Solutions to Homework 2

Solutions to Multiple Choice Questions

1. **B** – The definition of a boxplot is given on page 63 of the textbook.
2. **E** – The definition of a histogram and its construction is given on pages 54-67 of the textbook.
3. **D** – Both graphs are intended to represent the distribution of data but histograms are more informative about the shape. Consult the definition of a boxplot to confirm that it is based on the median, quartiles, and extremes.
4. **B** – The definition of the median is given on page 63 of the textbook.
5. **E** – Both the median and the mean are a measure of location so they are generally positioned near the center.
6. **C** – The data are skewed to the right when the histogram has a “fat tail” on the right hand side. The “fat tail” is attracting the mean value more than the median. The median is therefore smaller than the mean value.
7. **A** – The data are skewed toward the lower numbers and their values are between 0 and 7. Both the boxplot and histogram show this.
8. **D** – The boxplot for a data set is uniquely defined, so it cannot look different for the same data set.

Solutions to Problems

Answer to Question 1

The sample data from the four presses is shown below in table format. We have ordered the data for each press in ascending order. It should be noted that since we have sorted the data separately for each press, the field 'measurement number' in the table below does not correspond to the measurement number for the raw data given in the question. We only use the measurement number to more easily show how the data should be partitioned for determining the extrema, medians, and quartiles.

Measurement No.	Press 1	Press 2	Press 3	Press 4
1	71	69	67	64
2	81	73	76	67
3	81	76	76	70
4	82	77	77	70
5	83	77	77	70
6	84	81	78	73
7	85	81	83	73
8	85	84	85	73
9	86	86	87	74
10	87	86	90	74
11	87	88	92	75
12	89	88	92	76
13	90	89	93	76
14	91	94	93	77
15	91	95	95	78
16	93	95	97	80
17	94	96	100	84
18	95	96	103	84
19	99	101	109	85
20	101	102	110	90

Firstly, to determine the range of the data for each press we note that the range is simply the difference between the maximum value and the minimum value of the data.

- Range Press 1 = 101 – 71 = 30
- Range Press 2 = 102 – 69 = 33
- Range Press 3 = 110 – 67 = 43
- Range Press 4 = 90 – 64 = 26

The median is defined as the measurement value with half the values larger than it and half the values smaller. Since we are dealing with a sample size that is even (and not odd) we split the data set into two groups; measurements 1 – 10 and measurements 11 – 20. This partition is denoted by the horizontal green line in the table on the previous page. The median values for each press will therefore by the average of the values from measurements 10 and 11.

- Median Press 1 $= \frac{87+87}{2} = 87$
- Median Press 2 $= \frac{86+88}{2} = 87$
- Median Press 3 $= \frac{90+92}{2} = 91$
- Median Press 4 $= \frac{74+75}{2} = 74.5$

The quartiles are similar to the median; the lower quartile (Q1) is the measurement value with a quarter of the values smaller than it, and three-quarters of the values larger than it. Similarly, the upper quartile (Q3) is the measurement value with a quarter of the values larger than it, and three-quarters of the values smaller than it. Since we are dealing with a sample size that has four as a factor, we partition the data between measurements 5 and 6 (measurements 1 -5 represent the lower quarter) and between measurements 15 and 16 (measurements 15 – 16 represent the upper quarter). These partitions are denoted by the horizontal red lines in the table on the previous pages. The lower quartile values for each press will therefore be the average of the values from measurements 5 and 6. Similarly, the upper quartile values for each press will be the average of the values from measurements 15 and 16.

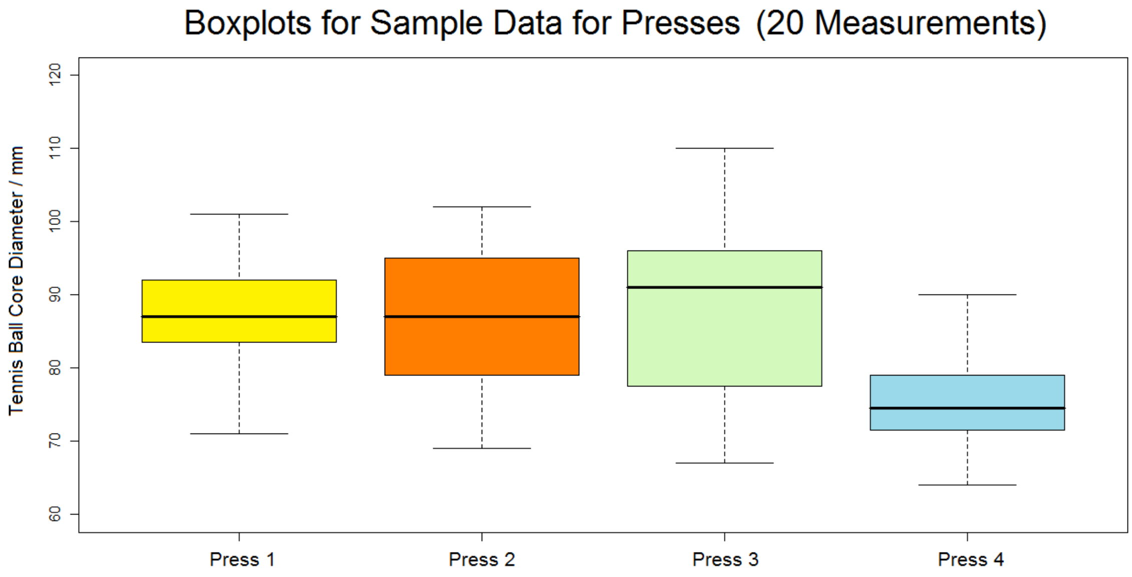
- Lower Quartile Press 1 $= \frac{83+84}{2} = 83.5$ Upper Quartile Press 1 $= \frac{91+93}{2} = 92$
- Lower Quartile Press 2 $= \frac{77+81}{2} = 79$ Upper Quartile Press 2 $= \frac{95+95}{2} = 95$
- Lower Quartile Press 3 $= \frac{77+78}{2} = 77.5$ Upper Quartile Press 3 $= \frac{95+97}{2} = 96$
- Lower Quartile Press 4 $= \frac{70+73}{2} = 71.5$ Upper Quartile Press 4 $= \frac{78+80}{2} = 79$

