Homework 2

Here we present solutions to the problems posted in the second homework assignment. The solutions and related commentary are put in italics. Remember that problems can have several different but correct ways of solving them.

Multiple choice questions

1. A boxplot is

   A a diagram, incorporating a box, which illustrates the centre and spread of a set of measurements in terms of the mean and standard deviation;

   ***B** a diagram, incorporating a box, which illustrates the centre and spread of a set of measurements in terms of the median, the quartiles and the extremes;

   C a plot, graph or chart, such as a histogram, pie-chart, etc., which is framed by a box;

   D a diagram consisting of a series of adjacent rectangles (boxes) in which the area of each rectangle represents the frequency of values in the corresponding interval on the horizontal axis.

   E a histogram where rectangles are made up of varying numbers of boxes of equal size.

   Solution: *Consult the definition of boxplot given in the lecture (slides) as well as the definition from the textboook, p. 63.*

2. A histogram is

   A a diagram which shows how the history of a process changes over time;

   B an arrangement of the steps in a process in order, with the steps displayed in a series of interconnected rectangles (or other shapes);

   C a line plot where the height of the line represents the frequency of the corresponding value;

   D an elaborate form of boxplot in which several rectangular boxes are used instead of just one;

   ***E** a diagram consisting of a series of adjacent rectangles in which the area of each rectangle represents the frequency of values in the corresponding interval on the horizontal axis.

   Solution: *Consult the definition of histogram and its construction given in the lecture (slides) as well as the definition from the textboook, p. 64-67.*

3. Histograms and boxplots

   A represent graphically completely different aspects of the data;

   B both represent frequency distributions of the data but histogram contains less information than boxplot;

C both are constructed using median, quartiles, and extremes;

***D** both represent frequency distributions of the data but histogram contains more information about the shape of distribution while boxplot is based solely on median, quartiles, and extremes;

E are not very useful tools to summarize data.

Solution: *Both graphs are intended to represent the distribution of data but histogram is more informative about the shape, consult the definition of boxplot to confirm that it is based on median, quartiles, and extremes.*

4. The median of a set of numbers is

A halfway between the quartiles;

***B** the value with half the numbers larger than it and half smaller;

C the most frequent value;

D the sum of the numbers divided by their number;

E the value which most evenly spaces the sample.

Solution: *The answer follows directly from the definition of median, page 63 of the textbook, or slides from the lecture.*

5. Which of the following is correct?

A the Range of a set of numbers is the list of numbers laid out in order from smallest to largest;

B the Interquartile range of a set of numbers is the list of numbers between the quartiles laid out in order from smallest to largest;

C the Median of a set of numbers is the value half way between the smallest and largest;

D the Standard Deviation of a set of numbers is the average of the deviations from the mean divided by the square root of their number;

***E** typically, the Mean of a set of numbers will be near the centre of the set, similar to the Median.

Solution: *Both the median and the mean are measure of location so they are positioned near the center.*

6. We say that data are skewed to the right if

A their median is larger than their mean;

B their interquartile range is larger than their standard deviation;

***C** the histogram has the right hand side portion prolongated more then the left hand side and the median is smaller than the mean;

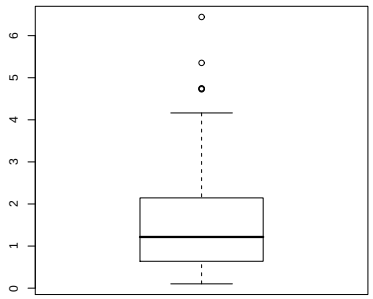
D the center of the frequency distribution is more to the right, i.e. toward the maximum, then to the left;

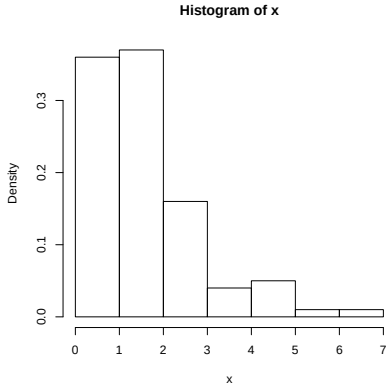
E on the horizontal coordinate the median is more to the right than the mean.

