Homework_2

Warren Geither

9/16/2020

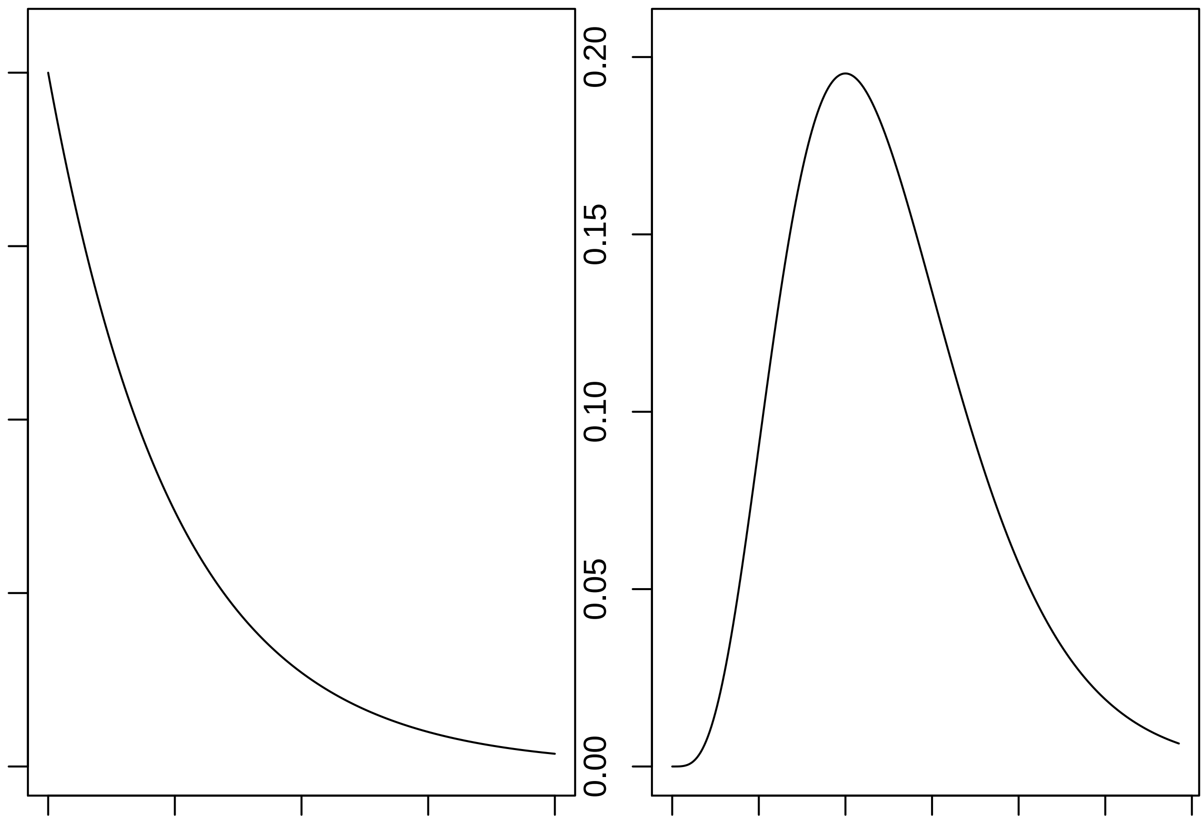
Problem 1

a.)

```r
# parameters for x1
x1_alpha <- 1
x1_beta <- 5
x1_mu <- x1_alpha*x1_beta
x1_sig <- sqrt(x1_alpha*x1_beta^2)
x1_var <- x1_alpha*x1_beta^2
x1_skew <- 2/sqrt(x1_alpha)
x1_params <- c(alpha = x1_alpha
               , beta = x1_beta
               , mu = x1_mu
               , sig = x1_sig
               , var = x1_var
               , skew = x1_skew)

# parameters for x2
x2_alpha <- 5
x2_beta <- 1
x2_mu <- x2_alpha*x2_beta
x2_sig <- sqrt(x2_alpha*x2_beta^2)
x2_var <- x2_alpha*x2_beta^2
x2_skew <- 2/sqrt(x2_alpha)
x2_params <- c(alpha = x2_alpha
               , beta = x2_beta
               , mu = x2_mu
               , sig = x2_sig
               , var = x2_var
               , skew = x2_skew)

# create param df
param_df <- as.data.frame(rbind(x1_params,x2_params))

# print values
knitr::kable(param_df)
```