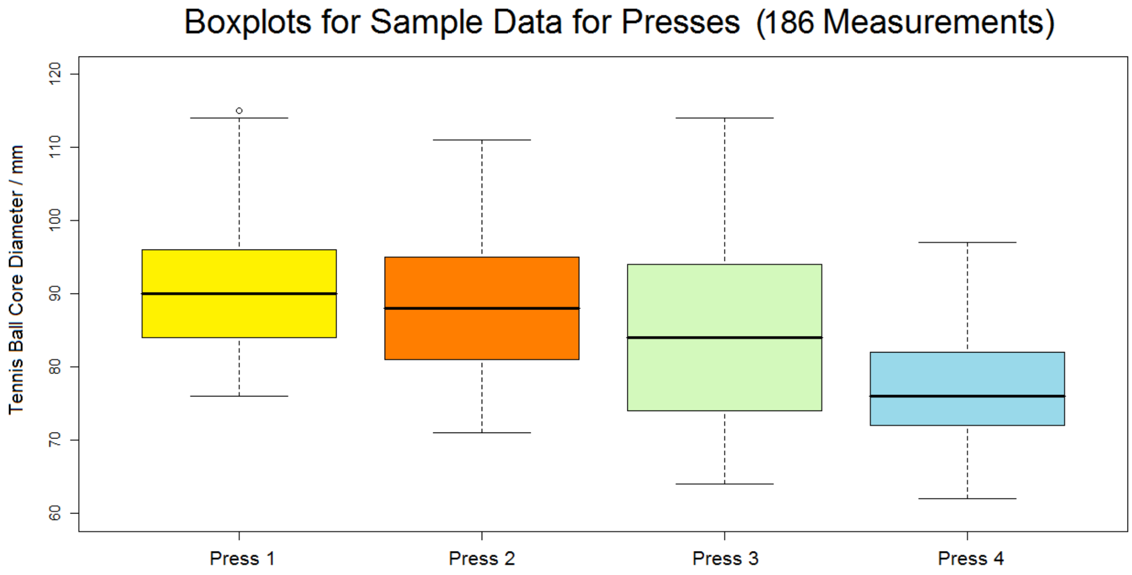
The interquartile range (IQR) for each press is the difference between the upper quartile and the lower quartile values. It defines the range where half the measurements reside.

- Interquartile Range Press 1 = 92 – 83.5 = 8.5
- Interquartile Range Press 2 = 95 – 79 = 16
- Interquartile Range Press 3 = 96 – 77.5 = 18.5
- Interquartile Range Press 4 = 79 – 71.5 = 7.5

All the statistical information we obtained for the problem is presented below in table form. Two box plots are also shown below; one for the data set given in the problem, the other for the full data set[1].

Press No.	Min	Max	Range	Median	Q1	Q3	IQR
1	71	101	30	87	83.5	92	8.5
2	69	102	33	87	79	95	16
3	67	110	43	91	77.5	96	18.5
4	64	90	26	74.5	71.5	79	7.5





[1] The full dataset can be found on the course website as an excel file called "Tennis.csv"

The top and bottom of each coloured box represents the upper and lower quartiles respectively. The height of the box is therefore equivalent to the interquartile range. The median is shown as a heavy black horizontal line that will always lie within the box. The thin horizontal line above each box is called a 'whisker'. The upper limit of the top whisker will be given by:

- Q3 + (1.5*IQR)

If there are no values in the dataset greater than this upper limit, then the actual top whisker will mark the maximum value in the dataset. For press three for the smaller dataset for example, the upper limit of the top whisker is:

- 96 + (1.5*18.5) = 123.75

But since the maximum measurement for press 3 is 110 and hence less than the upper limit of the top whisker, the top whisker will take on this value of 110. Any points which would lie above the top whisker are called outliers. There are two types of outliers; mild and extreme. A mild top outlier will lie within the range Q3 + (1.5*IQR) and Q3 + (3*IQR) and will be represented by a hollow dot. An extreme top outlier will have a value greater than Q3 + (3*IQR) and will be represented by a filled dot. On the second boxplot on the previous page we can that press 1 has a single top mild outlier.

