S520 Homework 5

Enrique Areyan
February 24, 2012

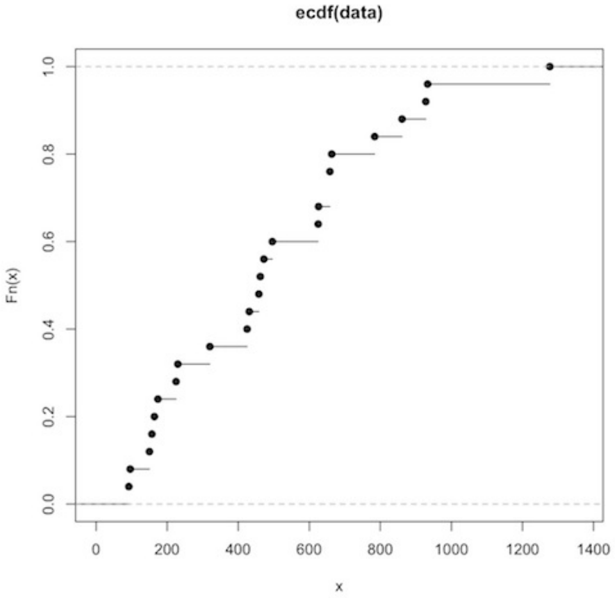
7.7.#1:

(a) Plugging in R the data:

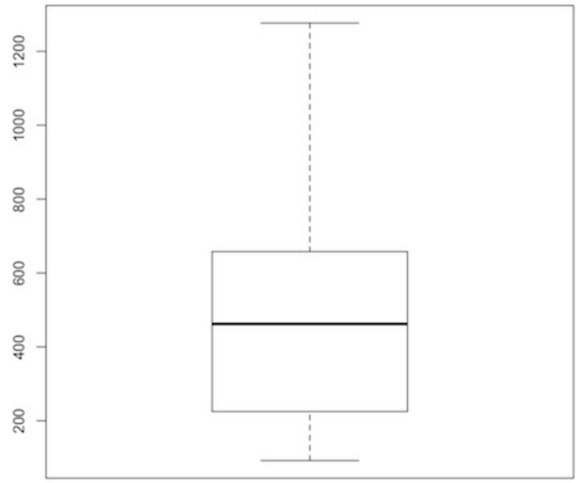
data ← c(462,425,164,784,625,472,658,658,663,928,92,230,
96,626, 1277,225,150,320,496,157,458,933,861,174,431)

Then, $Fn \leftarrow ecdf(data)$. The following are the graphs for this distribution :

