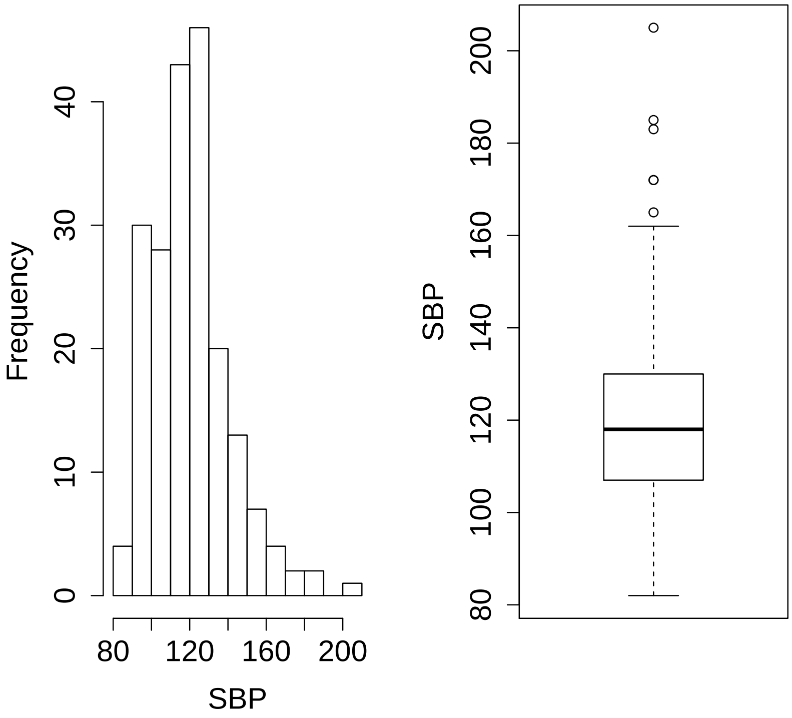
BIOS 662

Homework 2 Solution

September, 2018

Question 1

(a) Using the R functions hist() and boxplot(), we get the following output:



(b) Because $n = 200$ and $p = 0.25$, $np = 50$ is an integer, so the 25th percentile is given by

$$\hat{\zeta}_{0.25} = \frac{y_{(50)} + y_{(51)}}{2} = \frac{107 + 107}{2} = 107$$

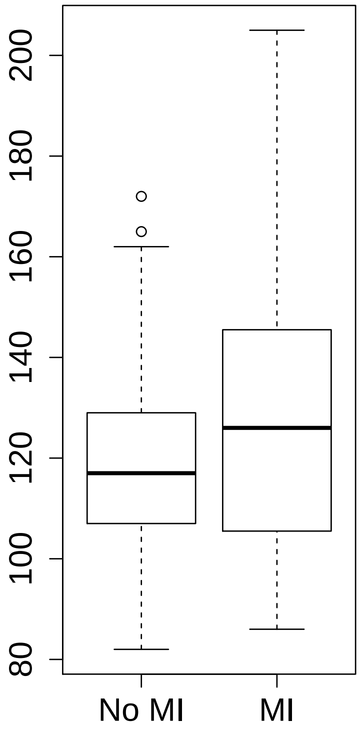
Similarly, one can show $\hat{\zeta}_{0.5} = 118$ and $\hat{\zeta}_{0.75} = 130$.

(c) Thus the IQR equals $130 - 107 = 23$.

(d) The 75th percentile + 1.5 IQR $= 130 + 1.5 \times 23 = 164.5$ and the largest observation less than this is 162. Likewise, the 25th percentile − 1.5 IQR $= 107 - 1.5 \times 23 = 72.5$ and the smallest observation greater than this is 82. Looking at the histogram we see that there are no outliers atthe lower end of the distribution, which is why in the boxplot there are no individual observations plotted below the lower whisker. The results agree with R exactly:

```
> boxplot(sbpall)$stats
     [,1]
[1,]   82
[2,]  107
[3,]  118
[4,]  130
[5,]  162
```

(e) The following figure shows side-by-side boxplots for the two groups. SBP tends to be higher in the MI group, with its median almost as high as the third quartile of the no MI group and its upper whisker extending substantially beyond the largest SBP in the no MI group.