Similarly, the thin horizontal line below each box is called the lower whisker, and the lower limit of this whisker is given by:

- Q1 - (1.5*IQR)

If there are no values in the dataset less than this lower limit, then the actual bottom whisker will mark the minimum value in the dataset. For press 3 for the smaller dataset for example, the lower limit of the bottom whisker is:

- 77.5 - (1.5*18.5) = 49.75

But since the minimum measurement for press three is 67 and hence greater than the lower limit of the bottom whisker, the bottom whisker will take on this value of 67. Any points which would lie below the bottom whisker are also called outliers. Again, there are two types of outliers; mild, and extreme. A mild top outlier will lie within the range Q1 - (1.5*IQR) and Q1 - (3*IQR) and will be represented by a hollow dot. An extreme bottom outlier will have a value smaller than Q1 - (3*IQR) and will be represented by a filled dot.

By comparing the data from the two boxplots we can clearly see that both figures reveal similar aspects about each of the presses. All the presses have very similar interquartile ranges. However, presses 1 and 2 show higher Q3-to-top-whisker ranges in the larger dataset than the smaller one. Also, the median value for press 3 is more centralised in the full dataset, than in the smaller dataset. In statistics it is important to remember that we tend to be more confident in the results from larger datasets, than smaller ones.

Answer to Question 2

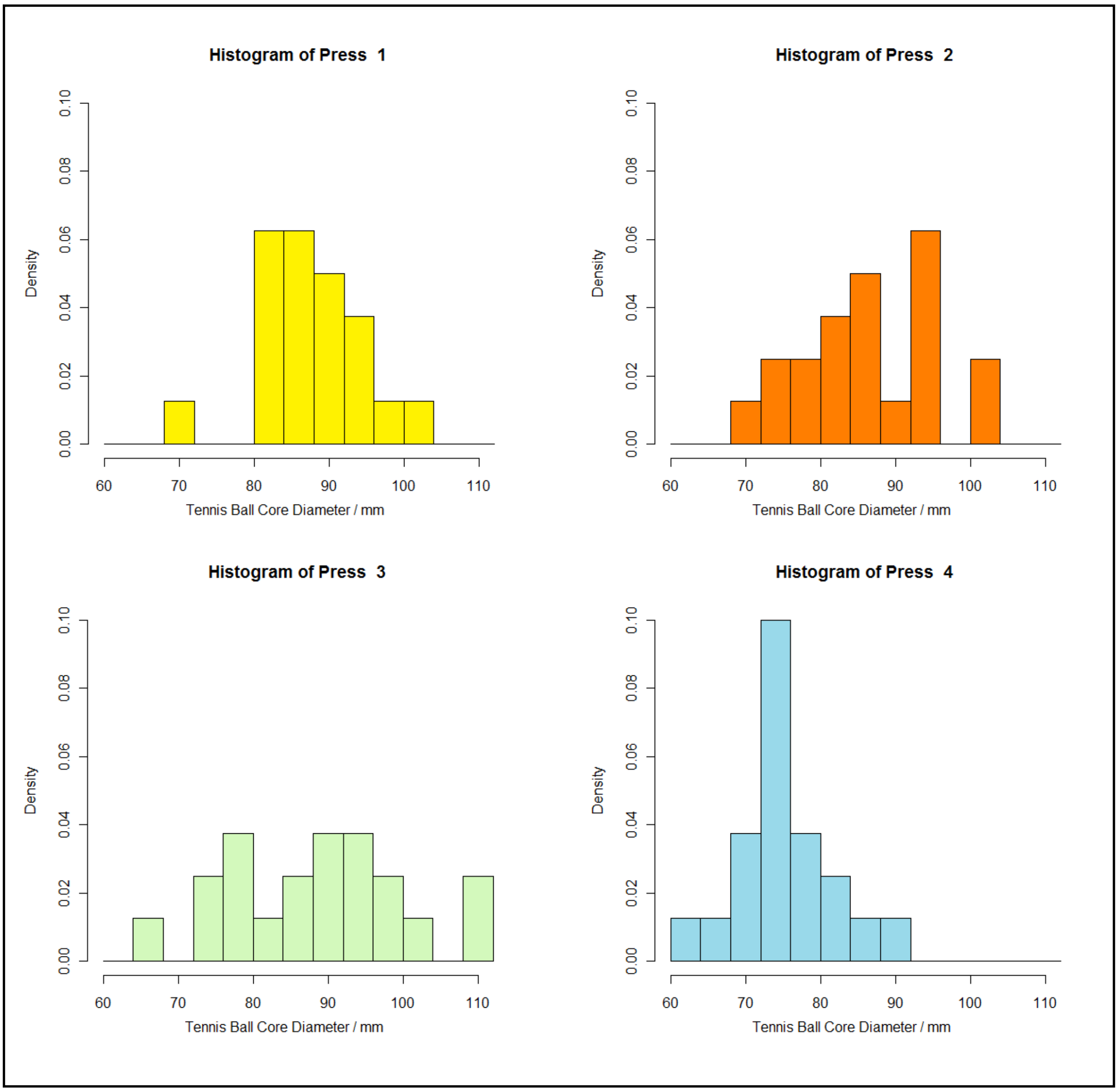
The first step in constructing histograms is to determine the width of the bin. We want to have the same bin for all four histograms. We set the bin width at an arbitrary value of 4 and range over the values of 60 to 112. We pick this range based on the maximum and minimum values for the dataset we determined in the last question.

