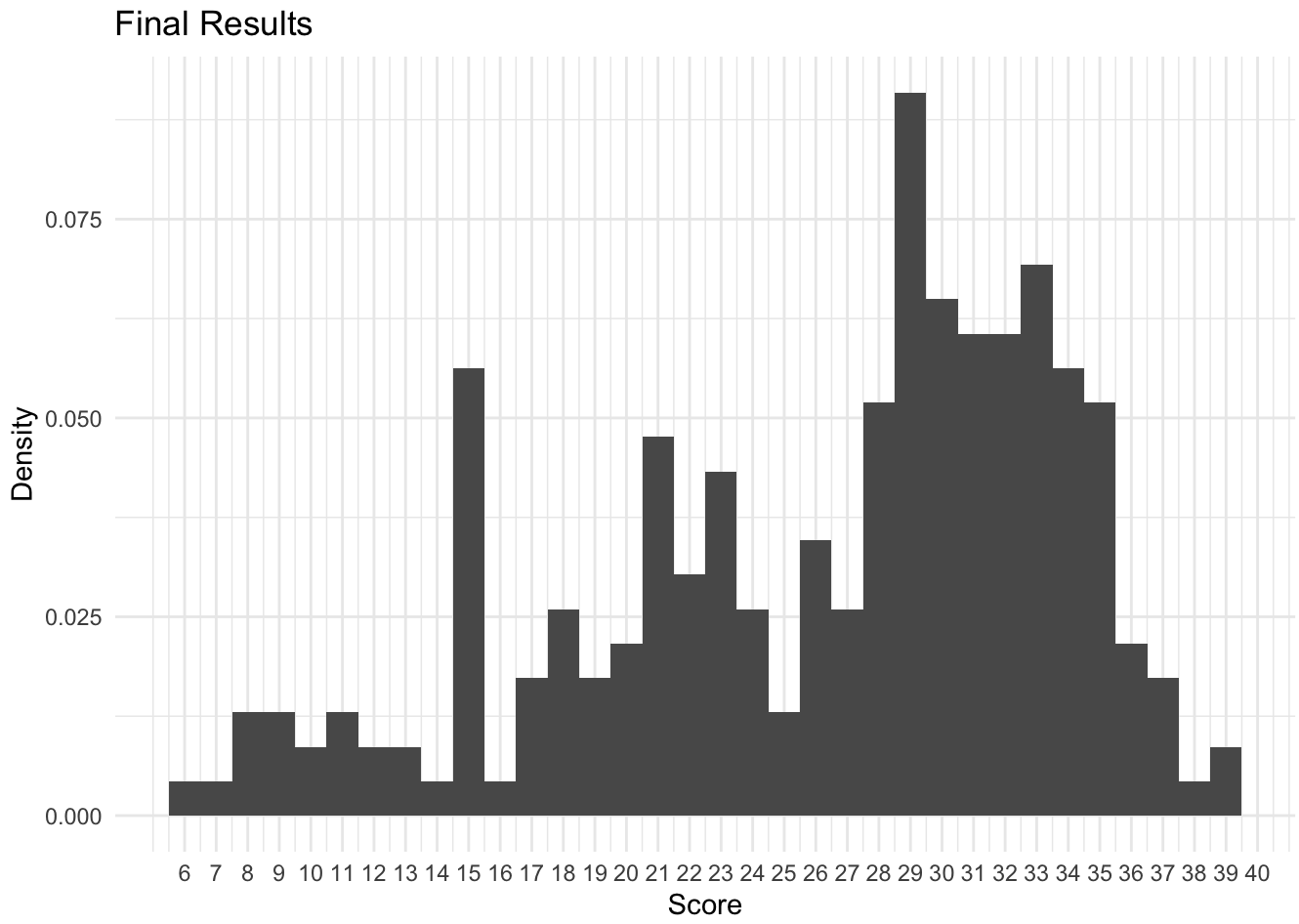
```
scores <- read.table("/Users/nataliebrewer/Desktop/Stat 135/HW5/data.scores.txt", sep="
")
colnames(scores) <- c("final", "midterm") #rename columns
scores <- scores[, c(2, 1)] #switch order of the columns
scores <- scores[-1, ] #remove first row to prevent NAs
scores$final <- as.numeric(scores$final) #convert strings to numbers
scores$midterm <- as.numeric(scores$midterm) #convert strings to numbers
scores <- scores[scores$final > 0 & scores$midterm > 0, ] # remove non-positive rows
scores$midterm <-  scores$midterm * 2 #multiply all midterm scores by 2
head(scores)
```

```
##   midterm final
## 2       8    13
## 3      10    12
## 4      14     6
## 5      16    27
## 6      14    23
## 7      36    32
```

```
#Histogram for midterm scores
ggplot(scores, aes(x=midterm)) +
  geom_histogram(aes(y=after_stat(density)), binwidth = 1) +
  labs(title="Midterm Results", x="Score", y="Density") +
  scale_x_continuous(breaks=seq(min(scores$midterm), max(scores$midterm), by=1)) +
  theme_minimal()
```