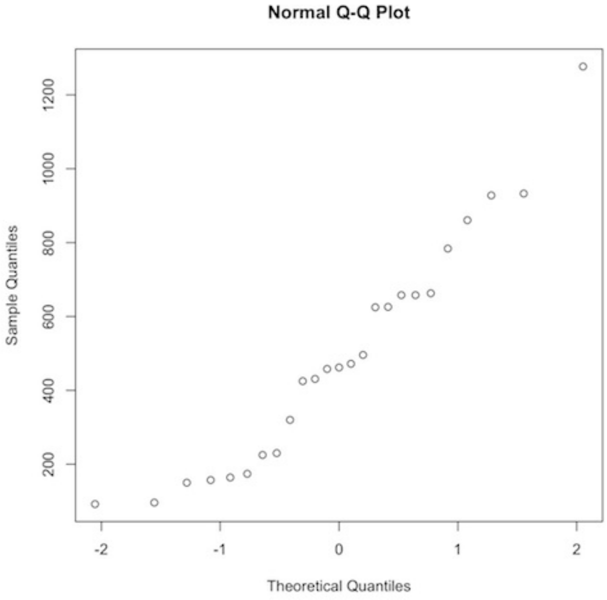
(a) $plot(Fn)$



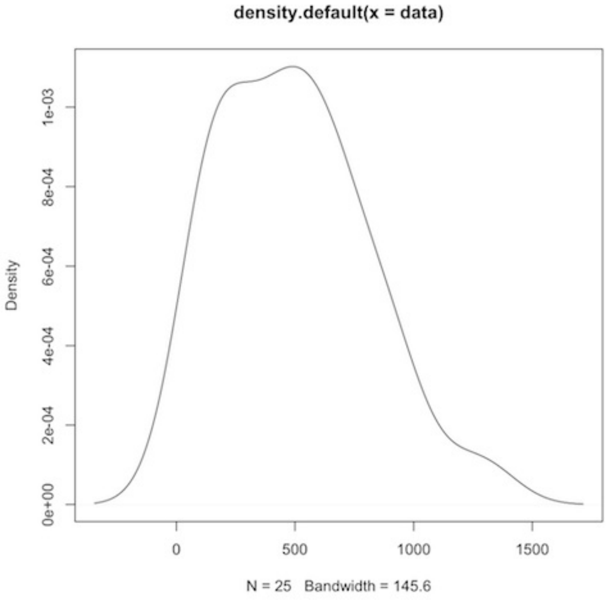
(e)$boxplot(data)$:



(f) $qqnorm(data)$:



(g)$plot(density(data))$:



(b) $mean(data) = 494.6$ and $var(data) = 94873.67$

(c) $median(data) = 462$ and $quantile(Fn)$:

0%	25%	50%	75%	100%
92	225	462	658	1277

So that $iqr \leftarrow 658 - 225 = 433$

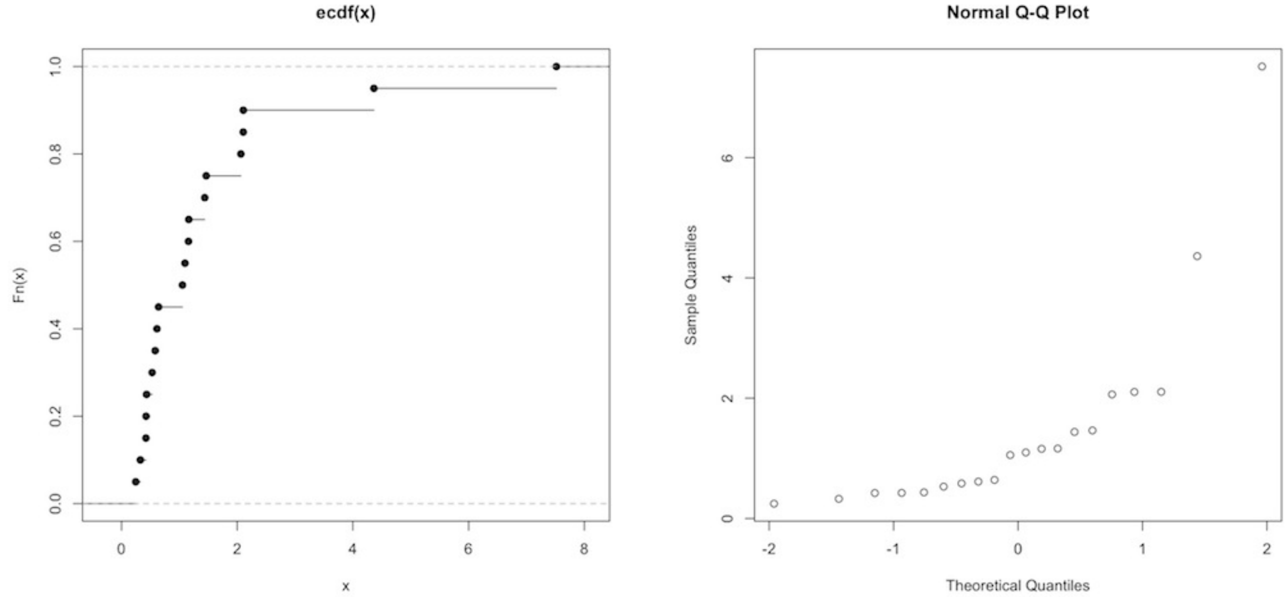
(d) $iqr/sqrt(var(data)) = 1.405773$

(h) It is plausible that the data was obtained from a normal distribution. The evidence supporting this is clear from the graphs as well as the iqr to stdev ratio.

