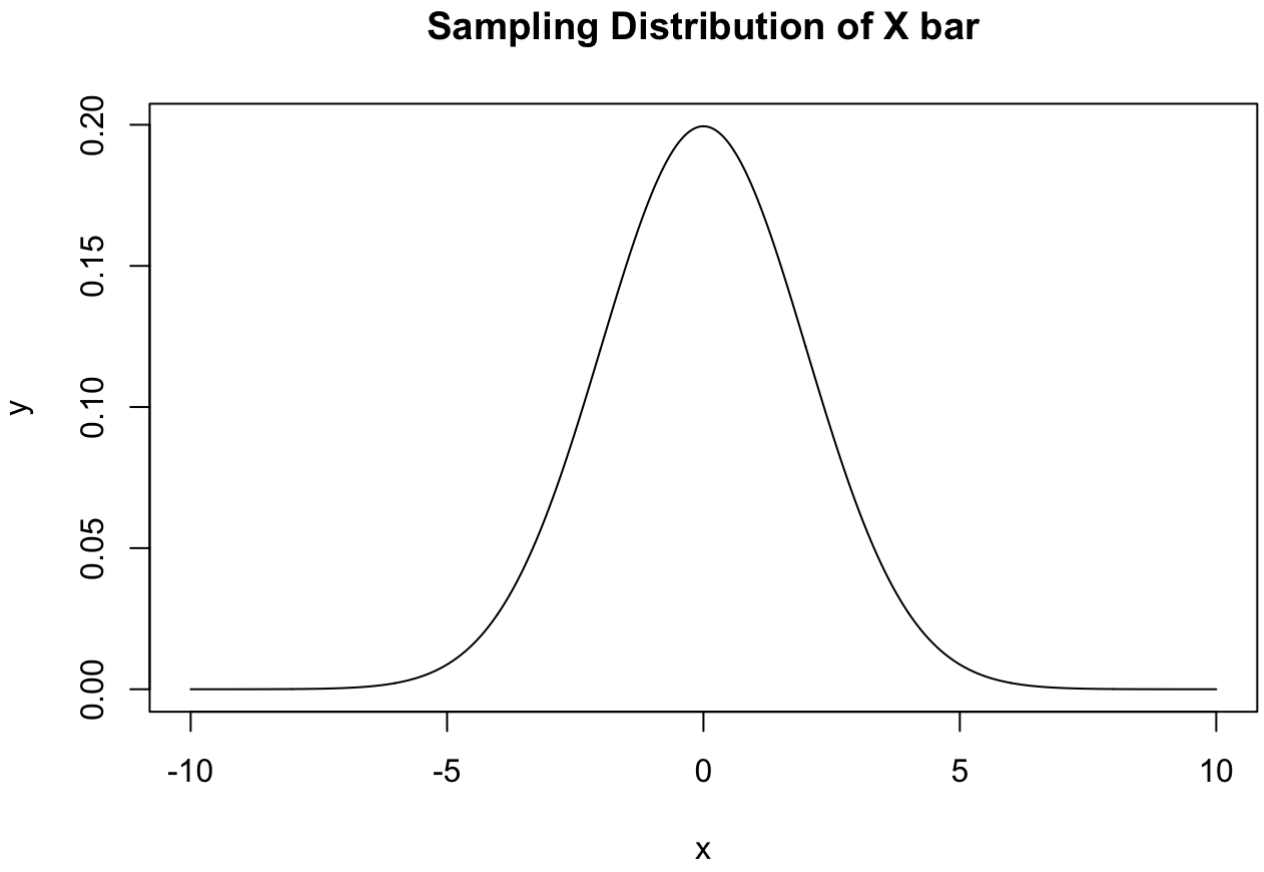
HW5

Ian Driscoll

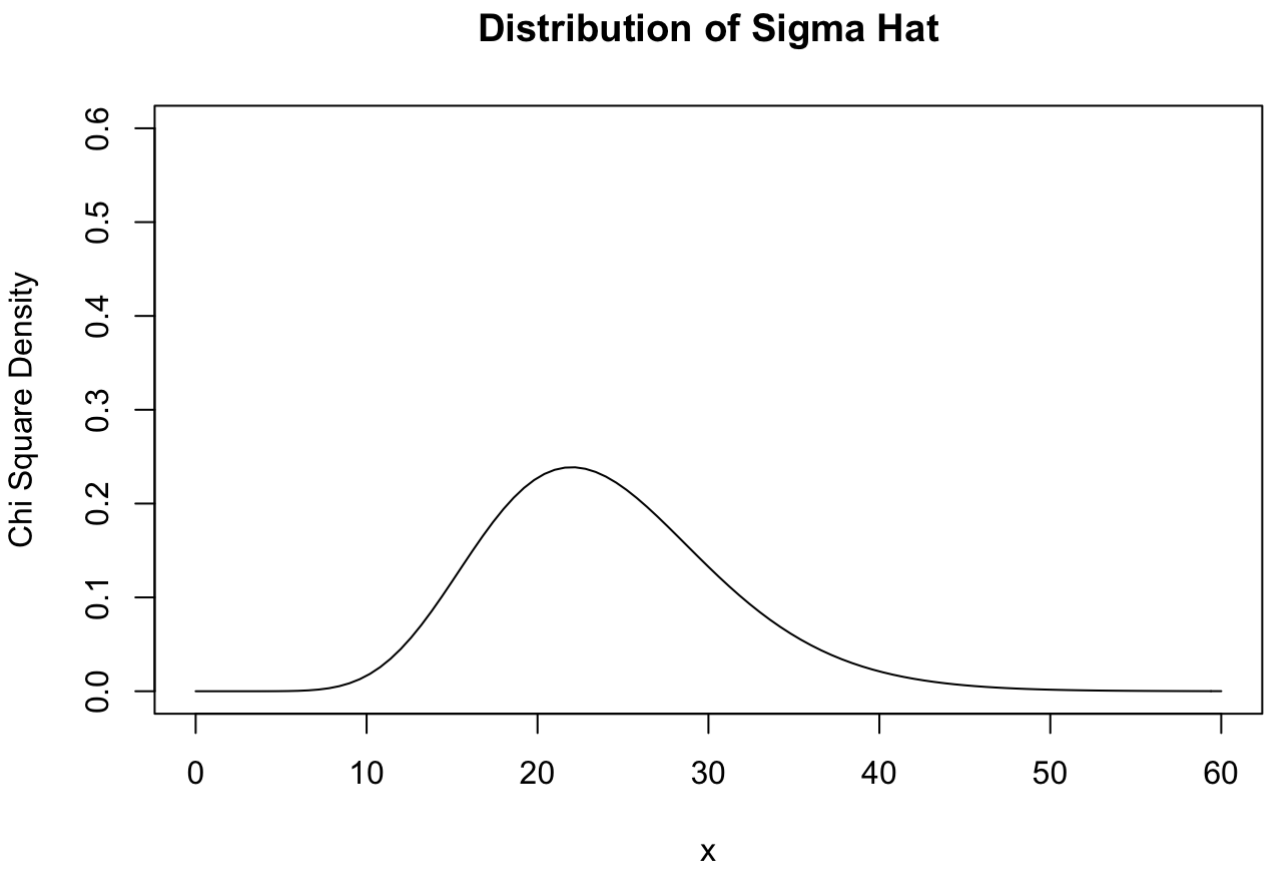
9/26/2018

5A.

```
# X bar
x = seq(-10, 10, length=1000)
y = dnorm(x, mean=0, sd=2)
plot(x, y, type="l", lwd=1, main = "Sampling Distribution of X bar")
```




```
# Sigma hat
curve(4 * dchisq(x,df = 24),xlim=c(0,60),ylim=c(0,0.6),ylab="Chi Square Density", main =
 "Distribution of Sigma Hat")
```



5D.

```
data = c(5.3299, 4.2537, 3.1502, 3.7032, 1.6070, 6.3923, 3.1181, 6.5941, 3.5281, 4.7433,
 0.1077, 1.5977, 5.4920, 1.7220, 4.1547, 2.2799)
moment1 = mean(data)
sample_var = var(data)
```

