APPM 4570/5570 Fall 2017: Statistical Methods Homework 1 Solutions

University of Colorado Boulder - Applied Mathematics Department

Theoretical Questions

1. Problem 1 - 15 Points

   Grading: 3 (a), 3 (b), 3 (c), 3 (d), 3 (e)

   (a) The populations of interest are college students that use Twitter at CU-Boulder campus and college students that use Twitter among other universities.

   (b) The population of CU students is a subset of the other population. Both populations are in an educational environment with access to technology.

   (c) Three possible habits of interest are: the amount of tweets sent per day, the amount of time spent on Twitter, and the amount of media uploaded onto Twitter.

   (d) Yes it will be possible with infinite time and resources.

   (e) A sample can be taken from both populations with the use of statistical inference to infer the characteristics about the entire populations.

2. Problem 2 - 15 Points

   Grading: 3 (a), (b), 3 (c), 3 (d), 3 (e)

   (a) The population are American voters.

   (b) The sample is 1000 American voters.

   (c) The sample frame is the set of American voters with a phone and are available between the hours of 6pm and 8pm.

   (d) The type of sample is every 100th number within the database in numerical order, this is called a systematic sample.

   (e) The variable of interest is whether the person would vote for a Congressional candidate who supports universal health care (yes or no).

3. Problem 3 - 25 Points

   Grading: 6 (a), 4 (b), 8 (c), 7 (d)

   (a) Recall from Calculus 1 that the minimum of an expression can be found by setting the first derivative equal to zero. So, letting $f(c) = \sum_{i=1}^{n}(x_i - c)^2$, we see that

$$\begin{aligned}
\frac{df}{dc} &= -2\sum_{i=1}^{n}(x_i - c) = 0 \\
&\implies \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} c \\
&\implies \sum_{i=1}^{n} x_i = nc \\
&\implies c = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.
\end{aligned}$$

The first derivative test implies that this is a minimum and not a maximum.

(b) We note that part (a) implies that $\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - \mu)^2$, so long as $\bar{x} \neq \mu$.

(c) The variance, s^2, of the x_i dataset is given by

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

Since $y_i = x_i - \bar{x}$, then

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}(\sum_{i=1}^{n} x_i - n\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i) - \bar{x} = \bar{x} - \bar{x} = 0$$

Now, we find the variance of the y_i dataset to be

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Since $\bar{y} = 0$, we get

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

(d) Note that the expression $z_i = \dfrac{x_i - \bar{x}}{s}$ can be expressed as $z_i = \dfrac{y_i}{s}$, and thus $\bar{z} = 0$.
So we find the variance of the z_i dataset to be

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

So it follows that $s_z^2 = s_z = 1$.

Computational Questions

1. Problem 1 - 30 (Undergrad.) / 45 (Grad.) Points

Grading: 2 (a), 5 (b), 4 (c), 4 (d), 5 (e), 3 (f), 2 (g), 5 (h), 6 (i), 4 (j), 5 (k)