Solution: *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, so the latter is smaller than the former.*

7. Of the following pairs of graphs, there is a pair, graphs of which correspond to the same data set. Identify this pair

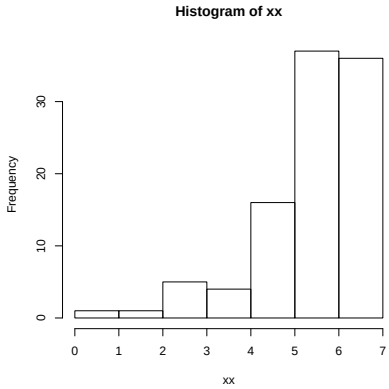
***A**

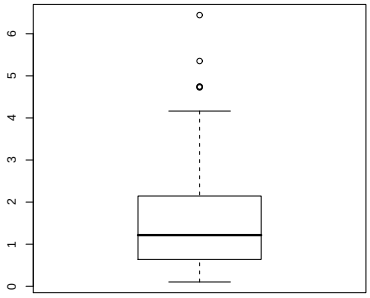




;

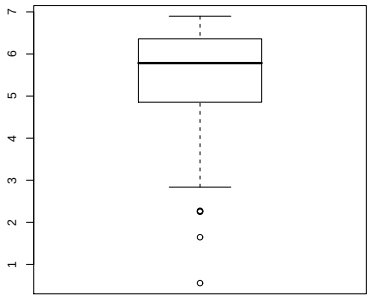
B

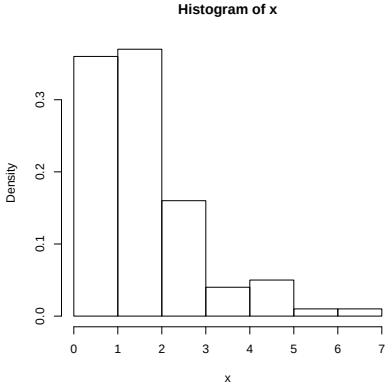




;

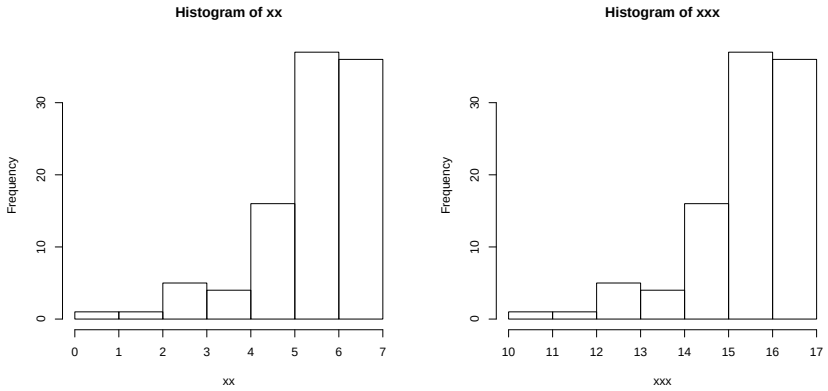
C





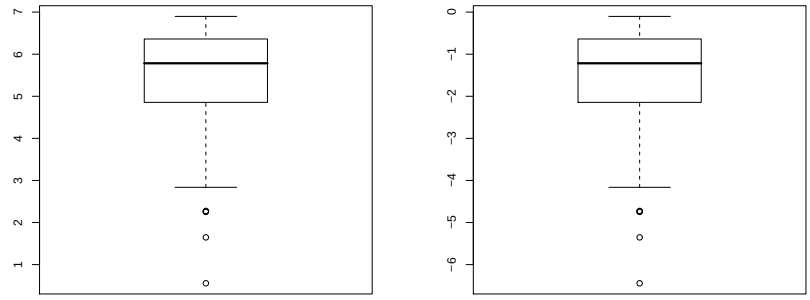
;