5G.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
scores = read.table("Desktop/data.scores.txt",header=TRUE)
scores = filter(scores, f > 0 & m > 0)
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

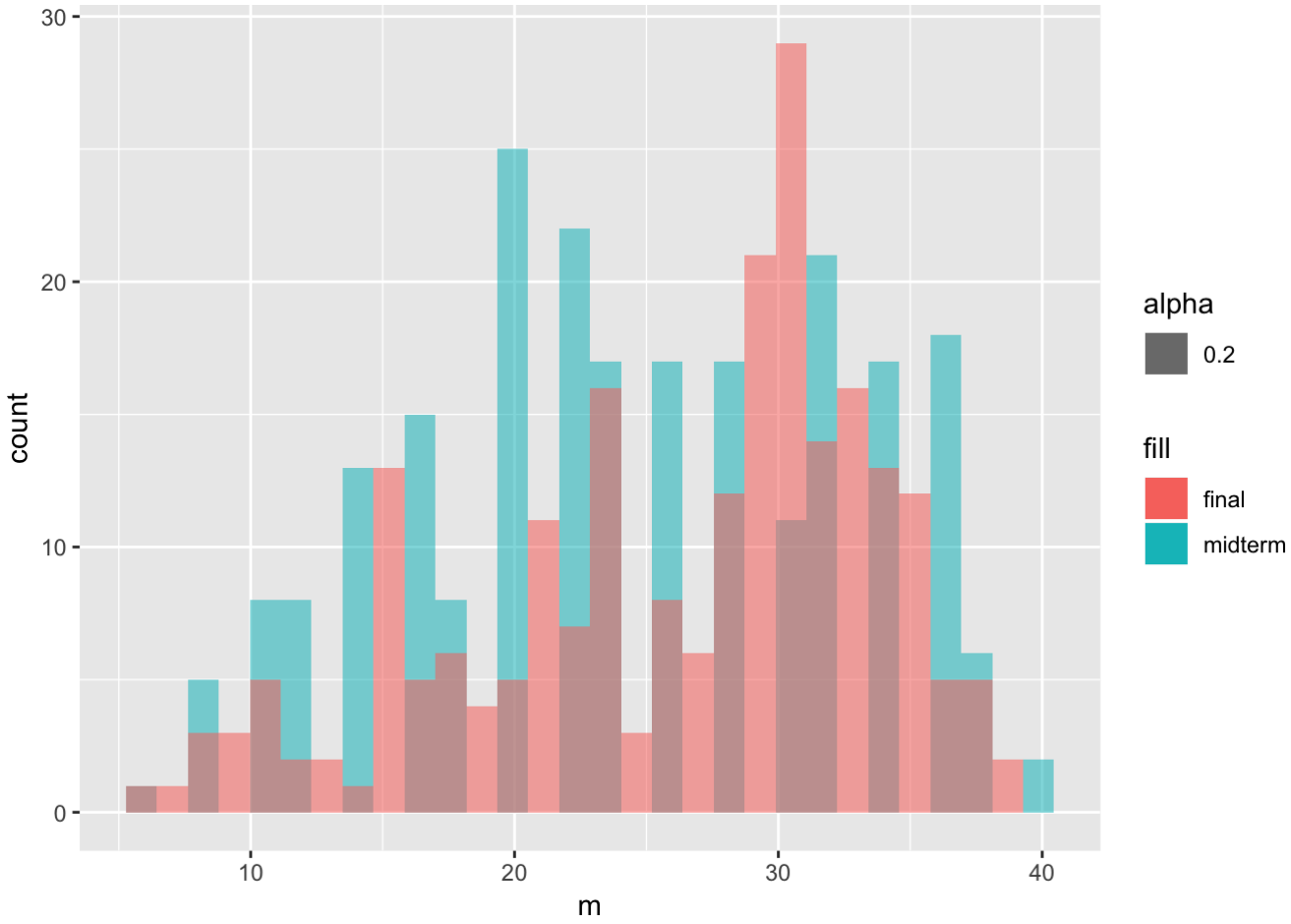
```
scores = data.frame(m = scores$m * 2, f = scores$f)

# Part A.

plot1 = ggplot(scores) + geom_histogram(aes(x = m, fill = 'midterm', alpha = 0.2)) + geo
m_histogram(aes(x = f, fill = 'final', alpha = 0.2))
plot1
```