	alpha	beta	mu	sig	var	skew
x1_params	1	5	5	5.000000	25	2.0000000
x2_params	5	1	5	2.236068	5	0.8944272

```r
# set range for distributions
x1_range <- seq(0, x1_mu + 3*x1_sig,0.01)
x2_range <- seq(0, x2_mu + 3*x2_sig,0.01)

# draw from distributions
x_1 <- dgamma(x1_range, x1_alpha, rate =1/x1_beta)
x_2 <- dgamma(x2_range, x2_alpha, rate =1/x2_beta)

# get 2 plots on same output window
par(mfrow=c(1,2))

# fix error for margin sizes
par(mar=c(1,1,1,1))
plot(x1_range, x_1, type='l', ylim=c(0,max(x_1)+0.01))
plot(x2_range, x_2, type='l', ylim=c(0,max(x_2)+0.01))
```



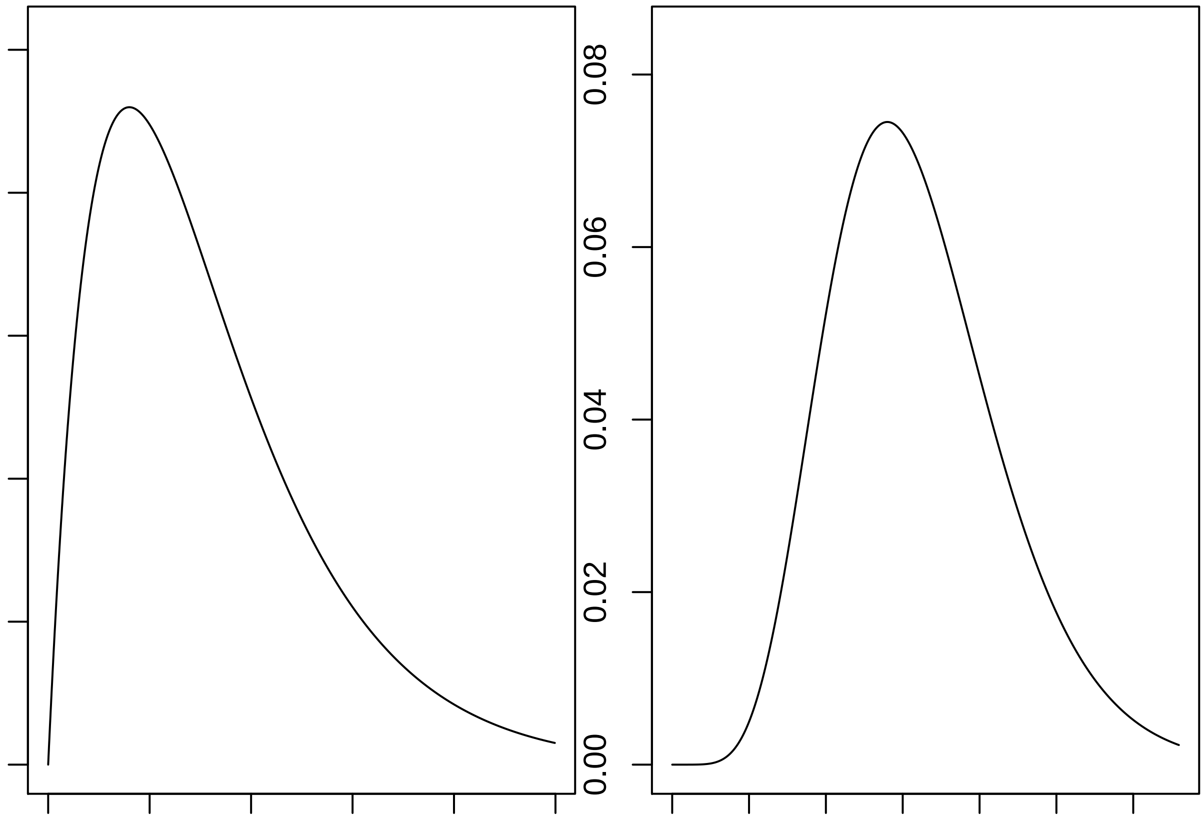
b.)

```r
# parameters for x1
x1_alpha <- 2
x1_beta <- 4
x1_mu <- x1_alpha*x1_beta
x1_sig <- sqrt(x1_alpha*x1_beta^2)
x1_var <- x1_alpha*x1_beta^2
x1_skew <- 2/sqrt(x1_alpha)
x1_params <- c(alpha = x1_alpha
               , beta = x1_beta
               , mu = x1_mu
```

```r
               , sig = x1_sig
               , var = x1_var
               , skew = x1_skew)

# parameters for x2
x2_alpha <- 8
x2_beta <- 2
x2_mu <- x2_alpha*x2_beta
x2_sig <- sqrt(x2_alpha*x2_beta^2)
x2_var <- x2_alpha*x2_beta^2
x2_skew <- 2/sqrt(x2_alpha)
x2_params <- c(alpha = x2_alpha
               , beta = x2_beta
               , mu = x2_mu
               , sig = x2_sig
               , var = x2_var
               , skew = x2_skew)

# create param df
param_df <- as.data.frame(rbind(x1_params,x2_params))

# print values
knitr::kable(param_df)
```