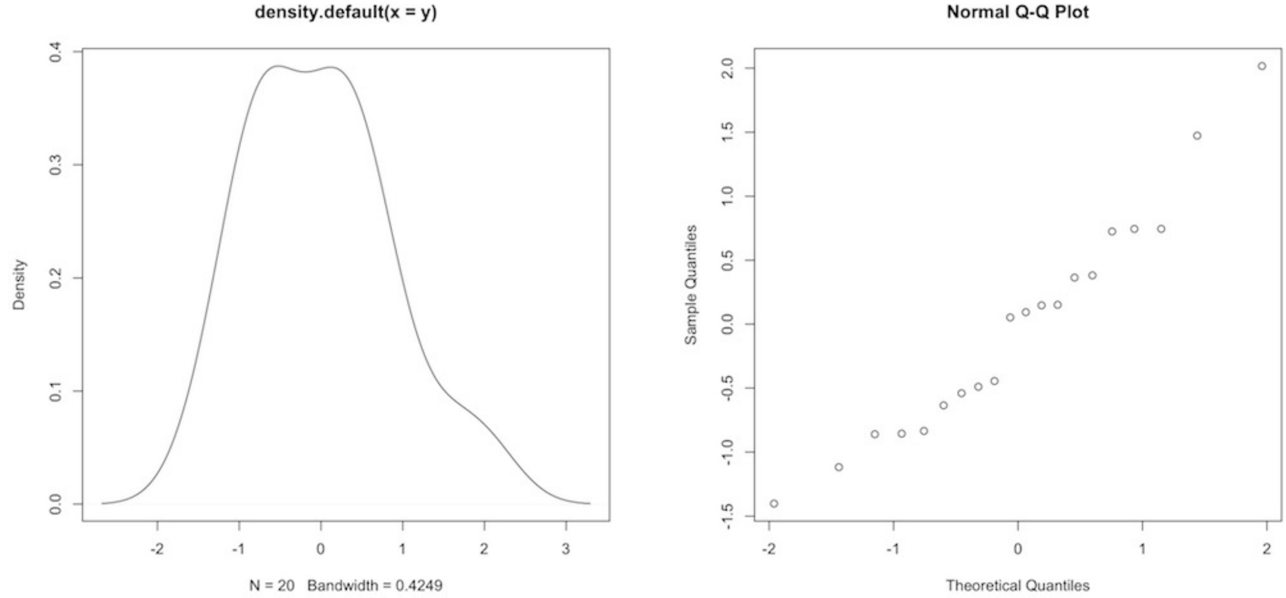
7.7.#4:

(a) Doing the same procedure as before but with the new data set:

For x



For y = log(x)



(b) $mean(data) = 1.4876$, $var(data) = 2.934267$, $median(data) = 1.076$ and $quantile(Fn)$:

0%	25%	50%	75%	100%
0.24600	0.50600	1.07600	1.61375	7.51700

So that $iqr \leftarrow 1.61375 - 0.50600 = 1.10775$

(c) $iqr/sqrt(var(data)) = 0.6466837$. The data does not seem to be drawn from a normal distribution because the iqr to stdev ratio does not conform to that of a normal distribution. Also, if we plot the kernel density estimate of x, as well as a box plot, it does not look remotely close to a normal distribution.

(d) The graph for $qqnorm(data)$ is on (a) above.
The normal probability plot does not conform to a normal distribution. Only looking at this piece of information, this data does not seem to be drawn from a normal distribution.

(e) Looking at the distribution of the transformed data, it does seem plausible that $\vec{y}$ was drawn from a normal distribution. The evidence in support of this conclusion is: (i) the kernel density plot on (a) (ii) the normal plot also on (a) and (iii) the iqr to stdev ratio. The quantiles of $\vec{y}$ are:

0%	25%	50%	75%	100%
-1.4024237	-0.6848364	0.0730414	0.4669196	2.0171671

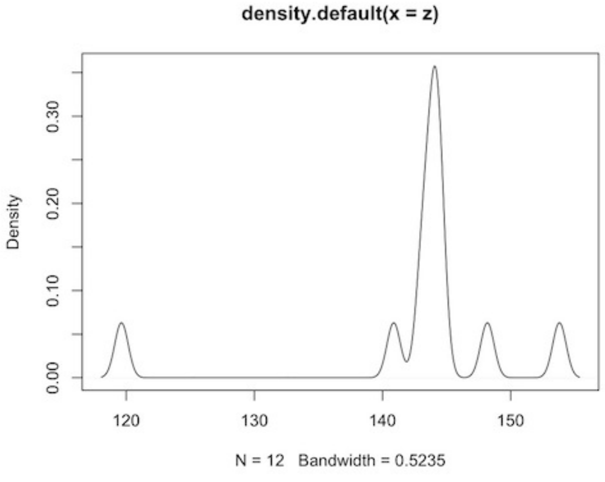
So that $iqr \leftarrow 0.4669196 + 0.6848364 = 1.151756$ and thus, $iqr/sqrt(var(data)) = 1.308589$