Question 2

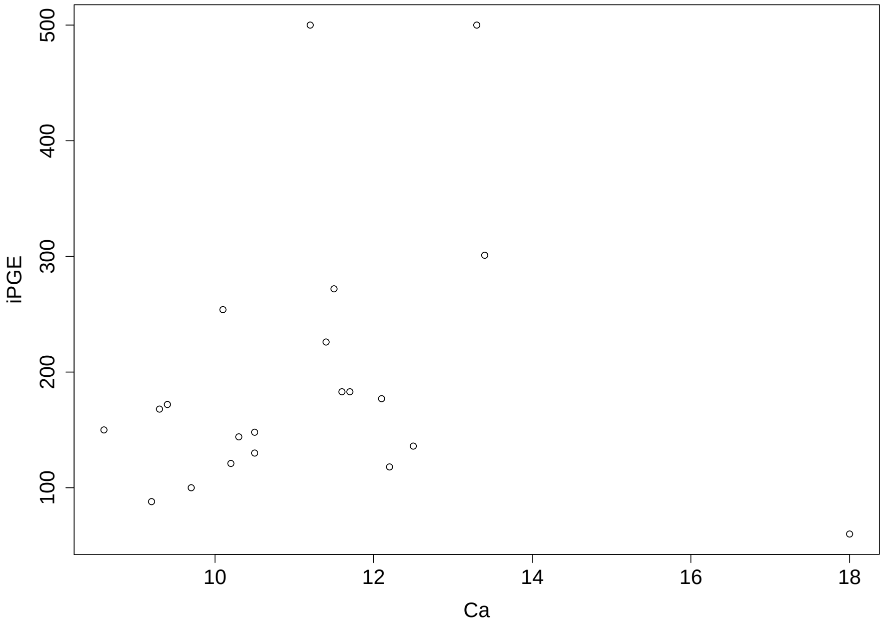
(a) $\bar{X}_{\text{Hypercalcemia}} = 2656/11 = 241.5; \quad \bar{X}_{\text{Normocalcemia}} = 1475/10 = 147.5.$

$s^2_{\text{Hypercalcemia}} = \frac{1}{11-1}\left(849988 - 11 \cdot 241.5^2\right) = 20868.47; \quad \text{so} \quad s = 144.46.$

$s^2_{\text{Normocalcemia}} = \frac{1}{10-1}\left(236749 - 10 \cdot 147.5^2\right) = 2131.83; \quad \text{so} \quad s = 46.17.$

The means seem to be substantially different. We'll see in the coming weeks that we need to use information about the standard deviations in order to decide whether the corresponding population means really do appear to differ.

(b) Below is a scatterplot of plasma iPGE against serum Ca.



If we ignore a few outliers, there is some evidence that higher plasma iPGE levels tend to be associated with higher serum Ca levels. There is substantial variability from person to person though, so if person A has higher serum Ca than person B it does not automatically follow that person A will have higher plasma iPGE than person B.

(c) Patient #11 has the highest serum Ca value among all patients yet has the lowest plasma iPGE level. A serum calcium level below 10 would be more in keeping with the tendency for lower plasma iPGE to be associated with lower serum Ca.

(d) Patients with serum calcium above 10.5 mg/dL are classified as hypercalcemic. If the serum calcium value for patient #11 is really below 10, then this patient would be classified as not having hypercalcemia and so would be moved from one group to the other. Because this patient has plasma iPGE level well below all the others in the

Hypercalcemia group, moving this patient out of the group would result in a larger mean plasma iPGE value for the group. By removing a value far from the mean, the standard deviation will decrease. The patient's plasma iPGE level is also lower than all values in the other group, so moving the patient into that group would lower its mean plasma iPGE value and because the newly added value is more extreme than other values in the group, the standard deviation will increase.

Suppose we change the serum calcium value for patient #11 from 18 to 10. Then the sample means and standard deviations of plasma iPGE change from

$\bar{X}_{\text{Hypercalcemia}} = 241.5; \quad \bar{X}_{\text{Normocalcemia}} = 147.5.$

$s_{\text{Hypercalcemia}} = 144.46; \quad s_{\text{Normocalcemia}} = 46.17$

to

$\bar{X}_{\text{Hypercalcemia}} = 259.6; \quad \bar{X}_{\text{Normocalcemia}} = 139.5.$

$s_{\text{Hypercalcemia}} = 138.43; \quad s_{\text{Normocalcemia}} = 57.13.$