D



;

E

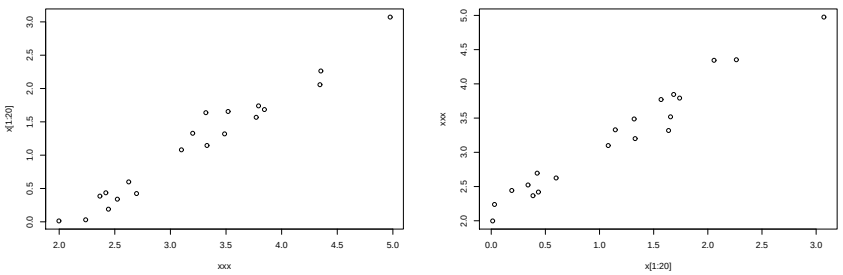


.

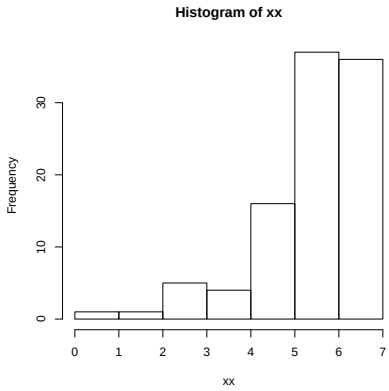
Solution: *The data are skewed to the right and their values are between 0 and 7.*

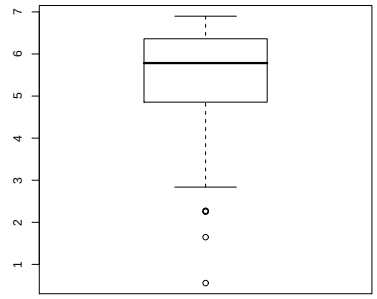
8. Of the following pairs of graphs, there is only one pair, graphs of which do not correspond to the same data set. Identify this pair

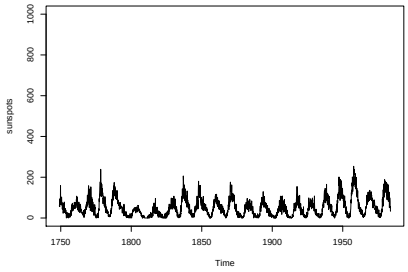
A

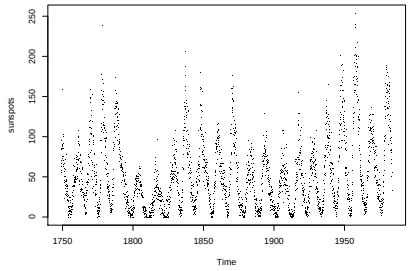


;

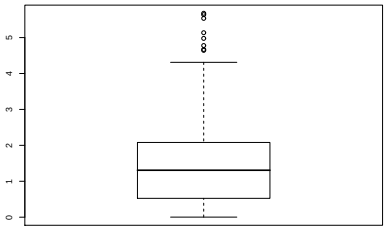
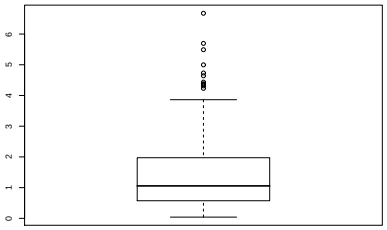
B 

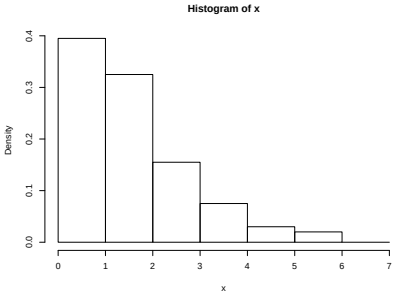

;

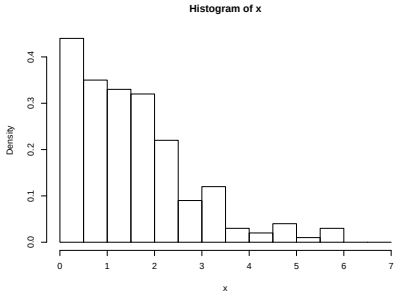
C 



;

***D** 

;

E 



.