	alpha	beta	mu	sig	var	skew
x1_params	2	4	8	5.656854	32	1.4142136
x2_params	8	2	16	5.656854	32	0.7071068

```r
# set range for distributions
x1_range <- seq(0, x1_mu + 3*x1_sig,0.01)
x2_range <- seq(0, x2_mu + 3*x2_sig,0.01)

# draw from distributions
x_1 <- dgamma(x1_range, x1_alpha, rate =1/x1_beta)
x_2 <- dgamma(x2_range, x2_alpha, rate =1/x2_beta)

# get 2 plots on same output window
par(mfrow=c(1,2))

# fix error for margin sizes
par(mar=c(1,1,1,1))
plot(x1_range, x_1, type='l', ylim=c(0,max(x_1)+0.01))
plot(x2_range, x_2, type='l', ylim=c(0,max(x_2)+0.01))
```



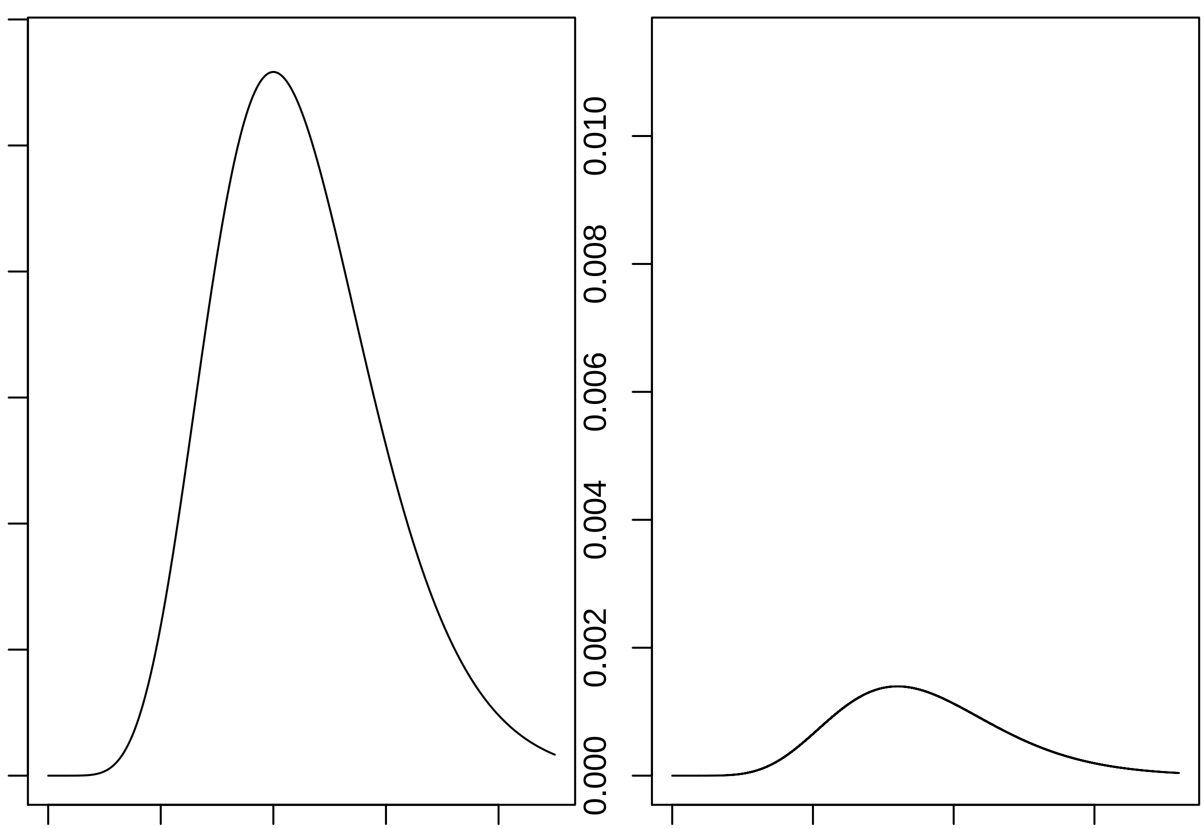
c.)

```r
# parameters for x1
x1_alpha <- 9
x1_beta <- 0.5
x1_mu <- x1_alpha*x1_beta
x1_sig <- sqrt(x1_alpha*x1_beta^2)
x1_var <- x1_alpha*x1_beta^2
x1_skew <- 2/sqrt(x1_alpha)
x1_params <- c(alpha = x1_alpha
               , beta = x1_beta
               , mu = x1_mu
               , sig = x1_sig
               , var = x1_var
               , skew = x1_skew)

# parameters for x2
x2_alpha <- 9
x2_beta <- 100
x2_mu <- x2_alpha*x2_beta
x2_sig <- sqrt(x2_alpha*x2_beta^2)
x2_var <- x2_alpha*x2_beta^2
x2_skew <- 2/sqrt(x2_alpha)
x2_params <- c(alpha = x2_alpha
               , beta = x2_beta
               , mu = x2_mu
               , sig = x2_sig
               , var = x2_var
               , skew = x2_skew)
```

```r
# create param df
param_df <- as.data.frame(rbind(x1_params,x2_params))

# print values
knitr::kable(param_df)
```