Bin Range[2]	Count Press 1	Count Press 2	Count Press 3	Count Press 4
(60 – 64]	0	0	0	1
(64 – 68]	0	0	1	1
(68 – 72]	1	1	0	3
(72 – 76]	0	2	2	8
(76 – 80]	0	2	3	3
(80 – 84]	5	3	1	2
(84 – 88]	5	4	2	1
(88 – 92]	4	1	3	1
(92 – 96]	3	5	3	0
(96 – 100]	1	0	2	0
(100 – 104]	1	2	1	0
(104 – 108]	0	0	0	0
(108 – 112]	0	0	2	0

If we divide each count by the size of the datasets, in this case 20, then we obtain relative frequencies. If we then divide the relative frequencies by the width of the bin, in this case 4, we obtain the density. The histograms we will generate will be density plots. The densities for each press for each bin are given below in table form. The actual histograms are shown on the following page.

Bin Range	Density Press 1	Density Press 2	Density Press 3	Density Press 4
(60 – 64]	0	0	0	0.0125
(64 – 68]	0	0	0.0125	0.0125
(68 – 72]	0.0125	0.0125	0	0.0375
(72 – 76]	0	0.025	0.025	0.1
(76 – 80]	0	0.025	0.0375	0.0375
(80 – 84]	0.0625	0.0375	0.0125	0.025
(84 – 88]	0.0625	0.05	0.025	0.0125
(88 – 92]	0.05	0.0125	0.0375	0.0125
(92 – 96]	0.0375	0.0625	0.0375	0
(96 – 100]	0.0125	0	0.025	0
(100 – 104]	0.0125	0.025	0.0125	0
(104 – 108]	0	0	0	0
(108 – 112]	0	0	0.025	0

[2] The bin range excludes the lower limit but includes the upper limit. For the first bin, this means values must be greater than 60 and less than or equal to 64. Therefore a value of 60 would not fall into the first bin, but a value of 64 would.


Histogram of Press 1
Histogram of Press 2
Histogram of Press 3
Histogram of Press 4
Density
0.00
0.02
0.04
0.06
0.08
0.10
60
70
80
90
100
110
Tennis Ball Core Diameter / mm

<u>Answer to Question 3</u>

Since we already determined the median and interquartile ranges in Question 1, we need to evaluate the mean and standard deviation values for each press. The mean is given by:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_n$$

So for each press we obtain the summation of all the measurements and then divide by the total number of measurements, in this case 20. We are given the summations in the problem for each press. Therefore the means are:

- Mean Press 1 $= \frac{1775}{20} = 87.75$
- Mean Press 2 $= \frac{1734}{20} = 86.7$
- Mean Press 3 $= \frac{1780}{20} = 89$
- Mean Press 4 $= \frac{1513}{20} = 75.65$

The standard deviation is given by:

$$S = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} {X_n}^2\right) - \bar{X}^2}$$

In the problem, for each press we are given the summation of the square of each measurement. Therefore the standard deviations for each press are:

- Standard Deviation Press 1 = $\sqrt{\frac{154911}{20} - (87.75)^2} = 6.74$
- Standard Deviation Press 2 = $\sqrt{\frac{152026}{20} - (86.7)^2} = 9.188$
- Standard Deviation Press 3 = $\sqrt{\frac{161016}{20} - (89)^2} = 11.39$
- Standard Deviation Press 4 = $\sqrt{\frac{115251}{20} - (75.65)^2} = 6.295$

Shown below is a table for comparing the median and mean, and the standard deviation and interquartile ranges for the four presses.

Press No.	Mean	Median	Standard Deviation	Interquartile Range
1	87.75	87	6.74	8.5
2	86.7	87	9.188	16
3	89	91	11.39	18.5
4	75.65	74.5	6.295	7.5

From the table we observe that both the mean and median are within close proximity to each other for each press. Standard deviations and interquartile ranges are consistent in reporting large spread for presses 2 and 3. The actual numerical values for the standard deviations and interquartile ranges differ, but this is not a surprise since their bases for measuring the spread of data are different.