Solution: *The boxplot for a data set is uniquely defined, so it can not look differently for the same data set.*

Problems

1. A sample of 20 measurements from each of the four presses are presented below as well as on page 64 in Exercise 2.2 of the textbook.

```
"Press 1",90,101,81,84,87,93,81,83,86,85,89,82,87,91,95,94,99,71,91,85
"Press 2",102,77,81,76,77,88,96,81,101,88,96,73,94,95,69,89,86,86,95,84
"Press 3",103,100,87,76,77,92,83,90,97,110,67,92,77,78,85,95,93,93,76,109
"Press 4",70,84,80,76,70,78,84,75,85,73,74,70,64,73,90,74,67,77,76,73
```

Find the medians, quartiles, ranges, and interquartile ranges for these data. Then plot the boxplots for these data sets and compare with the boxplots of the entire data sets as shown in Figure 2.2, p.64, of the textbook or in the lecture slides. Comment.

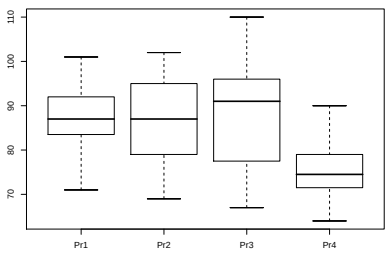
Solution: *The first step to finding quartiles and media is to order the data*

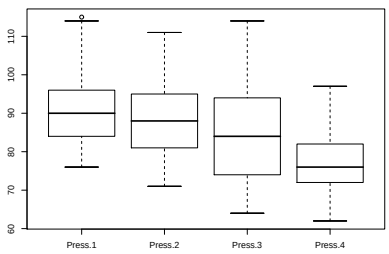
```
Press 1: 71  81  81  82  83  84  85  85  86  87  87  89  90  91  91  93  94  95  99  101
Press 2: 69  73  76  77  77  81  81  84  86  86  88  88  89  94  95  95  96  96 101  102
Press 3: 67  76  76  77  77  78  83  85  87  90  92  92  93  93  95  97 100 103 109  110
Press 4: 64  67  70  70  70  73  73  73  74  74  75  76  76  77  78  80  84  84  85   90
```

From the ordered data one can read that the extremal values and quartiles are as follows

	Press1	Press2	Press3	Press4
Min	71.0	69	67.0	64.0
Q1	83.5	79	77.5	71.5
Med	87.0	87	91.0	74.5
Q3	92.0	95	96.0	79.0
Max	101.0	102	110.0	90.0

Thus ranges for each press are: 30, 33, 43, 26, respectively, and the interquartile ranges are 8.5, 16, 18.5, 7.5. The boxplot for this data set is shown in the left hand side figure below and for the entire data set on the right hand side one. We clearly see that both figures reveals similar aspects of each of the presses although for Press 2 smaller data set shows somewhat overblown spread. In statistics, we tend to believe more to what larger data sets are showing – this time bigger is better.





;