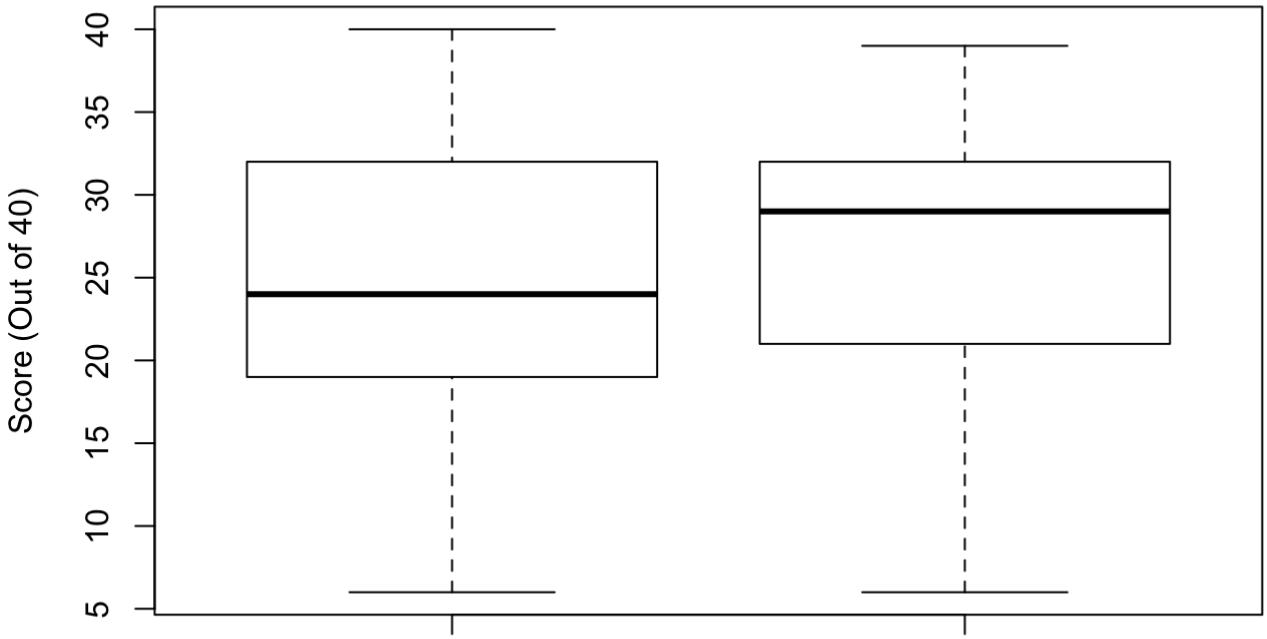
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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




```
# Part B.

boxplot(x = scores$m, scores$f, xlab = "Midterm vs. Final", ylab = "Score (Out of 40)")
```