7.7.#5:

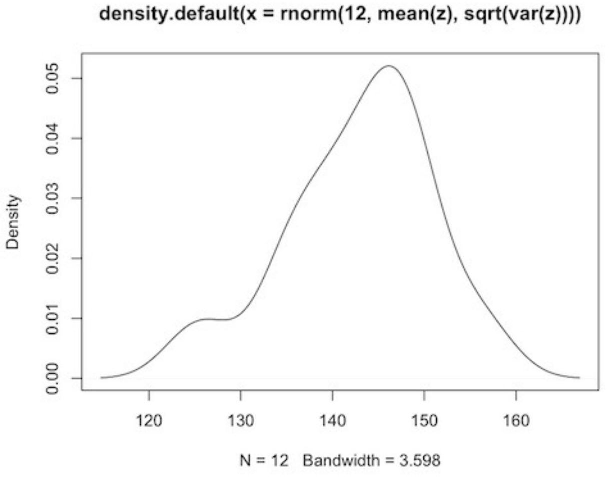
(a) I used the following command in R to analyze the data: z <- c(143 + (3/16) , 144 + (4/16), 140 + (14/16), 144 + (7/16), 143 + (12/16), 153 + (13/16), 119 + (10/16) , 143 + (1/16) , 143 + (14/16) , 144 + (3/16), 144 + (7/16) , 148 + (3/16)).

I am inclined to state that the measurements do not appear to be a sample from a normal distribution. This statement follows from a comparison between the data and a normal distribution with mean $\mu = 142.8073$, i.e., the mean of the measurements and standard deviation $\sigma = 7.992915$. The iqr of these measures is 1.2813, whereas the iqr of the theoretical normal distribution is approximately 10.7822, very far from that of the measurements. The iqr to stdev ratio is 0.1602982 for the data and 1.34897 for the theoretical distribution. Both of these numbers were compared not only with the theoretical distribution, but with concrete sample of size 12 from $Normal(142.8073, 7.992915)$. In both cases the numbers were very different.
The following graphs also support this reasoning: **Note**: although the graphs from the sample of the theoretical normal distribution may vary widely, the graphs in this document are just a reference and are not to be tought as the only possible graphs from this distribution. However, by looking at these graphs and other graphs which I do not add to the document because of space constrains, the difference between the measurements and the normal distribution is obvious.)