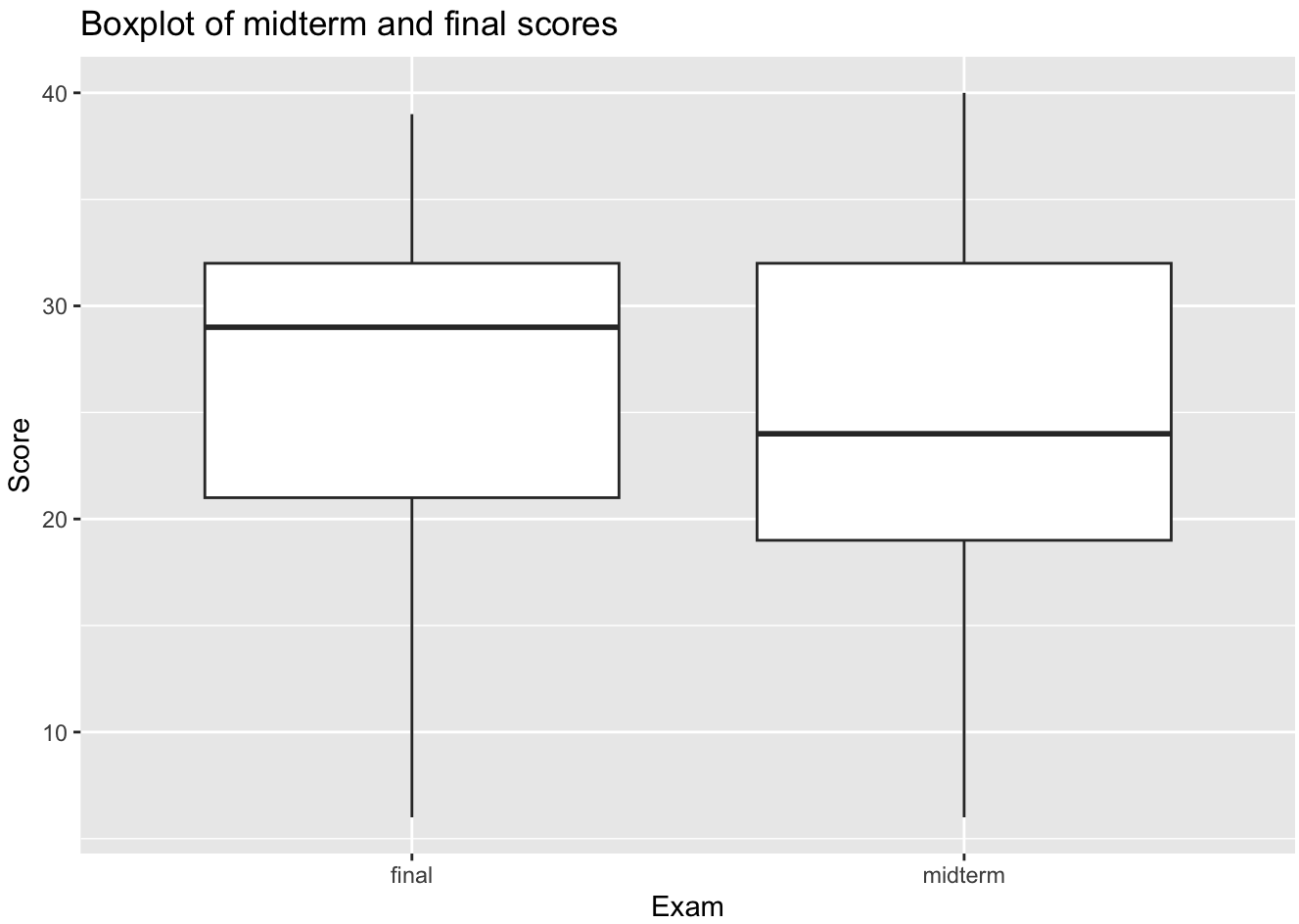

```
#Histogram for final scores
ggplot(scores, aes(x=final)) +
  geom_histogram(aes(y=after_stat(density)), binwidth = 1) +
  labs(title="Final Results", x="Score", y="Density") +
  scale_x_continuous(breaks=seq(min(scores$midterm), max(scores$midterm), by=1)) +
  theme_minimal()
```




```
#Convert data to narrow format
scores_narrow <-  scores %>%
  gather(key= when, value = score, `midterm`, `final`)
head(scores_narrow)
```

```
##      when score
## 1 midterm     8
## 2 midterm    10
## 3 midterm    14
## 4 midterm    16
## 5 midterm    14
## 6 midterm    36
```