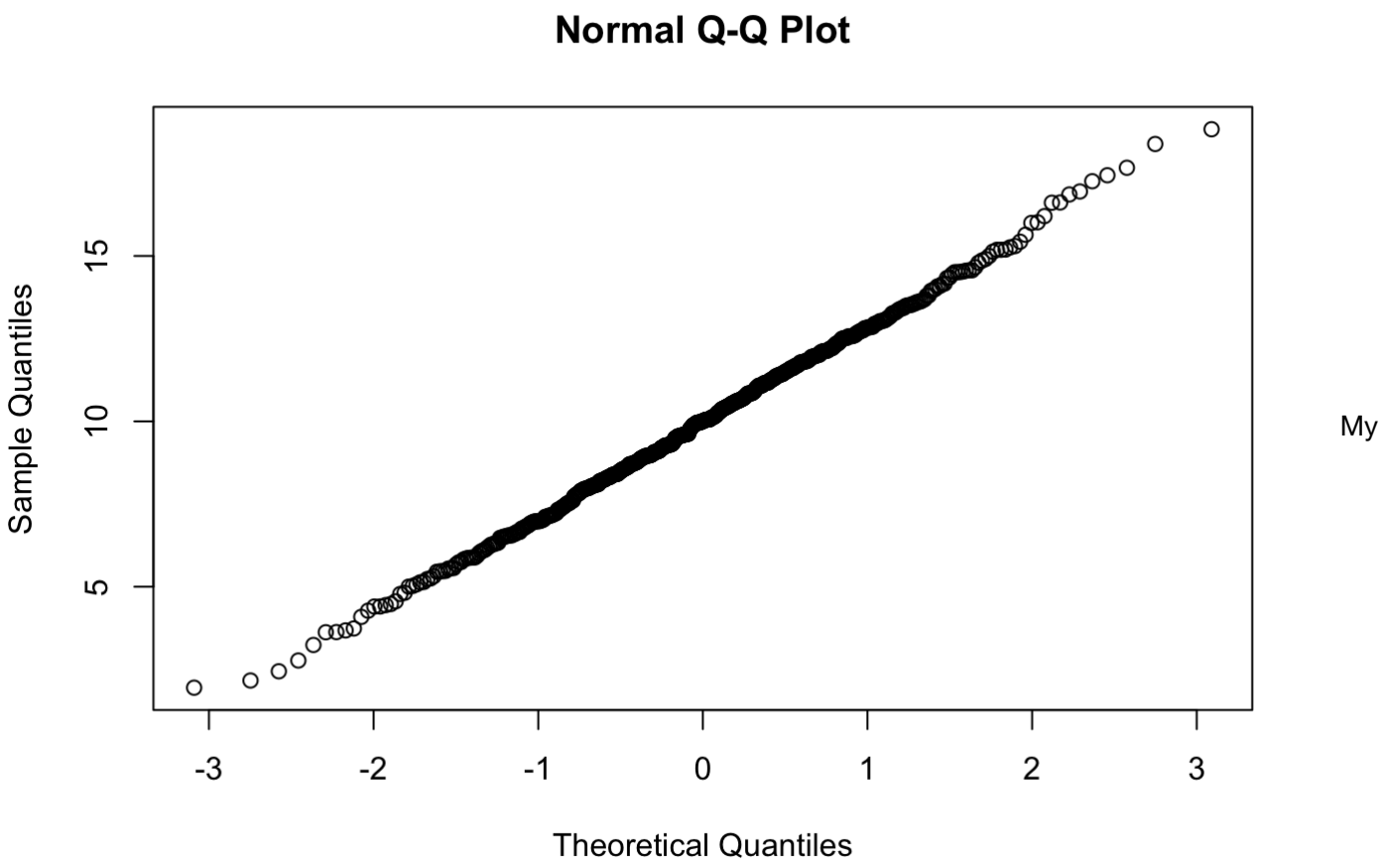
Midterm vs. Final

The spread is very similar for both the midterm and final, with people on average doing slightly better on the final. The upper third quartile is a smaller range for the final than the midterm, and the lower second quartile is larger than on the midterm. Both are centered around 24-26.