	alpha	beta	mu	sig	var	skew
x1_params	9	0.5	4.5	1.5	2.25	0.6666667
x2_params	9	100.0	900.0	300.0	90000.00	0.6666667

```r
# set range for distributions
x1_range <- seq(0, x1_mu + 3*x1_sig,0.01)
x2_range <- seq(0, x2_mu + 3*x2_sig,0.01)

# draw from distributions
x_1 <- dgamma(x1_range, x1_alpha, rate =1/x1_beta)
x_2 <- dgamma(x2_range, x2_alpha, rate =1/x2_beta)

# get 2 plots on same output window
par(mfrow=c(1,2))

# fix error for margin sizes
par(mar=c(1,1,1,1))
plot(x1_range, x_1, type='l', ylim=c(0,max(x_1)+0.01))
plot(x2_range, x_2, type='l', ylim=c(0,max(x_2)+0.01))
```



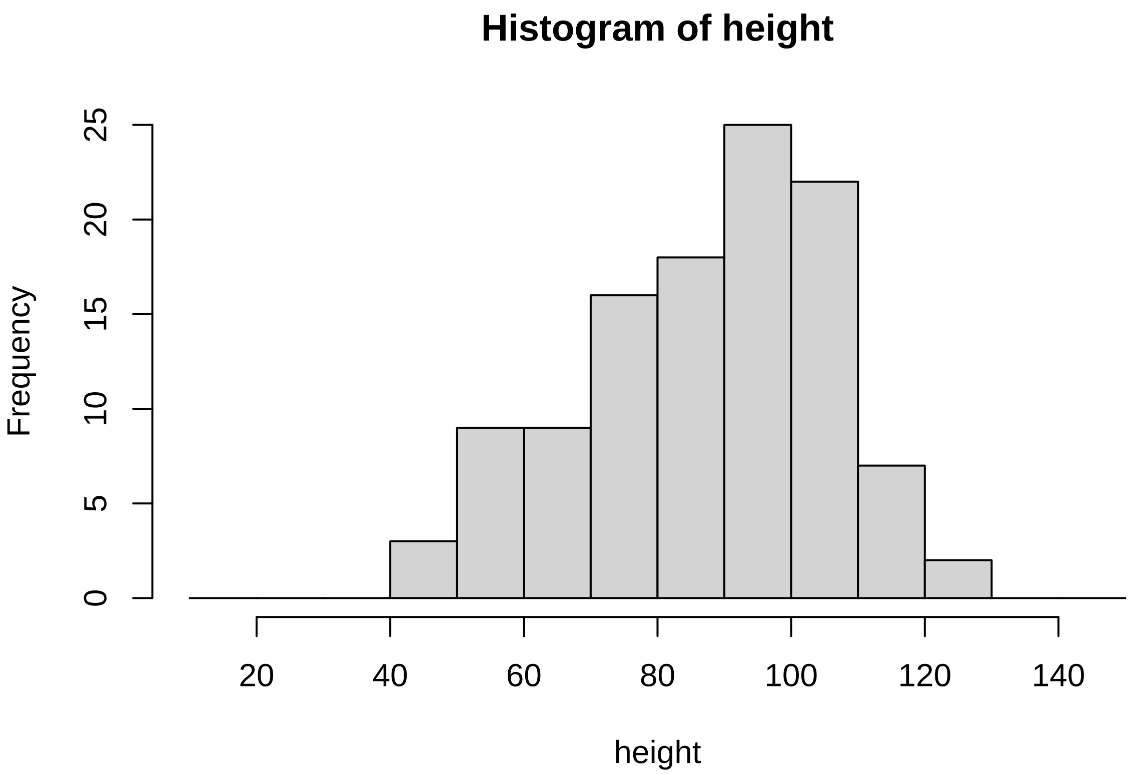
d.) Not possible. Different skew implies different alphas. Which implies you can get the same mean, but variance will never equal