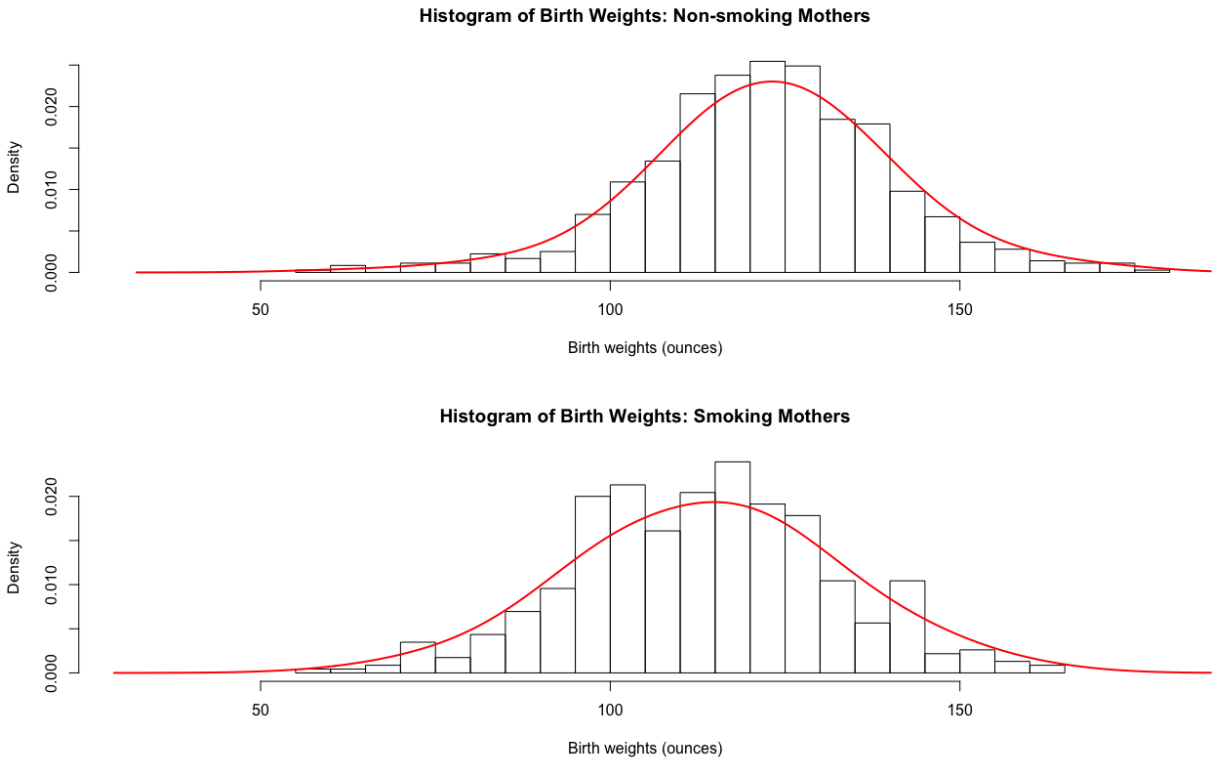
(a) From the website link,

bwt : Birth weight in ounces (999 unknown)
gestation : Length of pregnancy in days (999 unknown)
parity : 0= first born, 9=unknown
age : mother's age in years
height : mother's height in inches (99 unknown)
weight : Mother's prepregnancy weight in pounds (999 unknown)
smoke : Smoking status of mother 0=not now, 1=yes now, 9=unknown

Check code for process of removing unknown values.

```
clean = babies[!(babies$bwt == 999 | babies$gestation == 999 | ...
    babies$height == 99 | babies$weight == 999 | babies$smoke == 9),]
head(clean)
summary(clean)
```

(b)

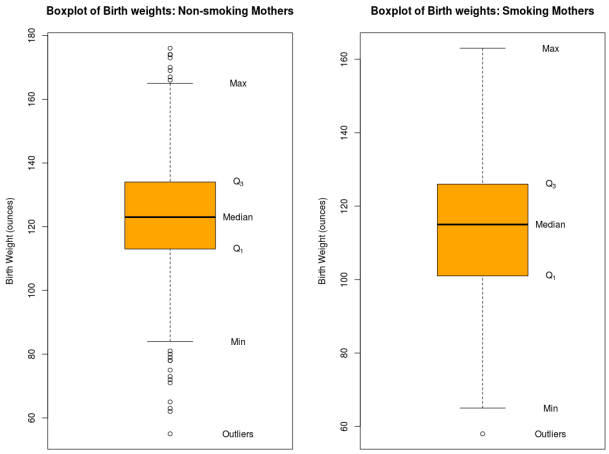



```
#Define variables for ease
x = clean$bwt[clean$smoke ==0]
y = clean$bwt[clean$smoke ==1]
#Create frame
par(mfrow = c(2,1))
#Plot histograms along with their density curves
hist(x, xlim = c(30,180), xlab = "Birth weights (ounces)", main = "...
    Histogram of Birth Weights: Non-smoking Mothers", freq = FALSE, breaks...
     = 20)
lines(density(x, adjust=2), col="red", lwd=2)
hist(y, xlim = c(30,180), xlab = "Birth weights (ounces)", main = "...
    Histogram of Birth Weights: Smoking Mothers", freq = FALSE, breaks = ...
    20)
lines(density(y, adjust=2), col="red", lwd=2)
```

(c) Weights seem to be roughly the same, with perhaps more heavier babies from non-smoking mothers. The histograms appear to be symmetric. The first histogram is unimodal with a mean of 123.08 and a standard deviation of 17.42. The second histogram is bimodal with a mean of 113.82 and a standard deviation of 18.27.

(d) The mean weight difference between babies of smokers and non-smokers is 9.261402 ounces. Using the mean as a measure of center to compare birth weights can be inaccurate if there are outliers.

(e)