5H

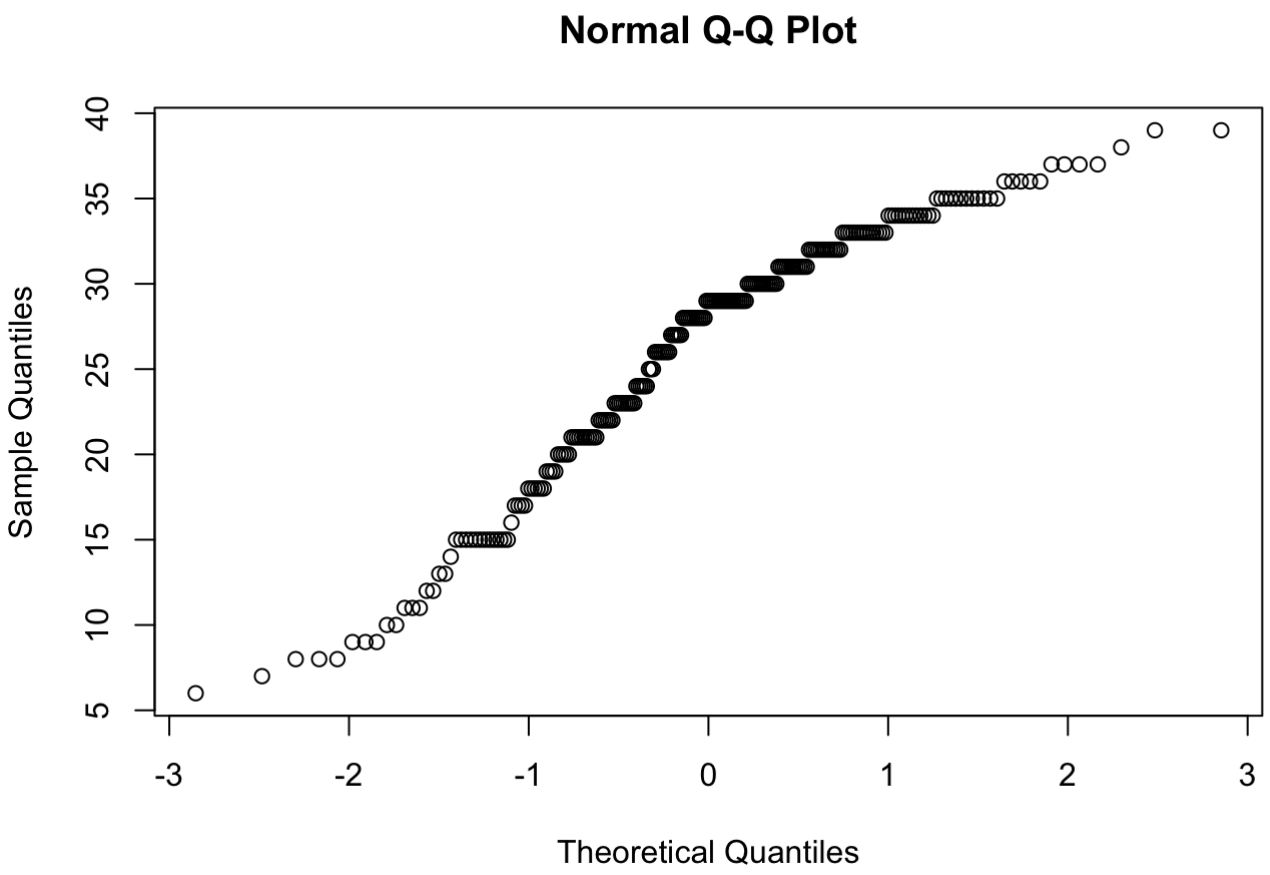
```
# Part A.

random500 = rnorm(500, 10, 3)
qqnorm(random500)
```



My plot does appear to be linear.

```
# Part B.
qqnorm(scores$f)
```



The histogram does not appear normal just like this plot, they both appear skewed to the left.

5l

```
stem(scores$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 | 00000000000000000000000000000000
##   30 | 0000000000000000000000000000
##   32 | 00000000000000000000000000000
##   34 | 000000000000000000000000
##   36 | 000000000
##   38 | 000
```