2. For the data in the previous problem construct histograms. Describe all steps that lead to the final graphs.

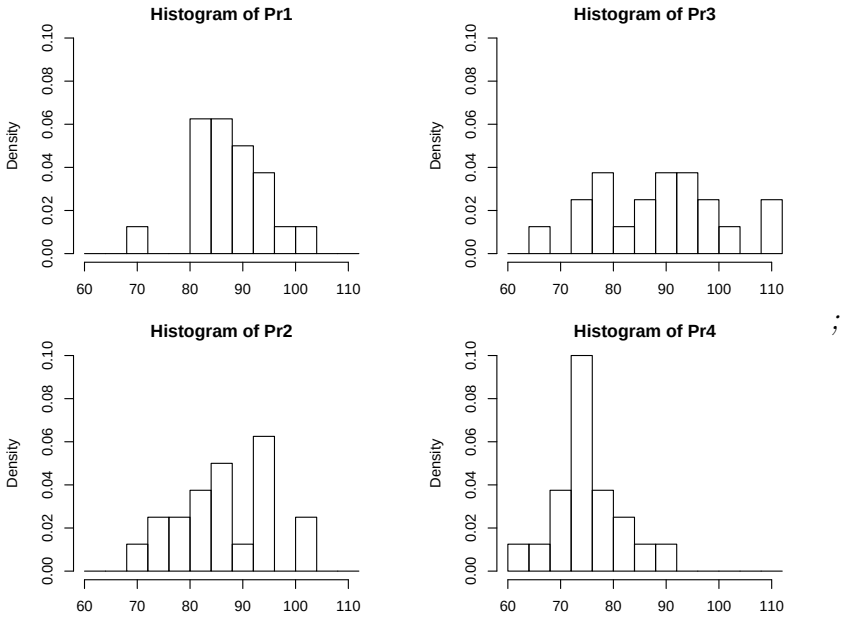
Solution: *The first step in constructing histogram is to determine the bins. We want to have the same bins for all four histograms. The information from the solution to the previous problem is very helpful here. We set for the bins having width 4 and starting from 60 and ending at 112, so there is total of 13 bins. For these bins we compute raw counts of the data falling into each bin, so called frequencies. Here they are*

Bins:	60	64	68	72	76	80	84	88	92	96	100	104	108	112
Press 1:	0	0	1	0	0	5	5	4	3	1	1	0	0	
Press 2:	0	0	1	2	2	3	4	1	5	0	2	0	0	
Press 3:	0	1	0	2	3	1	2	3	3	2	1	0	2	
Press 4:	1	1	3	8	3	2	1	1	0	0	0	0	0	

If these counts are divided by the size of data sets, in this case 20, then we obtain relative frequencies, if additionally they are divided by the width of the bin, in this case 4, than we obtain the heights of the rectangles suitable for the density plots of histograms:

```
Press 1: 0.0000 0.0000 0.0125 0.0000 0.0000 0.0625 0.0625 0.0500 0.0375 0.0125 0.0125 0.0000 0.0000
Press 2: 0.0000 0.0000 0.0125 0.0250 0.0250 0.0375 0.0500 0.0125 0.0625 0.0000 0.0250 0.0000 0.0000
Press 3: 0.0000 0.0125 0.0000 0.0250 0.0375 0.0125 0.0250 0.0375 0.0375 0.0250 0.0125 0.0000 0.0250
Press 4: 0.0125 0.0125 0.0375 0.1000 0.0375 0.0250 0.0125 0.0125 0.0000 0.0000 0.0000 0.0000 0.0000
```

The histograms are shown next



3. Continuing work on the previous data sets, evaluate and compare the following characteristics: medians and means, interquartile ranges and standard deviations. Which of the data set is the most spread? Which of the data is located the most

to the right on horizontal axis? The following computations can be helpful in this problem

$$\begin{aligned}
90+101+81+84+87+93+81+83+86+85+89+82+87+91+95+94+99+71+91+85 &= 1755,\\
102+77+81+76+77+88+96+81+101+88+96+73+94+95+69+89+86+86+95+84 &= 1734,\\
103+100+87+76+77+92+83+90+97+110+67+92+77+78+85+95+93+93+76+109 &= 1780,\\
70+84+80+76+70+78+84+75+85+73+74+70+64+73+90+74+67+77+76+73 &= 1513.
\end{aligned}$$