Answer to Question 4

a) We use the following equations to determine the mean and standard deviation:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_n \qquad S = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} {X_n}^2\right) - \bar{X}^2}$$

When we recalculate the mean and standard deviation for the smaller dataset we obtain a mean of 501.8333 and a standard deviation of 120.4866. Immediately we should notice a difference in the value for standard deviation. We suspect this is due to the so called *n-1* divisor variance which is frequently (and more properly) used for computation of variances and standard deviations (also called sample standard deviation[3]). Firstly, note that the equation for standard deviation above may be rewritten as:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_n - \bar{X})^2}$$

Sample standard deviation (SSD) however uses the equation:

$$SSD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_n - \bar{X})^2}$$

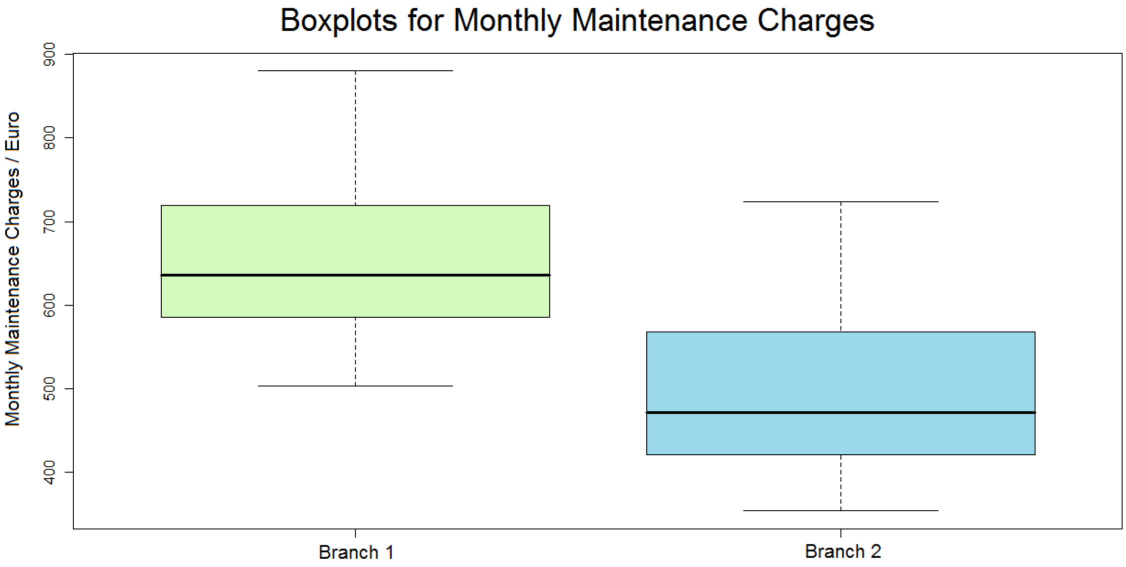
If we now use this equation to calculate the sample standard deviation the value we get is 131.9885. This value matches with that we obtained from the fellow branch manager. In conclusion, the second data set is correct. When we check the first set, we see that the mean is correct, but the sample standard deviation is 111.47. We conclude that the manager used the SSD equation (like the fellow branch manner) but missed a significant '1' when reporting the SSD.

b) Here is a summary of the correct results

Branch Number	Mean	Sample Standard Deviation
1	660.67	501.83
2	111.47	131.99

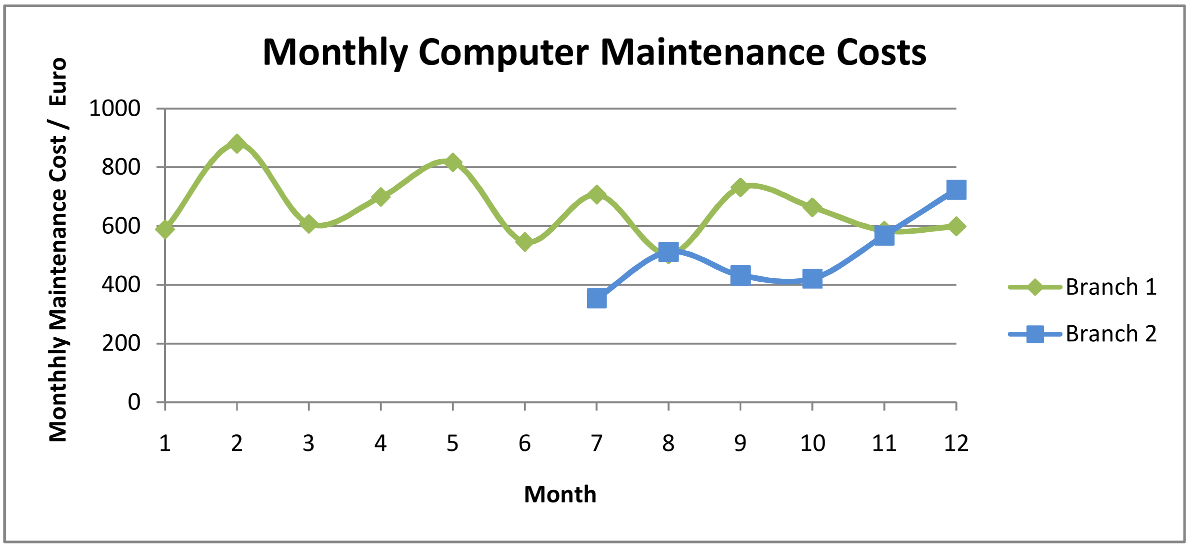
c) Boxplots are shown on the following page.