e.) Not possible. To have the same skew and mean, you need to fix both alpha and beta. Which implies variance will be the same as well

f.) Not possible. You would need to fix both alpha and beta to achieve the same variance and different skew. But this implies mean is the same as well.

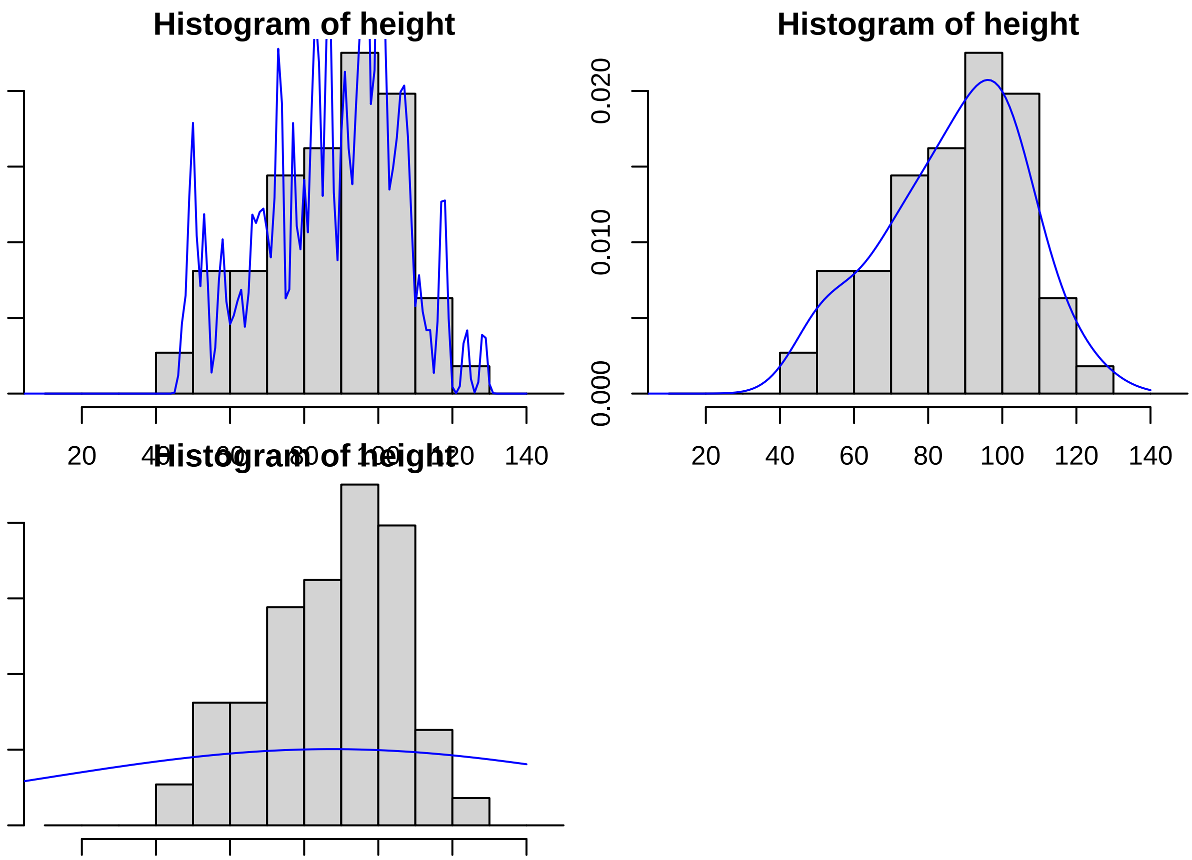
Problem 2

```
# load data
treedat <- read.csv2("/cloud/project/treedat.csv", sep=",")

# look at data
height <- as.vector(as.numeric(treedat$height))
hist(x=height, breaks=10*seq(1:15))
```




```
# distribution we are using in kde N(0,1)
norm_density <- function(x) {
                              value <- exp(-.5*x^2)/sqrt(2*pi)
                              return(value)
                            }

# default bandwidth
h = 0.77

# plug in x - x_i/h to the normal density function and find the mean
kde = function(x) {
```

```r
                    value_2 <- mean(norm_density((x-height)/h)/h)
                    return(value_2)
                  }

# apply this for all x
kde_2 = function(x) {
                      sapply(x,kde)
                    }

# set up plot
par(mar=c(1,1,1,1))
par(mfrow=c(2,2))
grid <- seq(1,140,by=1)

# set bandwidth and plot kde on histogram
h = 0.77
hist(x=height, freq=FALSE, breaks=10*seq(1:15))
lines(grid, kde_2(grid), col="blue")

# find mode
mode_1 <- which(kde_2(grid)==max(kde_2(grid)))
prob_1 <- sum(kde_2(grid)[1:91])

h = 7.7
hist(x=height, freq=FALSE, breaks=10*seq(1:15))
lines(grid, kde_2(grid), col="blue")

mode_2 <- which(kde_2(grid)==max(kde_2(grid)))
prob_2 <- sum(kde_2(grid)[1:91])

h = 77
hist(x=height, freq=FALSE, breaks=10*seq(1:15))
lines(grid, kde_2(grid), col="blue")

mode_3 <- which(kde_2(grid)==max(kde_2(grid)))
prob_3 <- sum(kde_2(grid)[1:91])
```