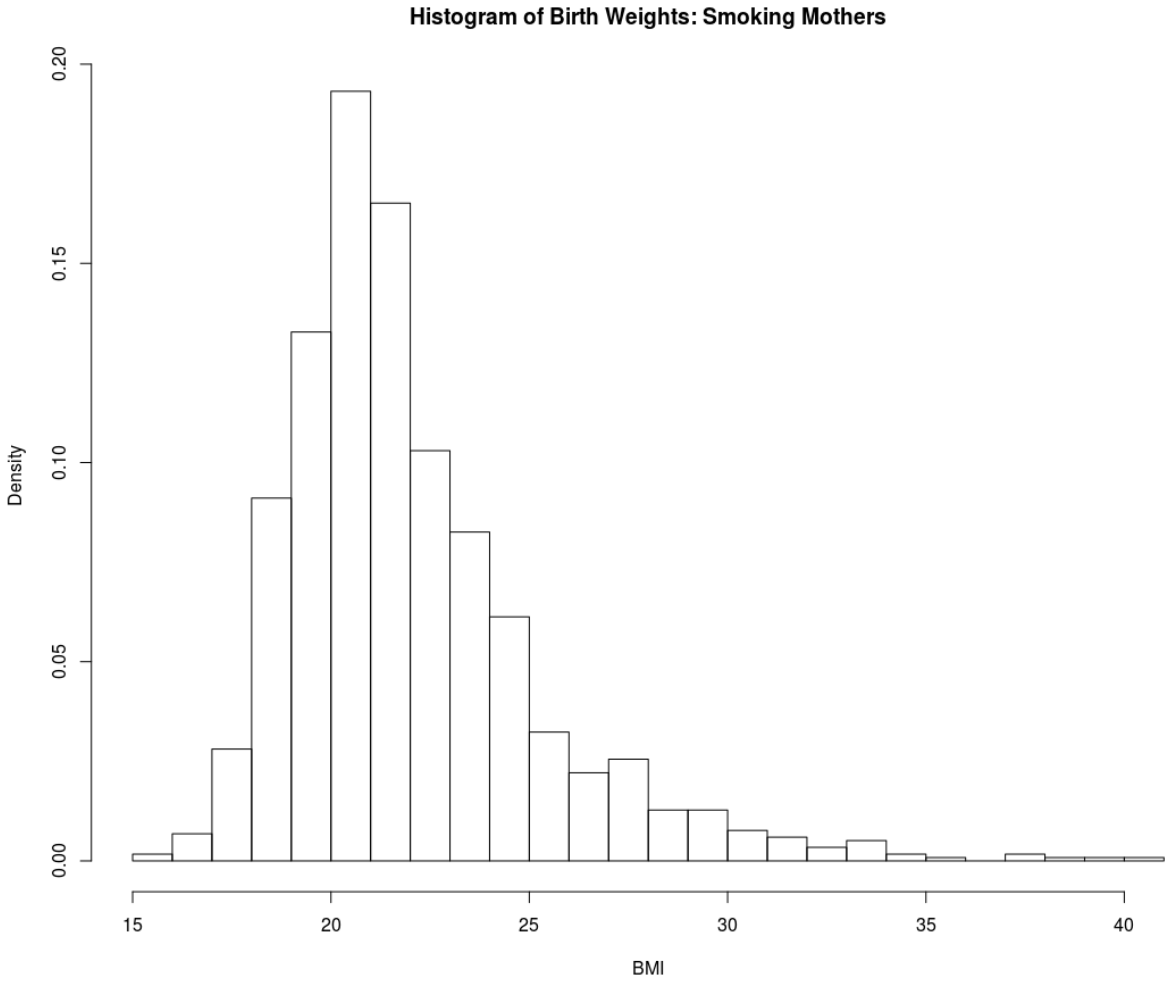

```
par(mfrow = c(1,2))
box=boxplot(x,main="Boxplot of Birth weights: Non-smoking Mothers",ylab="...
    Birth Weight (ounces)",col="orange")
text(x=rep(1.3,6),y=c(box$stats[,1],min(box$out)) ,labels=c("Min",...
    expression(Q[1]),"Median",expression(Q[3]),"Max","Outliers"))
box=boxplot(y,main="Boxplot of Birth weights: Smoking Mothers",ylab="Birth...
     Weight (ounces)",col="orange")
text(x=rep(1.3,6),y=c(box$stats[,1],mean(box$out)) ,labels=c("Min",...
    expression(Q[1]),"Median",expression(Q[3]),"Max","Outliers"))
```

(f) The median weight difference between babies who are firstborn and those who are not is 2 ounces.

(g) Check code

```
#BMI = kg/m^2 , 0.0254 m = 1 in, 0.45359 kg = 1 lb
clean$weight = clean$weight*0.45359
clean$height = clean$height*0.0254
clean$BMI = (clean$weight)/((clean$height)^2)
cleannp <- data.frame(clean$weight,clean$height,clean$BMI)
```

(h) The figure below is a histogram of the BMI for each mother. Looking at this, it can be seen that the distribution of BMI is positively skewed.




```
hist(clean$BMI, xlab = "BMI", main = "Histogram of Birth Weights: Smoking ...
    Mothers", freq = FALSE, breaks = 20)
```

Min	1st Quartile	Median	3rd Quartile	Max
15.66	19.94	21.28	23.36	40.35

Table 1: BMI Quartiles of Mothers

	Mothers (No Smoke)			
Quartiles	Min - 1st Qu.	1st Qu. - Median	Median - 3rd Qu.	3rd Qu. - Max
Mean (oz.)	121.7975	123.8359	120.5152	124.4869
Standard Deviation (oz.)	15.43411	17.17968	17.71565	19.04755
Median (oz.)	121	125	120	125
IQR	20	19.75	20.5	23

Table 2: Summary of Each Quartile of Baby Birth Weights for Mothers Who Do Not Smoke