$$\begin{aligned}
90^2+101^2+81^2+84^2+87^2+93^2+81^2+83^2+86^2+85^2+89^2+82^2+87^2+91^2+95^2+94^2+99^2+71^2+91^2+85^2 &= \\
&= 154911,\\
102^2+77^2+81^2+76^2+77^2+88^2+96^2+81^2+101^2+88^2+96^2+73^2+94^2+95^2+69^2+89^2+86^2+86^2+95^2+84^2 &= \\
&= 152026,\\
103^2+100^2+87^2+76^2+77^2+92^2+83^2+90^2+97^2+110^2+67^2+92^2+77^2+78^2+85^2+95^2+93^2+93^2+76^2+109^2 &= \\
&= 161016,\\
70^2+84^2+80^2+76^2+70^2+78^2+84^2+75^2+85^2+73^2+74^2+70^2+64^2+73^2+90^2+74^2+67^2+77^2+76^2+73^2 &= \\
&= 115251.
\end{aligned}$$

Solution: *The median and interquartile ranges have been compute in the previous problems. We need to evaluate the means and standard deviations. Recall that the mean $\bar{X}$ is defined as*

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

In our case $n = 20$ and the numerators for each data set are given above which leads to $1755/20 = 87.75$, $1734/20 = 86.7$, $1780/20 = 89$, $1513/20 = 75.65$, respectively.

For the standard deviations, we start with variances and use the formula that has been presented in the lecture (conslut slides)

$$S^2 = \frac{X_1^2 + \cdots + X_n^2}{n} - \bar{X}^2.$$

Thus using the values for the numerators for each data set as given above we obtain variances

$$\begin{aligned}
\frac{154911}{20} - 87.75^2 &= 45.4875,\\
\frac{152026}{20} - 86.7^2 &= 84.41,\\
\frac{161016}{20} - 89^2 &= 129.8,\\
\frac{115251}{20} - 75.65^2 &= 39.6275.
\end{aligned}$$

Standard deviations are square roots of variances, so we have 6.74, 9.188, 11.39, 6.295, respectively. We can place all computed characteristics in a table to make comparisons easier

	Press1	Press2	Press3	Press4
Mean	87.75	86.7	89	75.65

```
Med   87.0    87       91.0   74.5
Std    6.74    9.188   11.39   6.295
IQR    8.5    16       18.5    7.5
```

From the table we observe that both mean and median identify a similar location of each data set, with Press 4 having noticeable smaller value of its central location. In terms of comparisons of spreads of four data sets, standard deviations and interquartile ranges are consistent in reporting large spread for Press 3 and also somewhat large for Press 2. The actual numerical values of standard deviation and interquartile range differ but this is not a surprise as they were developed on different premises on how to measure spread of the data.

4. A building society (savings and loan association) branch manager is concerned about computer maintenance charges in his branch and decides to investigate. As a first step, he retrieves from the branch accounting system the monthly maintenance charges for the previous twelve months (in euros):

$$588, 880, 608, 699, 817, 546, 707, 504, 732, 664, 584, 599$$

It yielded values for the mean and standard deviation of 660.67 and 11.47, respectively. For comparison, he telephones a fellow branch manager and persuades her to give him her monthly maintenance charges for the previous six months. They were

$$354, 512, 432, 421, 568, 724.$$

He calculates the mean and standard deviation for the other branch, finding $\bar{X} = 501.83$ and $s = 131.99$.

(a) recalcualate the mean and standard deviation for the fellow manager branch,

(b) create a table summarizing the results for two branches,

(c) create boxplots for the two branches,

(d) comment what kind of conclusions you can draw from the analysis of the data for two branches,

(e) do you think that the manager can jump to any final conclusions about computer maintenance charges for his branch?

Solution:

(a) The mean and standard deviation are recalculated for the second smaller data set as follows

$$(354 + 512 + 432 + 421 + 568 + 724)/6 = 501.8333$$
$$\sqrt{(354^2 + 512^2 + 432^2 + 421^2 + 568^2 + 724^2)/6 - 501.8333^2} = 120.4886$$