[3] When calculating standard deviation in Microsoft Excel, the function stdev() assumes sample standard deviation.



d) The manager can be suspicious about the costs in his branch where the median value appears to be way above the median value of the other branch. It is even higher than the third quartile.

e) One should be careful to draw any conclusions for two important reasons. Firstly, the second sample is relatively small (only six values) and the first one is not big either. Secondly, and more importantly, when we look also into the time aspect of the collected data (monthly data), then we noticed that large costs for the first branch occurred in the first six months for which the data are not available for the second branch. One would advise obtaining the results of the other branch for the entire year and looking for any seasonal effect on the data. Shown below is a process view of the costs for the two branches. One may be less likely to draw conclusions from the boxplots once the high variation in costs is visible in the process view.



Answer to Question 5

For this problem, we are not given the actual dataset[4], just a histogram with a narrow bin width. Since the question asks us to look at the effect of doubling bin sizes, the best approach is to create a dataset that will generate the histogram given in the problem. Microsoft Excel would be a useful tool here. We could then use a statistical program, such as 'R', to create a histogram on the dataset with greater bin widths. To create the dataset we perform the following steps:

- In between values 8 and 10 we find 20 columns. Therefore the bin width is 0.1 or $100.
- Our total range is from 6 to 16, therefore we will have 100 bins
- Our individual bin ranges will be:
  - (6.0, 6.1]
  - (6.1, 6.2]
  - (6.2, 6.3]
  - ….
  - (15.9, 16.0]
- Each bin will have an associated frequency e.g. for bin (8.0,8.1] it is 4
- When we are creating the actual values for the dataset, for each bin we can pick any value that lies within the bin's range. Since we don't know the spread of the true values within each bin, we assume that all values correspond to the central value e.g. for bin (8.0,8.1] we will use a value of 8.05.
- For each bin we then generate *n* central values, where *n* is the associated frequency for the bin
  - For bin (8.0,8.1] with frequency 4 we generate (8.05, 8.05, 8.05, 8.05) for the dataset
  - For bin (6.0,6.1] with frequency 0 we generate no values for the dataset
  - For bin (14.4,14.5] with frequency 1 we generate (14.45) for the dataset
- When we do this for all the bins we should generate 150 values, the entire dataset

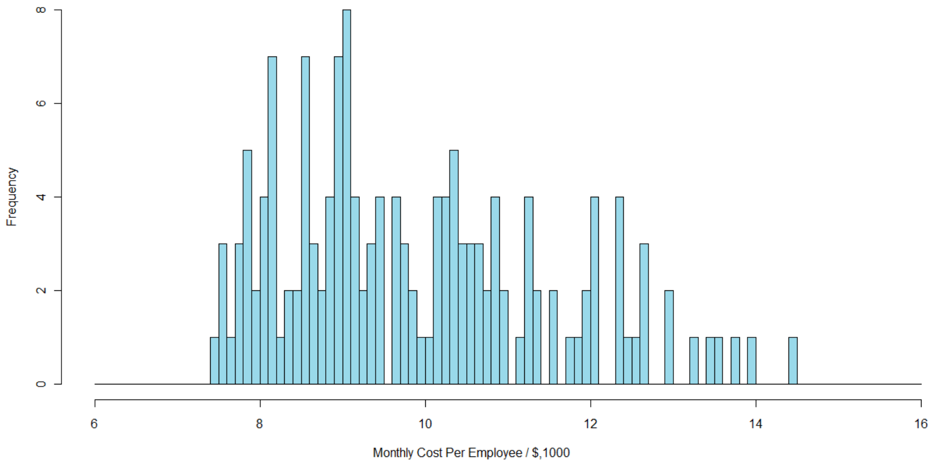
We could not use this approach if we wanted to look at histograms with bin widths less than $100. However, for bin widths greater than $100 the above approach is perfectly valid. Now that we have the dataset we can import it into 'R' and then create a histogram with a doubled bin width, $200 in this case. It can be seen in the second histogram that the frequency for each bigger bin is simply the sum of the frequencies for the corresponding two smaller bins in the first histogram.

a) The histogram with double the bin width is shown on the following page.

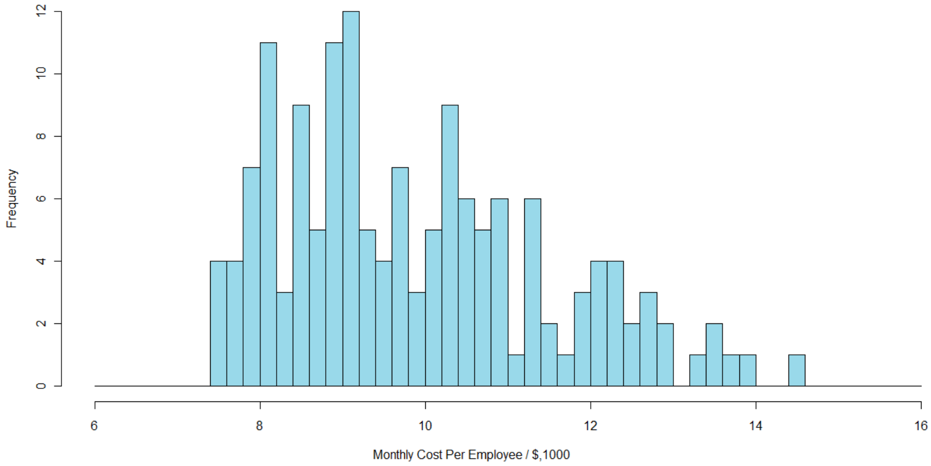
b) No, increasing the bin width to $200 did not have the desired smoothing effect. We need to double the bin width again to $400. The histogram with a bin width of $400 is also shown on the following page. This bin width appears to achieve the desired smoothing effect.