(i)
```
### Birth weight of babies from mothers who do not smoke: x
### BMI of mothers who do not smoke: a
#Quartile Values from summary(clean$BMI): 1st Qu. (19.94)  Median (22.00) ...
    Mean 3rd Qu. (23.36)
```

```
a = clean$BMI[clean$smoke ==0]
meanqu_x = c(mean(x[a< 19.94]), mean(x[a> 19.94 & a < 22]), mean(x[a > 22 & a < 23.36]), mean(x[a> 23.36]))
sdqu_x = c(sd(x[a< 19.94]), sd(x[a> 19.94 & a < 22]), sd(x[a > 22 & a < 23.36]), sd(x[a> 23.36]))
medianqu_x = c(median(x[a< 19.94]), median(x[a> 19.94 & a < 22]), median(x[a > 22 & a < 23.36]), median(x[a> 23.36]))
IQR_x = c(IQR(x[a < 19.94]), IQR(x[a > 19.94 & a < 22]), IQR(x[a > 22 & a < 23.36]), IQR(x[a> 23.36]))
```

Looking from the table above, it can be seen that the birth weight within the group that does not smoke is relatively symmetric. From the mean and median values, each quartile have similar values with each other. This suggests that the entire data set is centered around these values, and since each quartile share relatively similar values, the distribution is symmetric.

	Mothers (Smoke)			
Quartiles	Min - 1st Qu.	1st Qu. - Median	Median - 3rd Qu.	3rd Qu. - Max
Mean (oz.)	123.0141	121.4585	124.6207	124.8951
Standard Deviation (oz.)	17.72922	17.77291	17.54876	16.27237
Median (oz.)	123	122	124	124
IQR	21	21	24	18.75

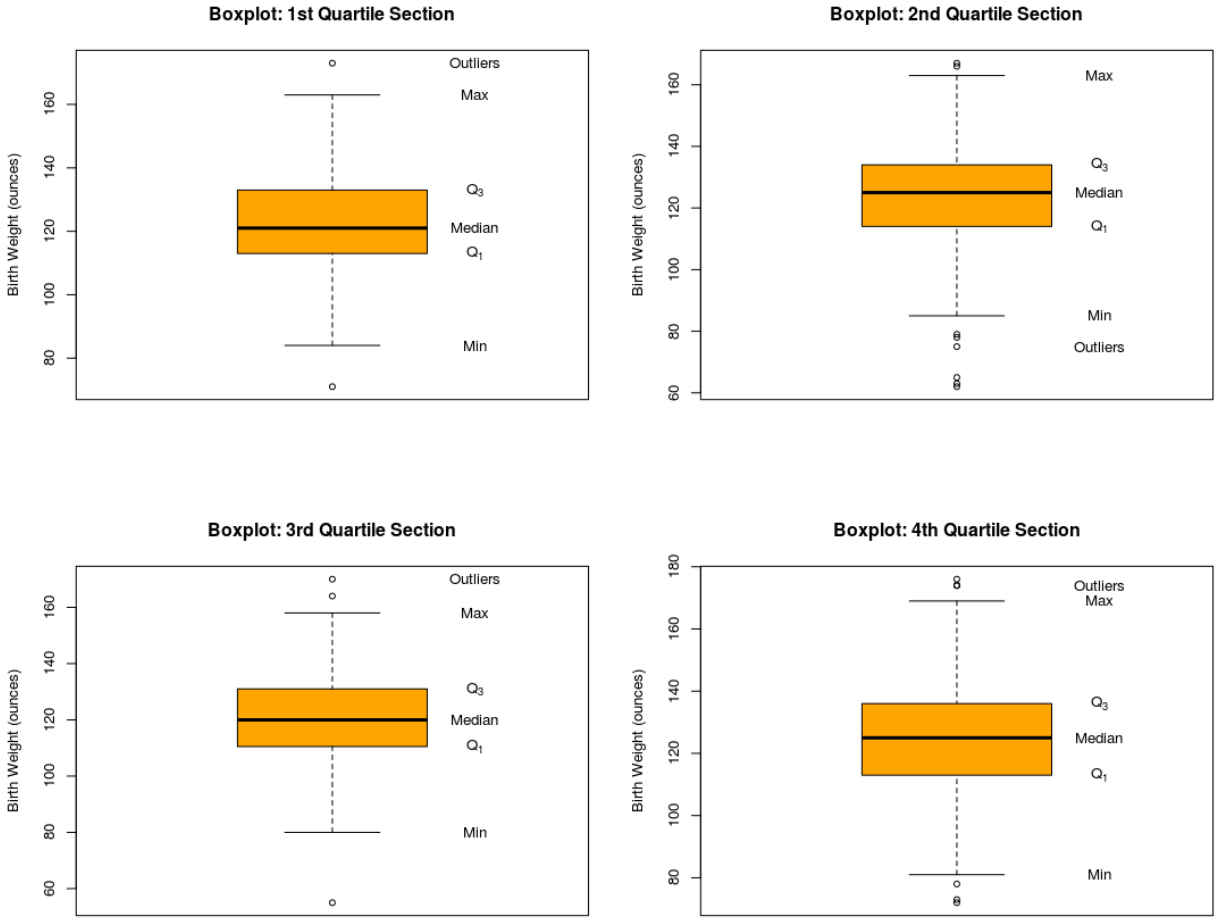
Table 3: Summary of Each Quartile of Baby Birth Weights for Mothers Who Do Smoke

```
### Birth weight of babies from mothers who do not smoke: y
### BMI of mothers who do smoke: b
#Quartile Values from summary(clean$BMI): 1st Qu. (19.94)  Median (22.00) ... Mean 3rd Qu. (23.36)
b = clean$BMI[clean$smoke ==1]
meanqu_y = c(mean(x[b < 19.94]), mean(x[b > 19.94 & b < 22]), mean(x[b > 22 & b < 23.36]), mean(x[b > 23.36]))
sdqu_y = c(sd(x[b < 19.94]), sd(x[b > 19.94 & b < 22]), sd(x[b > 22 & b < 23.36]), sd(x[b> 23.36]))
medianqu_y = c(median(x[b < 19.94]), median(x[b > 19.94 & b < 22]), median(x[b > 22 & b < 23.36]), median(x[b > 23.36]))
IQR_y = c(IQR(x[b < 19.94]), IQR(x[b > 19.94 & b < 22]), IQR(x[b > 22 & b < 23.36]), IQR(x[b > 23.36]))
```