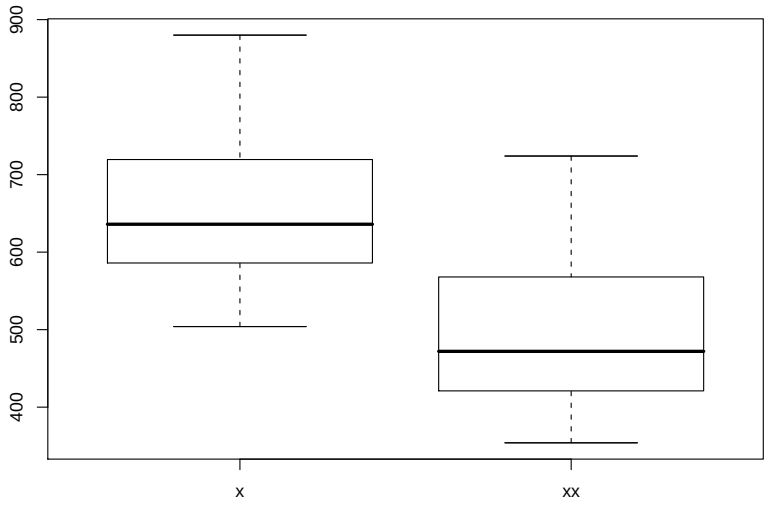
We note descripency between the reported value and the recomputed value of standard deviation. We suspect that this is due to so called $n-1$ divisor variance which is frequently (and more properly) used for computation of variances and standard deviations. Quick computation confirms our suspicion

$120.4885 * \sqrt{(6/5)} = 131.9885$. *In the conclusion, computations made by the manager are correct for the second data set. We can do the same computations for the first data set to find out that it actually yields the same mean as reported but the standard deviation is 111.47. We conclude that the manager (or the author of the textbook, see page 92) has missed one the most significant '1' when reporting standard deviation.*

(b) Here is a summary of the correct *results*

	Branch1	Branch2
Mean	660.67	501.83
Std	111.47	131.99

(c) Histograms are ploted below

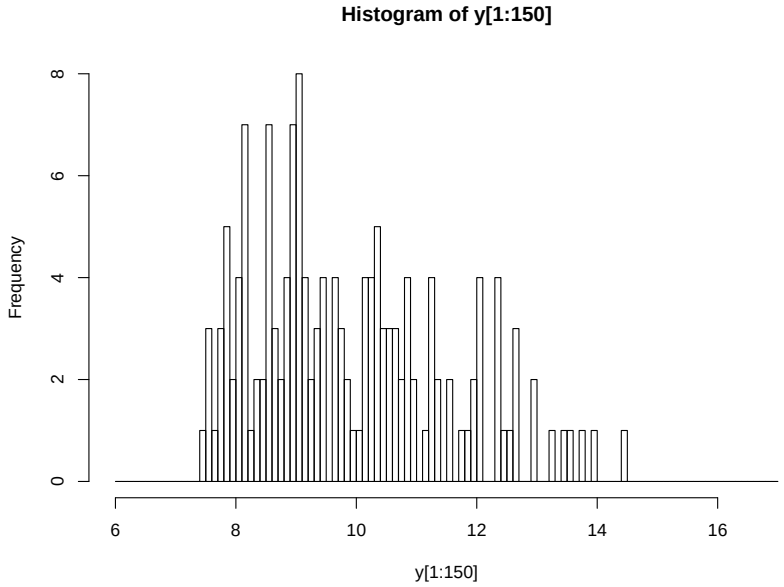


;

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

(e) However, one should be carful to draw any final conclusion 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 notice that large costs for the first branch occured in the first six months for which the data are not available for the second branch. One should advise obtaining the results of the other branch for the entire year and maybe studying seasonal effect on the data.

5. In an extensive study about the cost of running a real estate agency in the Bay Area, California, data have been collected on the monthly cost in k$ per an employee from 150 agencies. The histogram of raw counts (frequencies) resulting from this data set is presented below



The graph appears not to be smoothed enough to show the bell shaped distribution as it was expected.