[4] The raw data may be on the course website, but at the time of writing the author didn't have internet access to check. However, this answer to the question is valid, with the only caveat being that we don't know the true mean and standard deviation of the dataset.

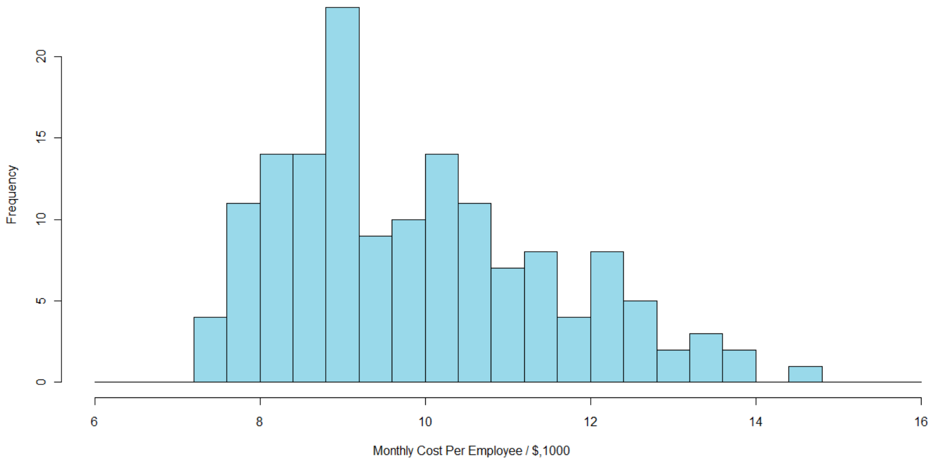
Histogram of Monthly Costs per Employee (Bin Width = $100)



Histogram of Monthly Costs per Employee (Bin Width = $200)



Histogram of Monthly Costs per Employee (Bin Width = $400)



c) By visual inspection of the third histogram, we estimate the mean to be at 10.0 or \$10,000. To determine the standard deviation we use the fact that almost all the data should be within the three standard deviations from the mean. So to quantitatively determine the standard deviation, we assume that all the data should lie within three standard deviations of the mean. The furthest non-empty bin is at (14.4, 14.8] so we will assume that the value in this bin is 14.6. This value is the furthest from the mean, therefore we determine the standard deviation as follows:

$3\sigma = 14.6 - 10.0$

$\sigma = 1.533$

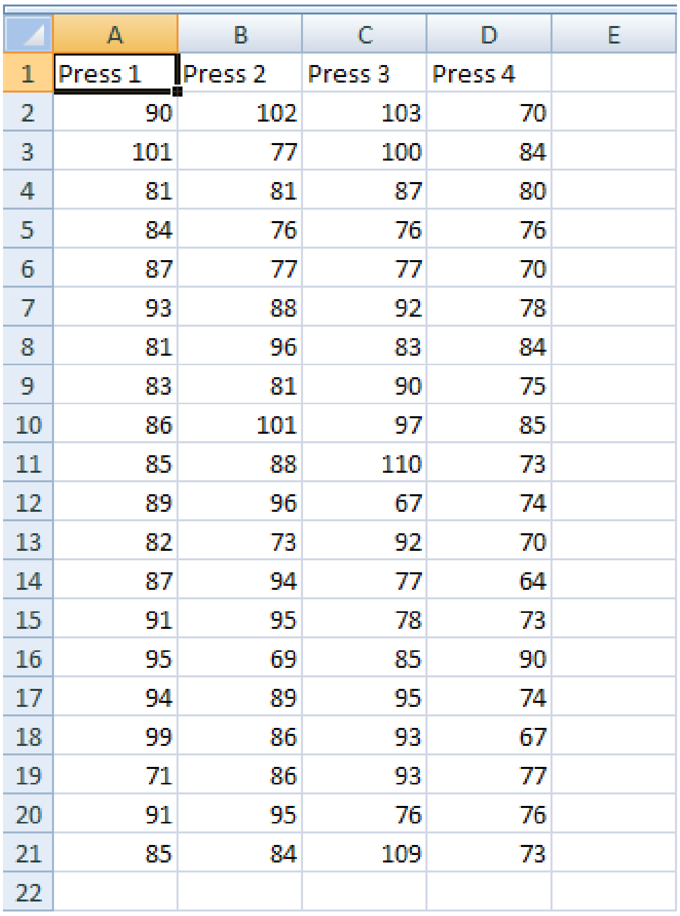
$\sigma = \$1{,}533$

So our estimates for the mean and standard deviation are \$10,000 and \$1,533 respectively. The raw data that we created from the first histogram generates a mean and standard deviation of \$9,886 and \$1,617 respectively. However, it is important to remember that we created the raw data from the first histogram, and we generated values knowing only what bin range they occurred in, so we don’t know the true mean and standard deviation.