Looking from the table above, it can be seen that the birth weight within the group that does smoke is relatively symmetric. From the mean and median values, each quartile have similar values with each other. This suggests that the entire data set is centered around these values, and since each quartile share relatively similar values, the distribution is symmetric.

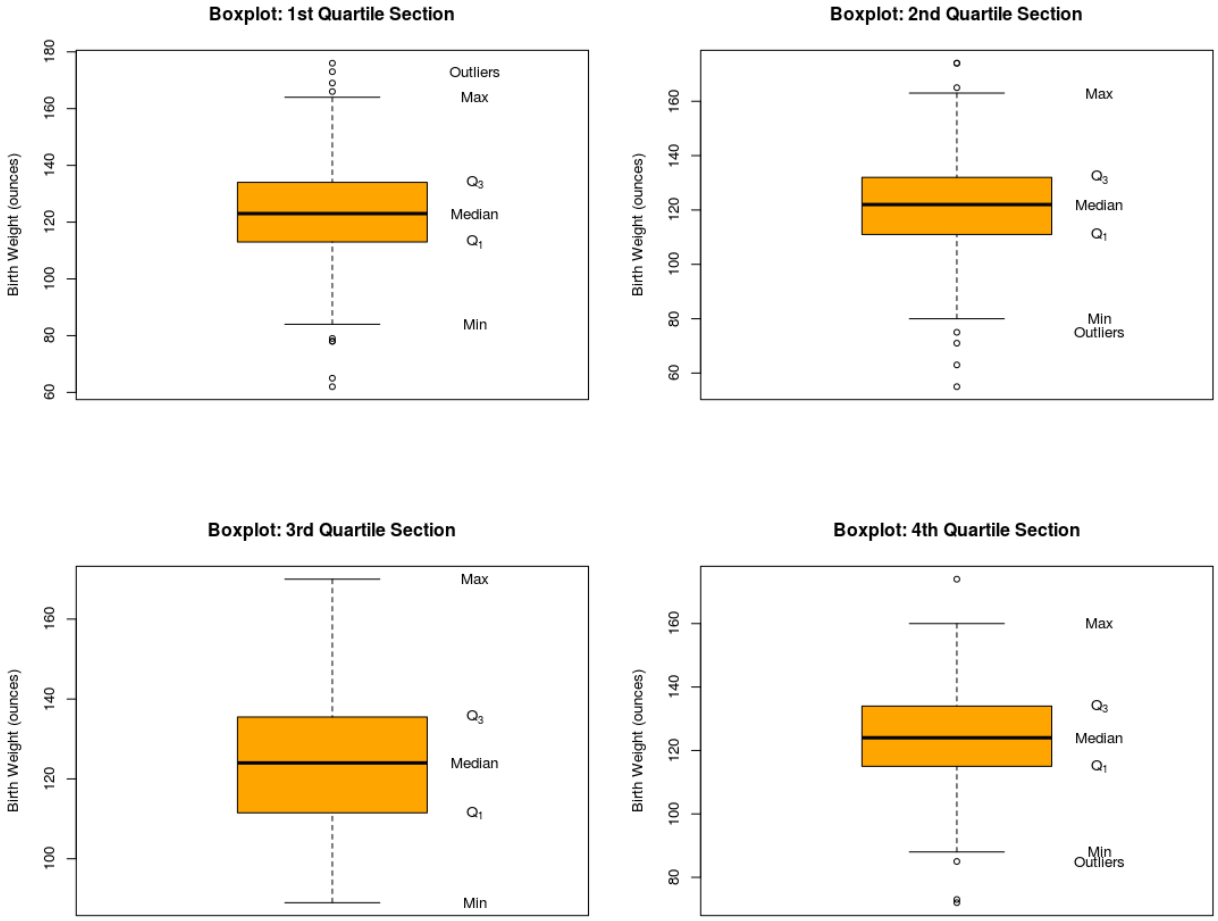
(j)