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

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 67888999
##   1 | 0011122334
##   1 | 55555555555567777888889999
##   2 | 000001111111111222222233333333333444444
##   2 | 55566666667777778888888888899999999999999999999
##   3 | 00000000000000111111111111122222222222223333333333333333344444444
##   3 | 555555555555666667777899
```

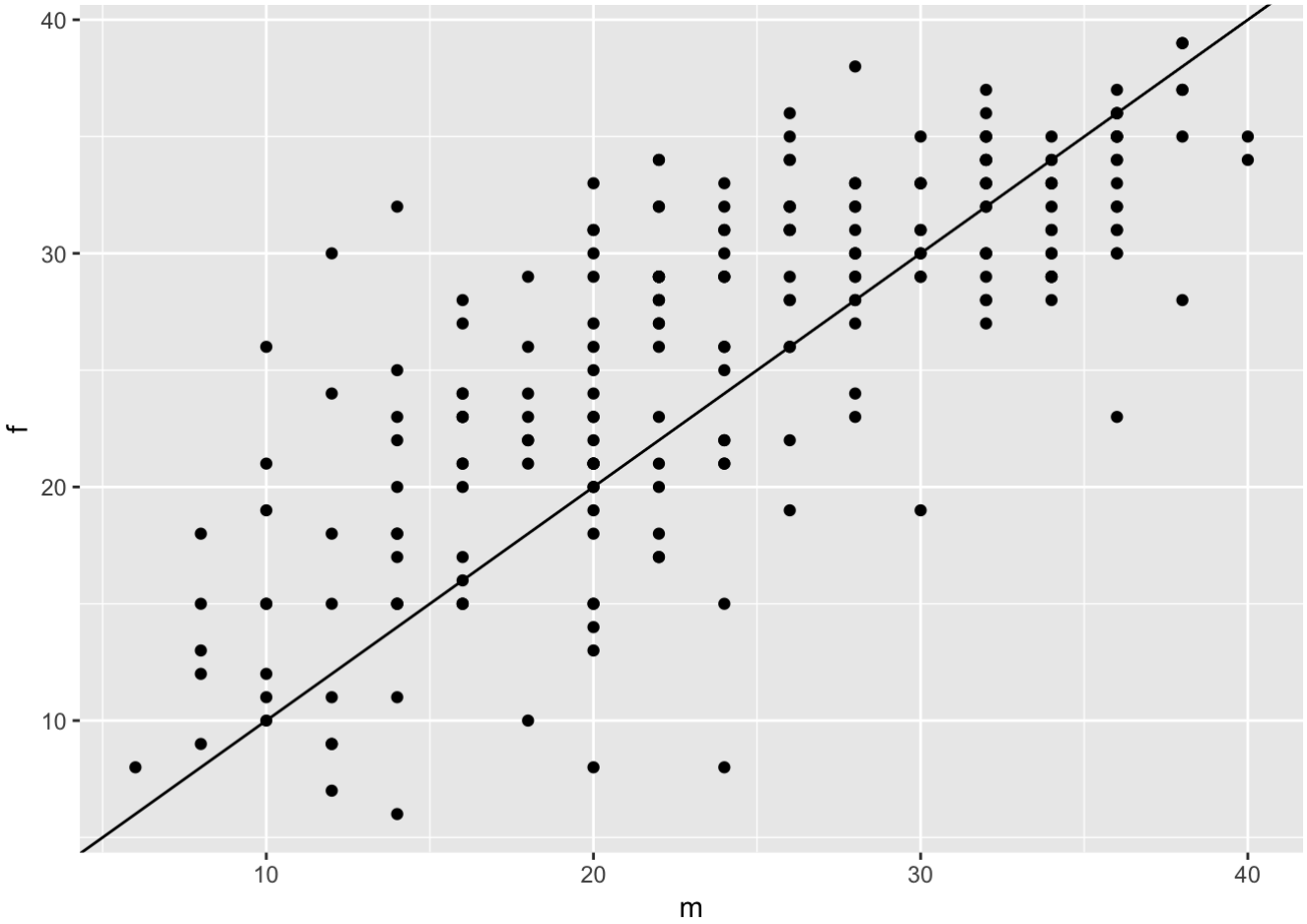
```
stem(scores$f, scale = 2)
```

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

The plot with scale 2 is the most detailed in the shape of the distribution of scores.

5J

```
# Part A.
ggplot(scores) + geom_point(aes(x = m, y = f)) + geom_abline(slope = 1, intercept = 0)
```



Eyeballing the graph, I would say the percentage who gained from the midterm is greater than 50%.

```
# Part B.
change = scores$f - scores$m
percent = length(change[change > 0]) / length(scores$f) * 100
percent
```

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

Part C.

For this plot, I would do a slightly less steep line with a larger y-intercept, since the line would have to be shifted up due to final scores being higher on average overall, and it would be less steep than a slope of 1 since scores would have to level off as they got closer and closer to 40 since they are starting from a higher y coordinate.