Data

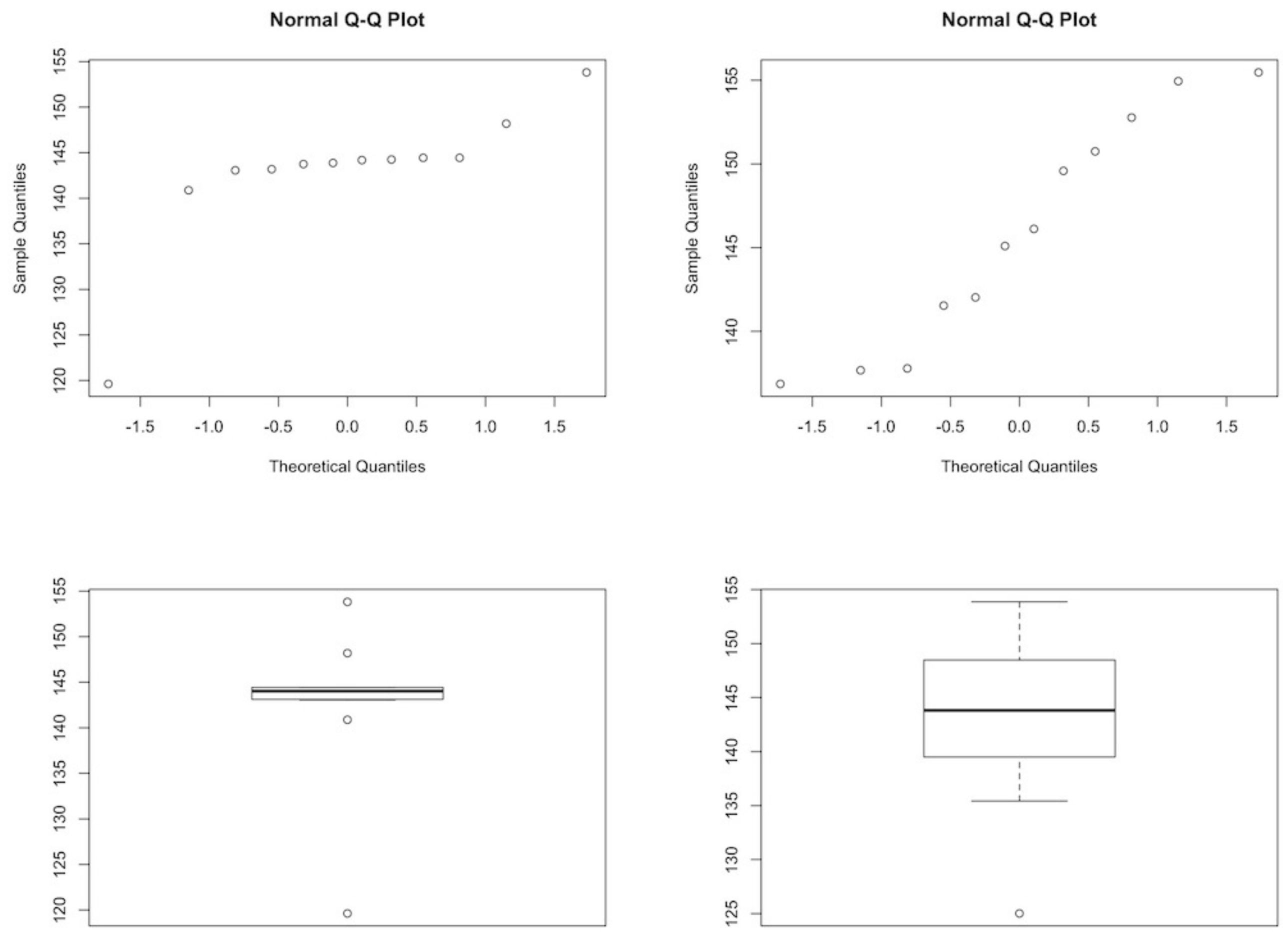


sample of size 12 from $Normal(142.8073, 7.992915)$



Data

sample of size 12 from $Normal(142.8073, 7.992915)$



(b) The variability in these measurements may be attributed to the lack of an standardized procedure used by the students to obtain the measurements. It may be the case that students used procedures that results in very different results as show in this data set. Also, it can be the case that the instrument used to obtain the measurements were very different among the students. Most likely, a combination of this two factors (procedure and instrument) lead to this amount of variability in the data.

(c) The true length of the table may lie between 143 and 145. Specifically, we can use the median to give a better estimate $median(z) = 144.031$. There are not too many points in this data set, and one can easily discard the more extreme values such as $119\frac{10}{16}$, $148\frac{3}{16}$ and $153\frac{13}{16}$. Just by inspecting the data set we can see that 9 out of 12 estimates are between 143 and 145. Finally, I believe the median is a better estimator than the mean ($mean(z) = 142.8073$) as it is not affected as much by outliers. In this particular case the mean drops to 142.8073 on account of one outlier $119\frac{10}{16}$. It is interesting to denote that by removing these three outliers, the distribution becomes a lot more like a normal distribution than it would otherwise.