Figure 1. Boxplots of bwt of Mothers Who Do Not Smoke Conditioned on BMI Quartiles




```
par(mfrow = c(2,2))
box=boxplot(x[a< 19.94],main="Boxplot: 1st Quartile Section",ylab="Birth ...
    Weight (ounces)",col="orange")
text(x=rep(1.3,6),y=c(box$stats[,1],box$out[1]) ,labels=c("Min",expression...
    (Q[1]),"Median",expression(Q[3]),"Max","Outliers"))
box=boxplot(x[a> 19.94 & a < 22],main="Boxplot: 2nd Quartile Section",ylab...
    ="Birth Weight (ounces)",col="orange")
text(x=rep(1.3,6),y=c(box$stats[,1],box$out[1]) ,labels=c("Min",expression...
    (Q[1]),"Median",expression(Q[3]),"Max","Outliers"))
box=boxplot(x[a > 22 & a < 23.36],main="Boxplot: 3rd Quartile Section",...
    ylab="Birth Weight (ounces)",col="orange")
text(x=rep(1.3,6),y=c(box$stats[,1],box$out[1]) ,labels=c("Min",expression...
    (Q[1]),"Median",expression(Q[3]),"Max","Outliers"))
box=boxplot(x[a> 23.36],main="Boxplot: 4th Quartile Section",ylab="Birth ...
    Weight (ounces)",col="orange")
text(x=rep(1.3,6),y=c(box$stats[,1],box$out[1]) ,labels=c("Min",expression...
    (Q[1]),"Median",expression(Q[3]),"Max","Outliers"))
```

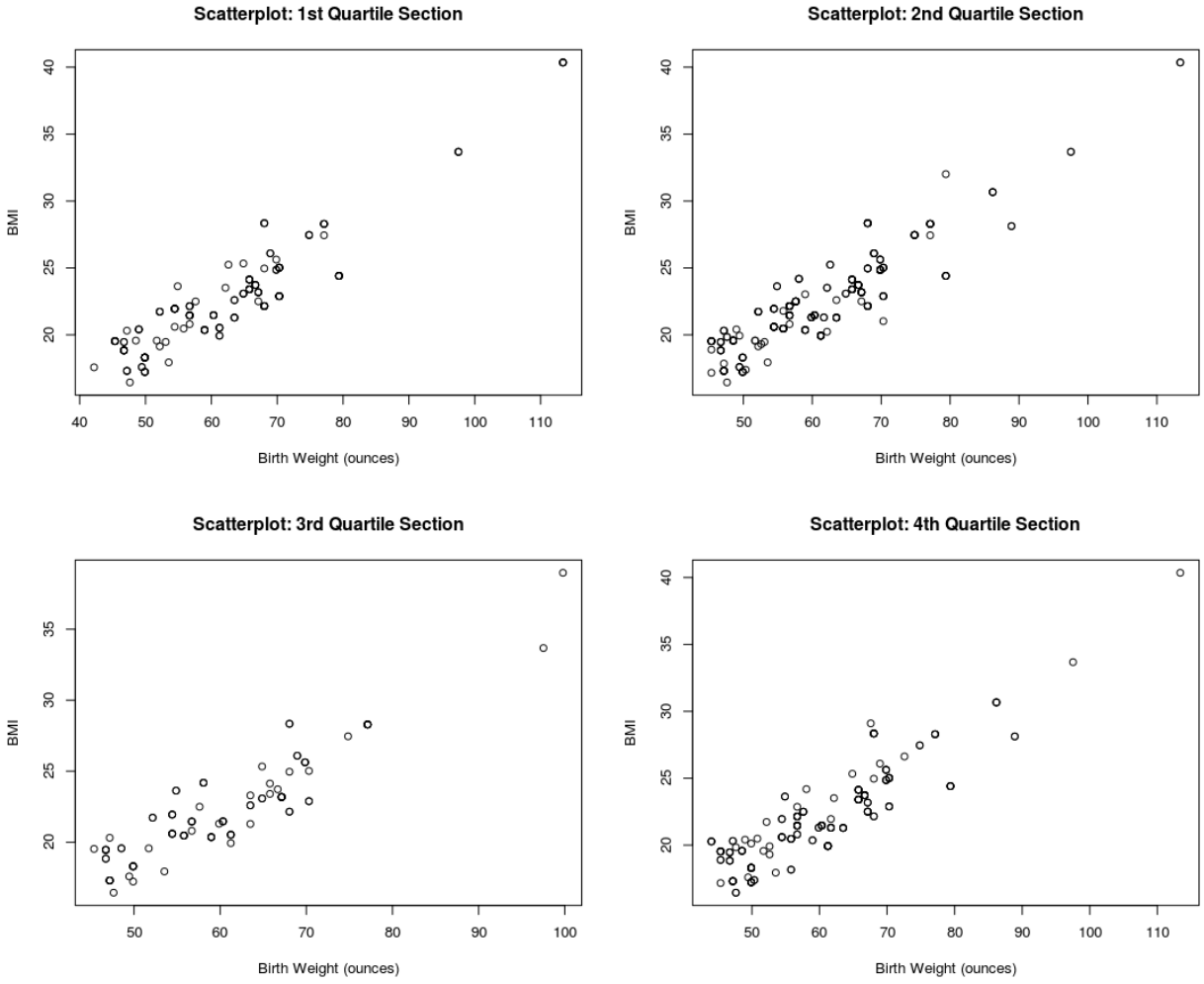
Figure 2. Boxplots of bwt of Mothers Who Do Smoke Conditioned on BMI Quartiles




```
1 par(mfrow = c(2,2))
2 box=boxplot(x[b< 19.94],main="Boxplot: 1st Quartile Section",ylab="Birth ...
      Weight (ounces)",col="orange")
3 text(x=rep(1.3,6),y=c(box$stats[,1],box$out[1]) ,labels=c("Min",expression...
      (Q[1]),"Median",expression(Q[3]),"Max","Outliers"))
4 box=boxplot(x[b> 19.94 & b < 22],main="Boxplot: 2nd Quartile Section",ylab...
      ="Birth Weight (ounces)",col="orange")
5 text(x=rep(1.3,6),y=c(box$stats[,1],box$out[1]) ,labels=c("Min",expression...
      (Q[1]),"Median",expression(Q[3]),"Max","Outliers"))
6 box=boxplot(x[b > 22 & b < 23.36],main="Boxplot: 3rd Quartile Section",...
      ylab="Birth Weight (ounces)",col="orange")
7 text(x=rep(1.3,6),y=c(box$stats[,1],box$out[1]) ,labels=c("Min",expression...
      (Q[1]),"Median",expression(Q[3]),"Max","Outliers"))
8 box=boxplot(x[b> 23.36],main="Boxplot: 4th Quartile Section",ylab="Birth ...
      Weight (ounces)",col="orange")
9 text(x=rep(1.3,6),y=c(box$stats[,1],box$out[1]) ,labels=c("Min",expression...
      (Q[1]),"Median",expression(Q[3]),"Max","Outliers"))
```