4.)

```
mode_results <- c("h=0.77"=mode_1, "h=7.7"=mode_2, "h=77"=mode_3)
knitr::kable(mode_results, col.names = c("mode"))
```

	mode
h=0.77	96
h=7.7	96
h=77	87

The change in bandwidth does not seem to heavily influence the mode. Which makes sense because the difference in values isnt changing so the h is just putting it on a different scale.

A practitioner could start with the rule of thumb 7.7 and adjust from there through trial and error depending what looks the best against the data.

A bad choice for bandwidth could result in very bad estimates for all values of x and affect your whatever analyses you may be doing after estimation.

5.)

```
prob_results <- c("h=0.77"=prob_1, "h=7.7"=prob_2, "h=77"=prob_3)
knitr::kable(prob_results, col.names = c("P(X<92)"))
```

	P(X<92)
h=0.77	0.5297984
h=7.7	0.5438774
h=77	0.3845063

Sum the values of the kde where X<92. This would give you the area under the curve. The probablity seems to drop relatively sharply as h went from 7.7 to 77. (-16%) From that drop, I would say yes probablity depends on the bandwidth.

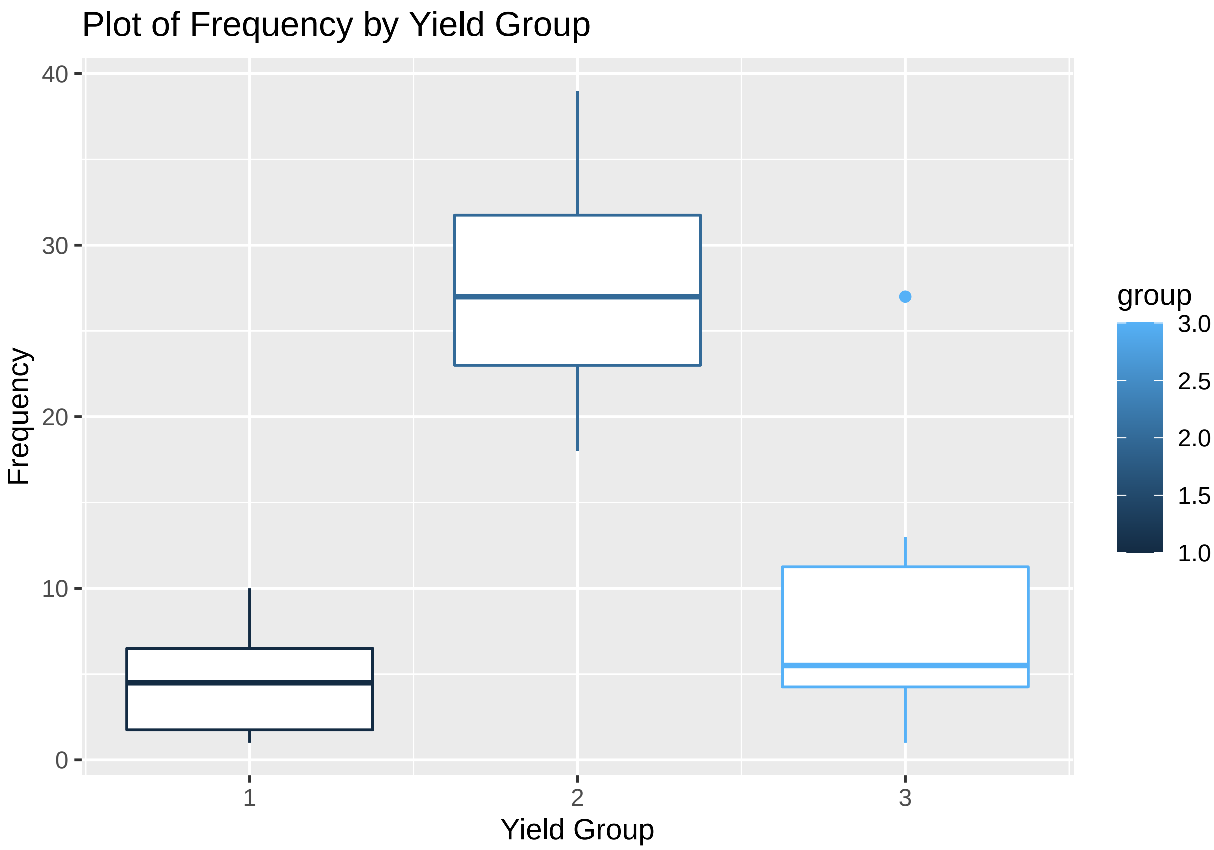
Problem 3

a.)

```
# create dataframe
soybean_df <- data.frame(yield=c(seq(14,48,2)), freq=c(1,4,1,5,7,10,26,22,39,33,28,18,27,13,4,5,6,1))