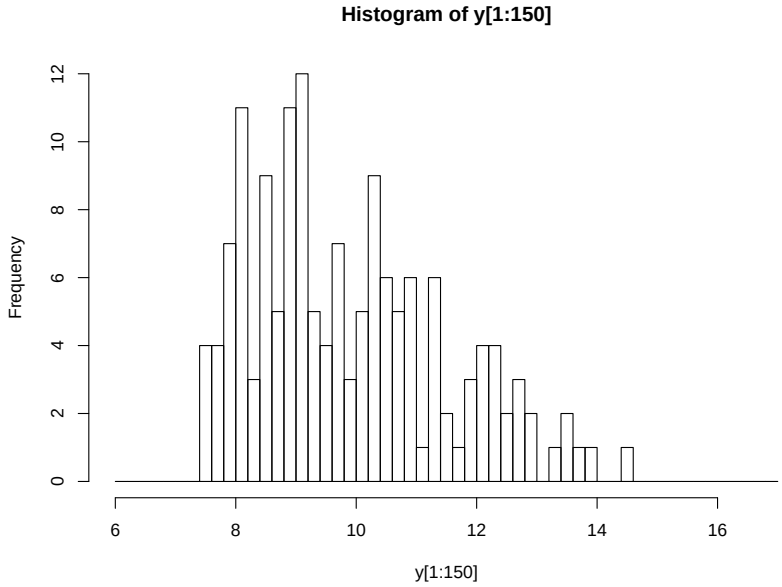
(a) Check how the graph will change if the bin sizes are doubled by constructing an appropriate histogram.

(b) Did the doubling bin sizes have the desired smoothing effect?

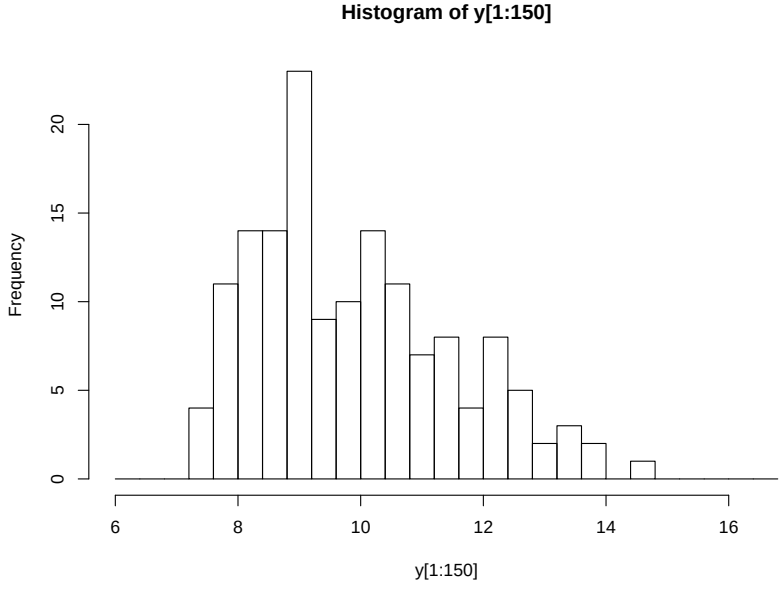
(c) By simply looking at the histograms, provide with your rough estimate of the mean and standard deviation of the underlying data on the costs of running real estate business.

Solution:

(a) It is easy to get the histogram with the doubled bin size. Simply each bigger bin is made of two smaller one so the counts for the bin translate to sum of the counts for smaller bin. Consequently the height for the new histogram will be the sum of heights for the corresponding bins in the old histogram (in the case of the density version of histogram this sum must be divided by two due to the increase in the length of the bin). Resulting histogram (raw counts version) is presented below



(b) Yes, the histogram is smoother now but one could argue that it is not smooth enough. We can repeat the procedure and this yields



(c) By inspecting histogram our rough estimate of the mean value and standard deviation would be k$10, k$1.5, respectively. The choice for the mean is obtain by visual assesment of the center of the data, while standard deviation was estimated using fact that almost all of the data should be within three standard deviations from the mean. The true values of this two parameters for this data set are k$9.88325, k$1.618740, respectively.