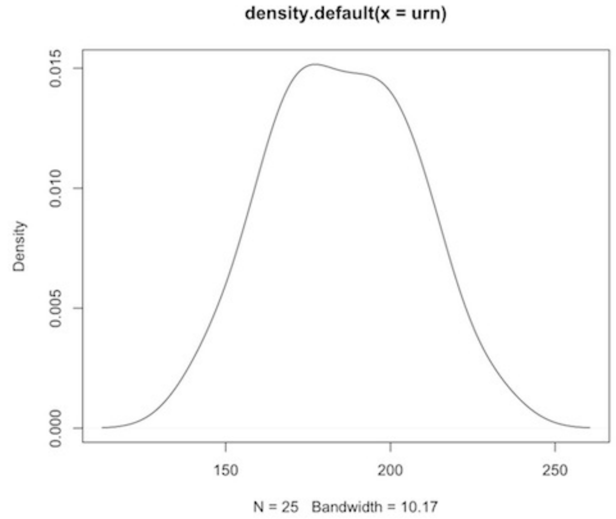
7.7.#7:

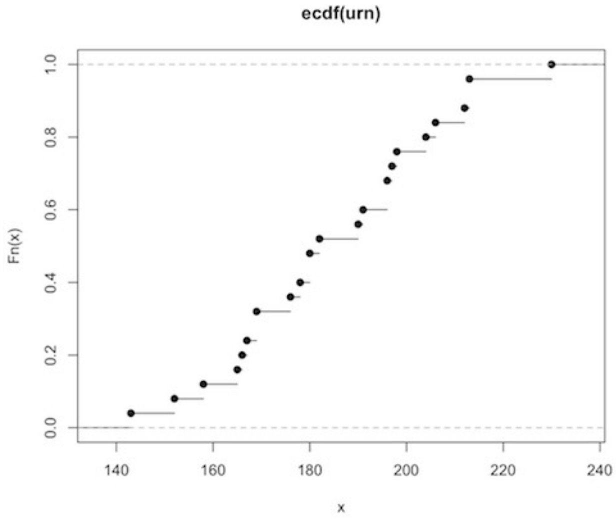
(a)
```
urn.model <- function (){
urn <- c(1,1,1,1,2,5,5,10,10,10)
return <- 0
for(i in 1:40){
draw <- sample(urn,1,TRUE)
return <- return + draw
}
return (return)
}
```

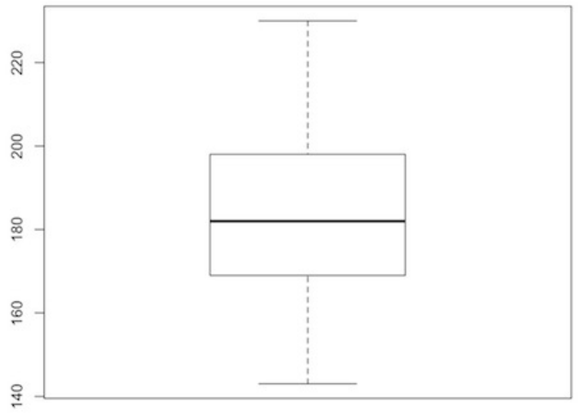
```
urn.random.var <- function(n){
result <- NULL
for(i in 1:n){
result <- c(result,urn.model())
}
return (result)
}

urn <- urn.random.var(25)
```

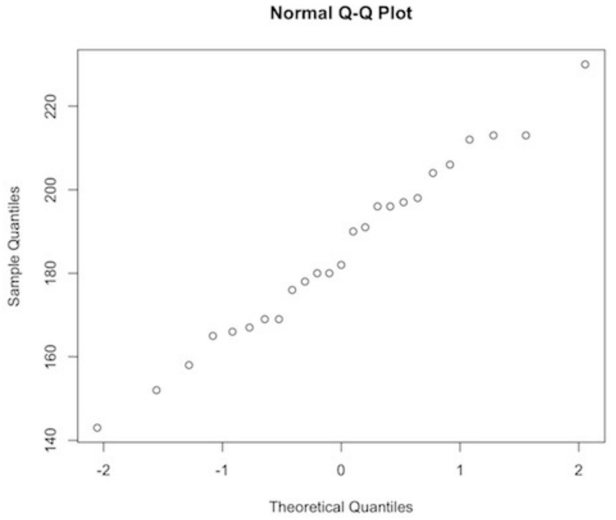
(b) It does appear that the distribution of Y can be approximated by a normal distribution. The following graphs and calculations support such conclusion:

$$mean(urn) = 185.24, \quad var(urn) = 462.7733, \quad iqr = 29, \quad \text{and } iqr/stdev = 1.348074$$









8.4.#3:

(a) i. Expected value of X:

$$EX_i = \sum_{x \in X(S)} xf(x) = 1 \cdot f(1) + 3 \cdot f(3) + 4 \cdot f(4) + 6 \cdot f(6)$$
$$= 1 \cdot 0.1 + 3 \cdot 0.4 + 4 \cdot 0.4 + 6 \cdot 0.1 = 0.1 + 1.2 + 1.6 + 0.6 = 3.5$$

ii. Variance of X:

$$VarX_i = EX_i^2 - (EX_i)^2 = 1^2 \cdot f(1) + 3^2 \cdot f(3) + 4^2 \cdot f(4) + 6^2 \cdot f(6) - 3,5^2$$

$$= 0.1 + 3.6 + 6.4 + 3.6 - 12.25 = 13.7 - 12.25 = 1.45$$

(b)

$$\begin{array}{rcll}
P(\bar{X}_{100} > 3.6) & = & P(\frac{\bar{X}_{100}-\mu}{\sigma/\sqrt{100}} > \frac{3.6-\mu}{\sigma/\sqrt{100}}) & \text{Substracting } \mu \text{ and dividing by } \frac{\sigma}{\sqrt{100}} \\
 & = & P(Z_n > \frac{3.6-3.5}{\sqrt{1.45}/10}) & \text{Plugging in the numbers} \\
 & = & P(Z_n > \frac{0.1}{0.1204159}) & \text{Performing the operations} \\
 & = & P(Z_n > 0.8304548) & \text{Performing the operations} \\
 & \approx & P(Z > 0.8304548) & \text{Central Limit Theorem } (Z \sim Normal(0,1)) \\
 & = & 1 - P(Z \leq 0.8304548) & \text{Elementary properties of probabilities} \\
 & = & 1 - pnorm(0.8304548) & \text{Using R} \\
 & = & 0.2031 &
\end{array}$$

8.4.#4: Let X_i = hours that two AAAA batteries will power the pointer and let $S_n = \sum_{i=1}^{n} X_i$. From the data we know that $n = 20$, $EX_i = \mu = 5$ hours and $\sigma = 0.5$ hours. We want to find out the following probability:

$$P(S_{20} = \sum_{i=1}^{20} X_i \geq 105)$$

By the CLT:

$$\sum_{i=1}^{n} X_i \approx Normal(n\mu, n\sigma^2)$$

Thus,

$$\begin{array}{rcll}
P(S_{20} \geq 105) & = & 1 - P(S_{20} \leq 105) & \text{Fundamental properties of probabilities} \\
 & \approx & 1 - pnorm(105, 100, \sqrt{5}) & \text{By the CLT} \\
 & = & 0.01267 &
\end{array}$$

8.4.#5:

(a) Her reasoning is supported by the Weak Law of Large Numbers. By calculating $P(170.5 < Y < 199.5)$ as the proportions or counts observed, i.e., $P(170.5 < Y < 199.5) = \frac{\delta}{n}$, where $\delta = 1$ if $y_i \in (170.5, 199.5)$ and 0 otherwise; the student is interpreting the probabilities as the observed frequency of the event $A = \{y_i \in (170.5, 199.5)\}$. In the long run, this proportion should match the true proportion or probability.

(b) I agree. These calculations follow directly from the CLT with $n = 40$, $Y = \sum_{i=1}^{40} X_i$
$\mu = EX_i = 0.4 + 0.2 + 1 + 3 = 4.6$,
$\sigma^2 = VarX_i = EX_i^2 - (EX_i)^2 = 0.4 + 0.4 + 5 + 30 - (4.6)^2 = 35.8 - 21.16 = 14.64$,

By the CLT, $Y \approx Normal(n\mu, n\sigma^2) = Normal(184, \sqrt{585.6})$, and thus,

$P(170.5 < Y < 199.5) = P(Y \leq 199.5) - P(Y \leq 170.5) = pnorm(199.5, 184, \sqrt{585.6}) - pnorm(170.5, 184, \sqrt{585.6})$

(c) The approach in (b) will produce a more accurate approximation. The reason is that in (b) the student uses the CLT and correctly identify the distribution of Y, which is not the case in (a). In (a) the student is approximating the value by making an inference in a single run, which can vary widly due to chance variation. In other words, in (a) the student is making her calculations based on an specific instance or experiment, whereas in (b) the student is relying on the theoretical normal approximation, which should give a better approximation.