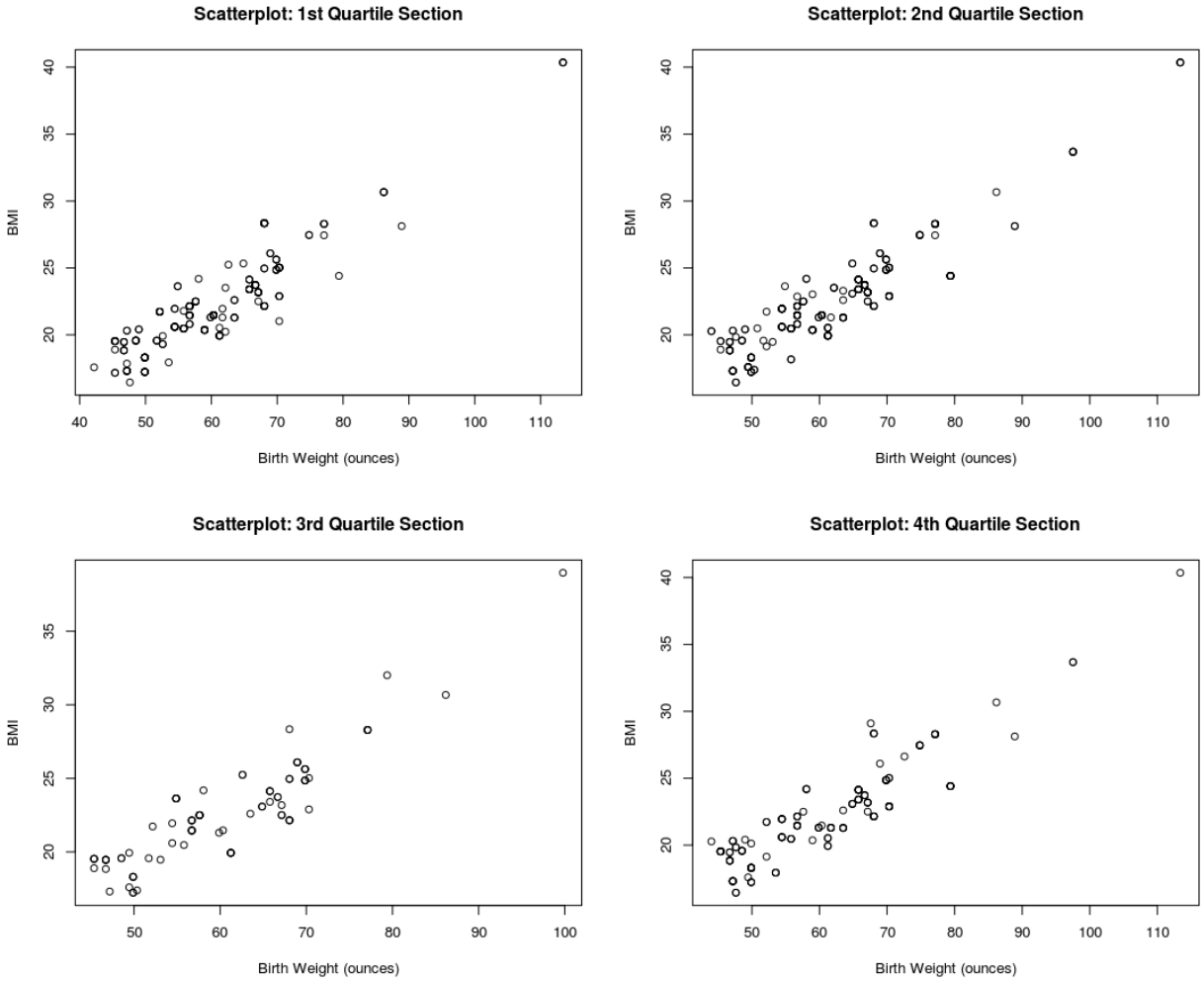
The boxplots above verify the conclusions from the previous part; such that the data is relatively symmetric for both smoking conditions. Each quartile relatively shares the same median and have symmetry about the median.