# tidyverse solution to find mode
sb_mode = soybean_df %>%
  filter(freq==max(freq)) %>%
  select(yield)

# calculate mean
sb_x_bar = with(soybean_df, sum(freq*yield)/sum(freq))

# list all elements out in vector, not just by frequencies
expanded_soybean_vector <- rep(soybean_df$yield, soybean_df$freq)

# find median
sb_median <- median(expanded_soybean_vector)

# dataframe of results
results_df <- data.frame(mean = sb_x_bar, median=sb_median, mode=sb_mode[1,1])

# calculate sample standard deviation
sb_s_bar <- with(soybean_df, sqrt(sum(freq*(yield-sb_x_bar)^2)/(sum(freq)-1)))

# calculate sample variance
sb_variance <- sb_s_bar^2

# calculate coefficient of variation
sb_s_bar/sb_x_bar
```

```
## [1] 0.1968789
```

```
# group data
grouped_soybean_df <- cbind(soybean_df, group=c(rep(1,6), rep(2,6), rep(3,6)))

# create boxplot
ggplot(grouped_soybean_df, aes(x=group, y=freq, group=group, col=group)) +
  geom_boxplot() +
  ggtitle("Plot of Frequency by Yield Group") +
  xlab("Yield Group") +
  ylab("Frequency")
```



The data were split into three equal yield groups (n=6). Group1=Yields 14-24, Group2=Yields 26-36, and Group3 Yields 38-48. Yield sizes in the middle group (group 2) appear to be far more frequent than the other groups

Problem 4

a.)