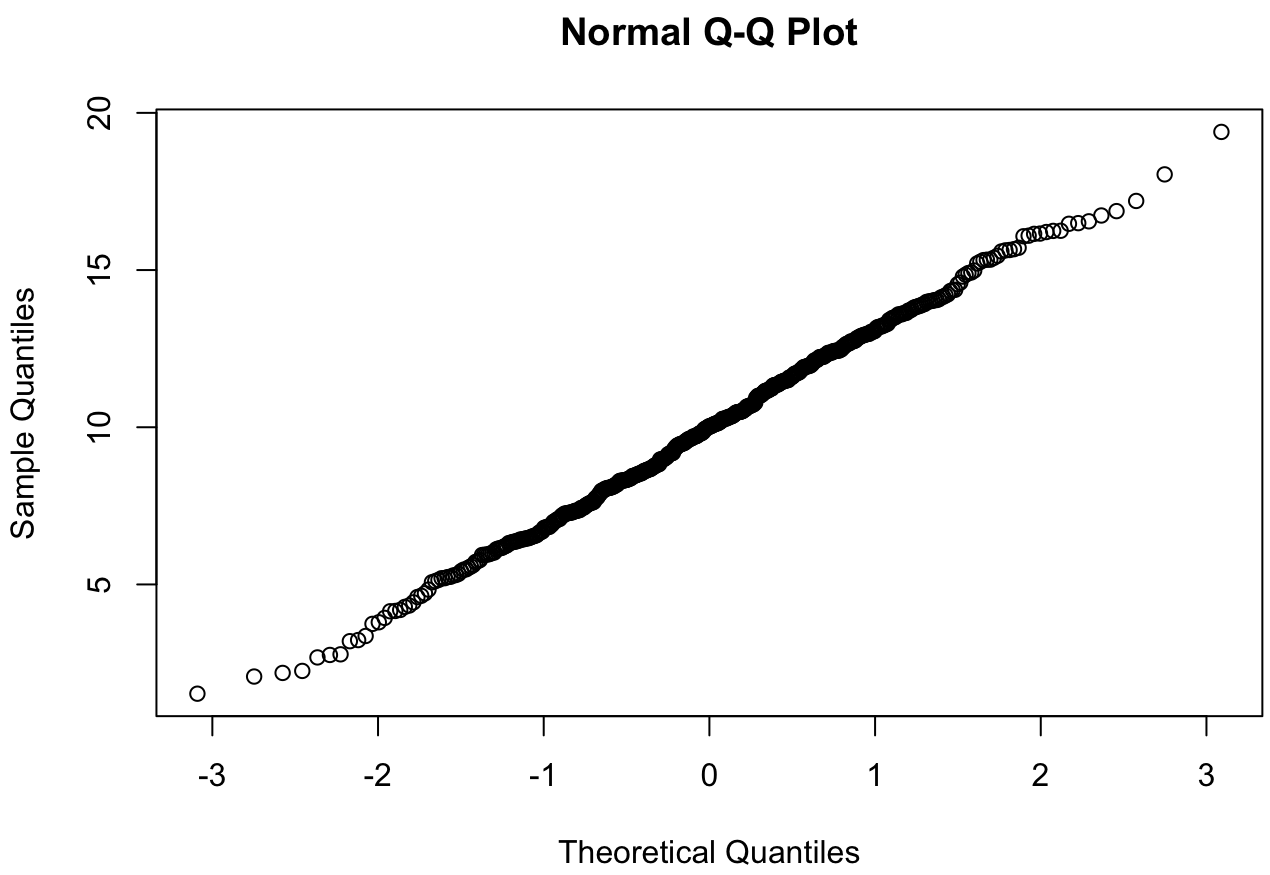
```
#Boxplot for midterm and final scores
ggplot(scores_narrow, aes(x = factor(when), y = score)) +
  geom_boxplot() +
  xlab("Exam") +
  ylab("Score") +
  ggtitle("Boxplot of midterm and final scores")
```