Figure 3. Scatterplots of bwt vs. BMI of Mothers Who Do Not Smoke Conditioned on BMI Quartiles



(k)
```
par(mfrow = c(2,2))
plot(clean$weight[x[a< 19.94]] , clean$BMI[x[a < 19.94]], xlab = "Birth ...
    Weight (ounces)", main = "Scatterplot: 1st Quartile Section", ylab = "...
    BMI")
plot(clean$weight[x[a> 19.94 & a < 22]] , clean$BMI[x[a > 19.94 & a < ...
    22]], xlab = "Birth Weight (ounces)", main = "Scatterplot: 2nd ...
    Quartile Section", ylab = "BMI")
plot(clean$weight[x[a > 22 & a < 23.36]] , clean$BMI[x[a > 22 & a < 23.36...
    ]], xlab = "Birth Weight (ounces)", main = "Scatterplot: 3rd Quartile ...
    Section", ylab = "BMI")
plot(clean$weight[x[a> 23.36]] , clean$BMI[x[a > 23.36]], xlab = "Birth ...
    Weight (ounces)", main = "Scatterplot: 4th Quartile Section", ylab = "...
    BMI")
```

Figure 4. Scatterplots of bwt vs. BMI of Mothers Who Do Smoke Conditioned on BMI Quartiles




```
1 par(mfrow = c(2,2))
2 plot(clean$weight[x[b < 19.94]] , clean$BMI[x[b < 19.94]], xlab = "Birth ...
      Weight (ounces)", main = "Scatterplot: 1st Quartile Section", ylab = "...
      BMI")
3 plot(clean$weight[x[b > 19.94 & b < 22]] , clean$BMI[x[b > 19.94 & b < ...
      22]], xlab = "Birth Weight (ounces)", main = "Scatterplot: 2nd ...
      Quartile Section", ylab = "BMI")
4 plot(clean$weight[x[b > 22 & b < 23.36]] , clean$BMI[x[b > 22 & b < 23.36...
      ]], xlab = "Birth Weight (ounces)", main = "Scatterplot: 3rd Quartile ...
      Section", ylab = "BMI")
5 plot(clean$weight[x[b > 23.36]] , clean$BMI[x[b > 23.36]], xlab = "Birth ...
      Weight (ounces)", main = "Scatterplot: 4th Quartile Section", ylab = "...
      BMI")
```

The scatterplots above all yield linear relationships.

2. Problem 2 - 15 Points Grading: 10 (Code) , 5 (Results)

There are many ways to generate 25 data points in R for this simulation. One way would be to use the "runif" function to populate x_i, then use that to perform all further calculations. This might look like:

```
    #Generate random sample of data
xi=runif(25,1,500)

#Calculate mean
xbar=(1/25)*sum(xi)

#Calculate other two variables
yi=(xi-xbar)
zi=((xi-xbar)/sd(xi))

#Calculate statistical values
var(xi)
var(yi)
sd(xi)
sd(yi)
var(zi)
sd(zi)
```

This creates the random sample of 25 objects, calculates the mean of the sample, and defines y_i and z_i as before. Then, when we print the variances and standard deviations we find that:

$$\begin{aligned} s_x &= s_y \\ s_x^2 &= s_y^2 \\ s_z &= s_z^2 = 1 \end{aligned}$$

Which confirms our previous results.