R-Code for Problems[5]

R-Code for Creating Boxplots for Question 1

Store the data in an excel file as shown below, and to save it in CSV format as “Tennis Small.csv”



	A	B	C	D	E
1	Press 1	Press 2	Press 3	Press 4	
2	90	102	103	70	
3	101	77	100	84	
4	81	81	87	80	
5	84	76	76	76	
6	87	77	77	70	
7	93	88	92	78	
8	81	96	83	84	
9	83	81	90	75	
10	86	101	97	85	
11	85	88	110	73	
12	89	96	67	74	
13	82	73	92	70	
14	87	94	77	64	
15	91	95	78	73	
16	95	69	85	90	
17	94	89	95	74	
18	99	86	93	67	
19	71	86	93	77	
20	91	95	76	76	
21	85	84	109	73	
22					

```
// GET THE SMALL DATASET FROM THE CSV FILE
tennis_small=read.csv(“G:/Business Statistics/Tennis Small.csv”)

// CREATE BOXPLOTS FOR THE SMALLER DATASET, AND SET THE LIMITS ON THE Y-AXIS TO BETWEEN 60
// AND 120, SINCE THE FULL DATASET WILL ALSO LIE WITHIN THESE LIMITS.
boxplot(tennis_small,ylim=range(60,120))

// GET THE FULL DATASET FROM THE CSV FILE
tennis_full =read.csv(“G:/Business Statistics/Tennis.csv”)

// CREATE BOXPLOTS FOR THE FULL DATASET
boxplot(tennis_full,ylim=range(60,120))
```

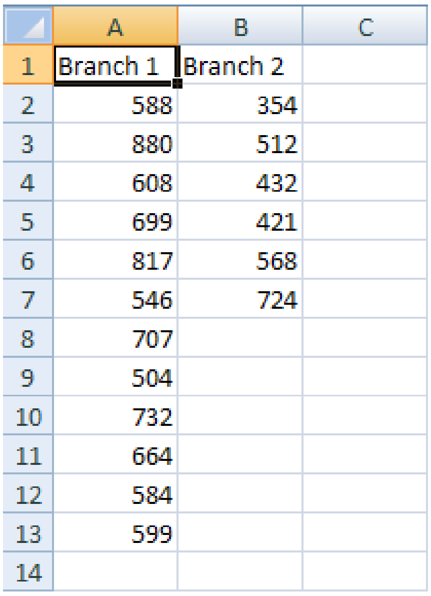
[5] Comments are preceded by “//” and are shown in green font.

R-Code for Creating Histograms for Question 2

```
// CREATE HISTOGRAMS FOR THE SMALLER DATASET, AND SET THE RANGE BETWEEN 60 AND
// 112, AND THE BIN WIDTH AS 4. SET THE HISTOGRAMS TO PLOT DENSITY, BY SETTING THE VALUE prob
// AS TRUE.  THE LIMITS ON THE Y-AXIS ARE SET TO BETWEEN 0 AND 0.1 FOR ALL THE HISTOGRAMS.
hist(tennis_small$Press.1,seq(60,112,4),prob=TRUE,ylim=range(0,0.1))
hist(tennis_small$Press.2,seq(60,112,4),prob=TRUE,ylim=range(0,0.1))
hist(tennis_small$Press.3,seq(60,112,4),prob=TRUE,ylim=range(0,0.1))
hist(tennis_small$Press.4,seq(60,112,4),prob=TRUE,ylim=range(0,0.1))
```

R-Code for Creating Boxplot for Question 4

Firstly, store the data in an excel file as shown below, and to save it in CSV (comma delimited) format.



	A	B	C
1	Branch 1	Branch 2	
2	588	354	
3	880	512	
4	608	432	
5	699	421	
6	817	568	
7	546	724	
8	707		
9	504		
10	732		
11	664		
12	584		
13	599		
14			

```
// GET THE DATASET FROM THE CSV FILE, THEN CREATE A BOXPLOT
branches=read.csv(“G:/Business Statistics/Branch.csv”)
boxplot(branches)
```

R-Code for Creating Histograms for Question 5

```
// CREATE AN ARRAY OF DATA
costs<-scan()
// THEN COPY & PASTE DIRECTLY INTO THE COMMAND LINE, THE COLUMN OF COST DATA

// CREATE THREE HISTOGRAMS WITH INCREASING BIN WIDTHS
hist(costs,seq(6,16,0.1))
hist(costs,seq(6,16,0.2))
hist(costs,seq(6,16,0.4))
```