By comparing these two boxplots, we can make the following observations about the data. Firstly, we can see that the median score for the final was higher than the median score for the midterm. We can also see that the midterm scores exhibit a slightly larger spread. For the final scores, the lower quartile is much further than the upper quartile from the median. As for the midterm, the upper quartile is slightly further from the median than the lower quartile.

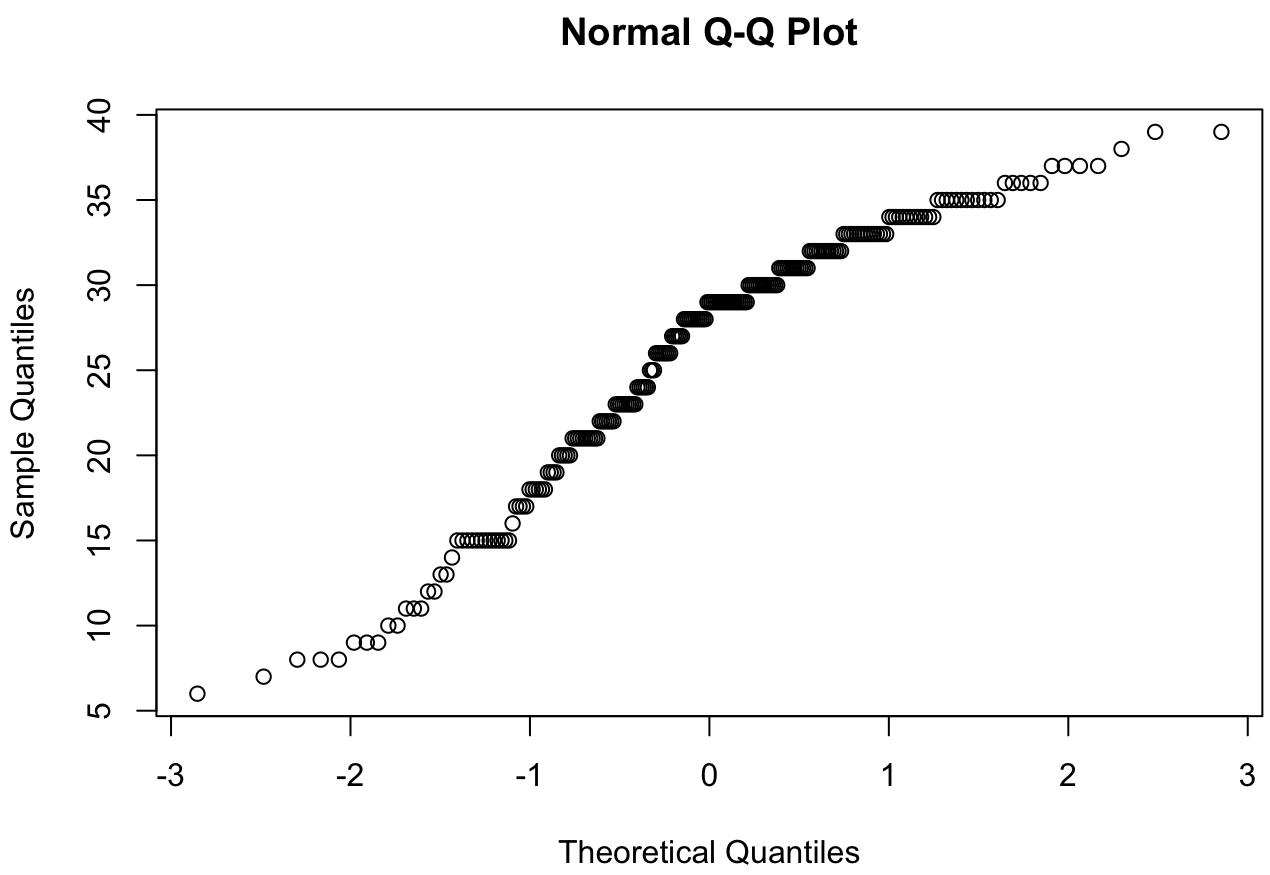
Problem 5H

```
normal_sample <- rnorm(500, 10, 3)
qqnorm(normal_sample)
```