$$\text{Sample mean for } X:$$
$$X = a + Y$$
$$\text{We know, } \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
$$\implies \bar{X} = \frac{1}{n}\sum_{i=1}^{n} a + Y_i$$
$$= \frac{1}{n}(\sum_{i=1}^{n}(Y_i) + na)$$
$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i) + a$$
$$= \bar{Y} + a \blacksquare$$
$$\text{Sample mean for } U:$$
$$U = b * Z$$
$$\text{Similarly, } \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$$
$$\implies \bar{U} = \frac{1}{n}\sum_{i=1}^{n} bZ_i$$
$$= b\frac{1}{n}\sum_{i=1}^{n} Z_i$$
$$= b\bar{Z} \blacksquare$$

b.)

Sample variance for X :

$$X = a + Y$$

$$\text{We know, } s_Y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

$$\implies s_X^2 = \frac{\sum ((a + y_i) - (\bar{Y} + a))^2}{n-1} \text{ , distribute the negative}$$

$$= \frac{\sum (a + y_i - \bar{Y} - a)^2}{n-1} \text{ , a's cancel}$$

$$= \frac{\sum (y_i - \bar{Y})^2}{n-1} \blacksquare$$

Sample standard deviation for X :

$$\sqrt{s_X^2} = s_X = \sqrt{\frac{\sum (y_i - \bar{Y})^2}{n-1}} \blacksquare$$

Sample variance for U :

$$U = b * Z$$

$$\text{We know, } s_Z^2 = \frac{\sum (z_i - \bar{z})^2}{n-1}$$

$$\implies s_U^2 = \frac{\sum ((bz_i) - (b\bar{z}))^2}{n-1}$$

$$= \frac{\sum (b^2 z_i^2 - 2b^2 z_i \bar{z} + b^2 \bar{z}^2)}{n-1}$$

$$= \frac{b^2 \sum (z_i^2 - 2 z_i \bar{z} + \bar{z}^2)}{n-1}$$

$$= \frac{b^2 \sum (z_i - \bar{Z})^2}{n-1} \blacksquare$$

Sample standard deviation for U :

$$\sqrt{s_U^2} = s_U = \sqrt{\frac{b^2 \sum (z_i - \bar{Z})^2}{n-1}} \blacksquare$$

c.)

The median's position remains unchanged for both X and U

Denote median as M

If n is odd, $M = \frac{n+1}{2} th$ observation

If n is even, $M = \dfrac{\frac{n+1}{2} th\ obs + \frac{n}{2} th\ obs}{2}$

Problem 5

1.)

Prove: $E(\bar{X}) = \mu$

$$\begin{aligned}
E(\bar{X}) &= E(\frac{1}{n}\sum_{i=1}^{n} x_i) \text{ , pull the constant out} \\
&= \frac{1}{n}E(\sum_{i=1}^{n} x_i) \text{ , by linearity we can move E()} \\
&= \frac{1}{n}\sum_{i=1}^{n} E(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \frac{1}{n}n\mu \\
&= \mu \blacksquare
\end{aligned}$$

2.)

Prove: $E(S^2) = \hat{\sigma}^2$

$$\begin{aligned}
E(S^2) &= E(\frac{\sum(x_i - \bar{x})^2}{n-1}) \\
&= \frac{1}{n-1}E(\sum(x_i - \bar{x})^2) \text{ , expanding we get} \\
&= \frac{1}{n-1}E(\sum x_i^2 - 2x_i\bar{x} + \bar{x}^2) \text{ , distribute sigma} \\
&= \frac{1}{n-1}E(\sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2) \\
&= \frac{1}{n-1}E(\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2) \text{ , solving for } \sum x_i \\
&= \frac{1}{n-1}E(\sum x_i^2 - 2\bar{x} * n\bar{x} + n\bar{x}^2) \\
&= \frac{1}{n-1}E(\sum x_i^2 - n\bar{x}^2) \text{ , using linearity} \\
&= \frac{1}{n-1}\sum E(x_i^2) - E(n\bar{x}^2) \text{ , subbing in for known identites we get} \\
&= \frac{1}{n-1}\sum(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2) \text{ , evaluating sigma for the left term} \\
&= \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) \\
&= \frac{1}{n-1}(n\sigma^2 - \sigma^2) \\
&= \frac{1}{n-1}(n-1)\sigma^2 \\
&= \sigma^2 \blacksquare
\end{aligned}$$