This plot appears to be approximately linear.

```
qqnorm(scores$final)
```



This plot appears to be less linear than the previous plot. The final scores data plotted has slight concavity in the curvature meaning many of the scores are higher than the quantiles of the standard normal.This aligns with our observations from the boxplot, which as median that is closer to the upper quantile, making the data slightly upwards skewed.

Problem 5l

```
stem(scores$final, scale=0.5)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 67888999
##   1 | 0011122334
##   1 | 55555555555567777888888999
##   2 | 000001111111111222222233333333333444444
##   2 | 55566666666777777888888888889999999999999999999
##   3 | 000000000000001111111111111222222222222233333333333333333444444444
##   3 | 555555555555666667777899
```

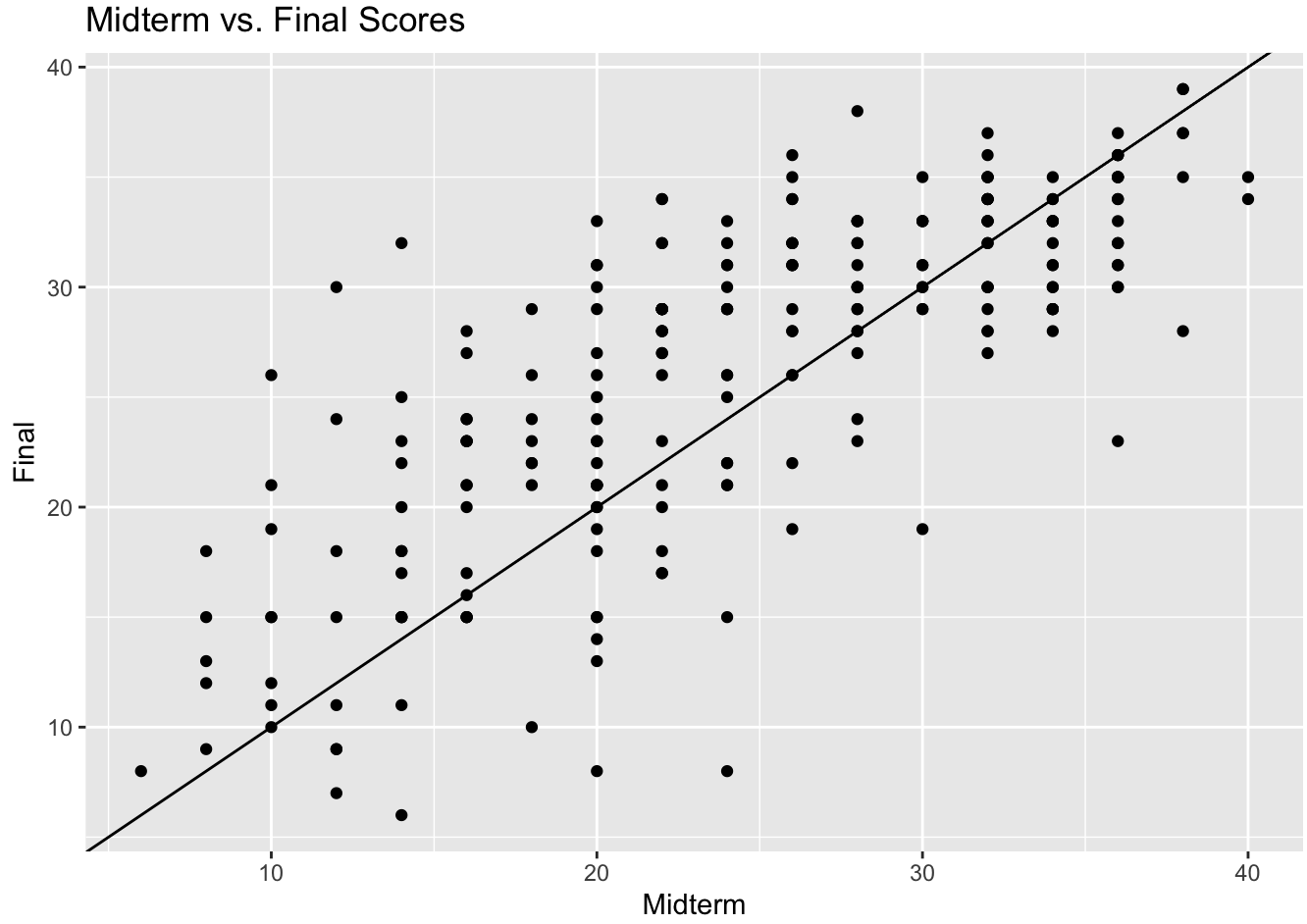
```
stem(scores$final, scale=2)
```

```
##
##   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 | 00000000000000000000
##   30 | 000000000000000
##   31 | 00000000000000
##   32 | 00000000000000
##   33 | 0000000000000000
##   34 | 0000000000000
##   35 | 000000000000
##   36 | 00000
##   37 | 0000
##   38 | 0
##   39 | 00
```

I think that both of these plots give useful insight into the nature of the distribution of the data. The 0.5 scale stem and leaf is useful for getting a very compressed view of the data. We can easily see that the range containing the most data values is 30 to 34, since this is the longest leaf. The second stem and leaf gives a more detailed view of the data. We can see exactly how many students received each score as well as the overall, rough way the scores are distributed. It is also easy to note slight abnormalities like the peculiarly high frequency of 15s. However, this plot is not as efficient at compressing the data into a succint summary.

Problem 5J

```
#Plot the midterm on x axis and final on the y axis

ggplot(scores, aes(x=midterm,y=final)) +
  geom_point() +
  geom_abline(slope=1, intercept=0) +
  ggtitle("Midterm vs. Final Scores") +
  xlab("Midterm") +
  ylab("Final")
```



Just from eyeballing this plot, it looks like more than 50% students gained from this grading scheme. This is because there are more students whose final score was greater than twice their midterm score (i.e. more data points lying above the x = y line).

```
percentage <- (sum(scores$final > scores$midterm) / nrow(scores)) * 100
percentage
```

```
## [1] 59.30736
```

This is consistent with my eyeballed estimate.

To try and estimate final score based on midterm score using a straight line, I would use a line with a steeper slope. This is because the current line has less than 50% of the data points below it, so a steeper line would split the data more evenly in two. A flatter slope would just result in more of the data lying above the estimate. A line with a slightly higher y-intercept would also result in more of the data lying beneath the line.