Homework 2

Here is the list of problems constituting the second homework assignment. First, please, try to find your own solutions and after this effort consult these solutions with the ones presented during tutorials and, finaly, check the solutions that will be posted on the course webpage. 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.

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.

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.

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.

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.

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;

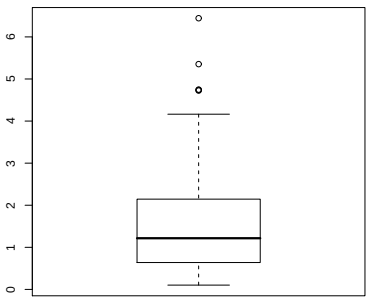
   C the histogram has the right hand side portion prolongated more then the left hand side and the median is smaller than the mean;

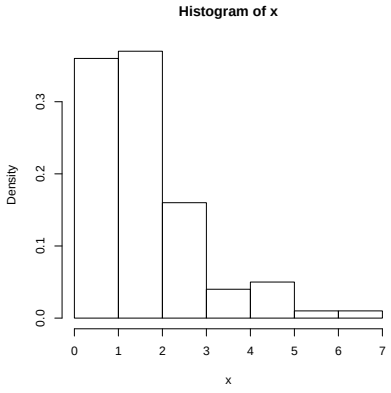
   D the center of the frequency distribution is more to the right, i.e. toward the maximum, then to the left;

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

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

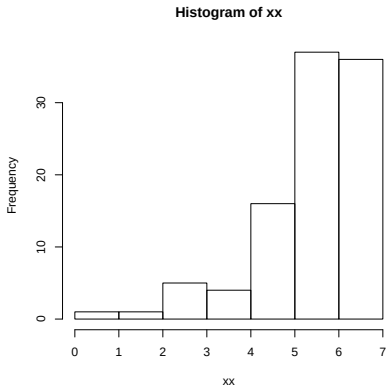
A

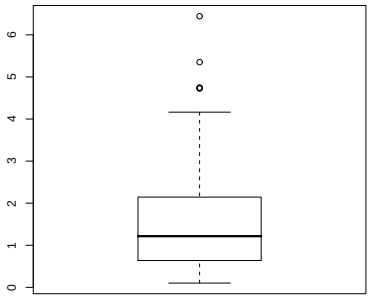




;

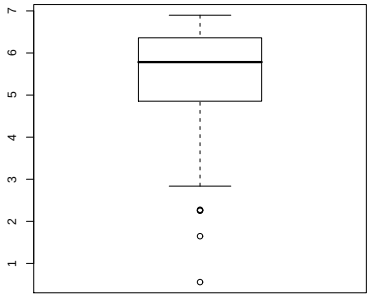
B

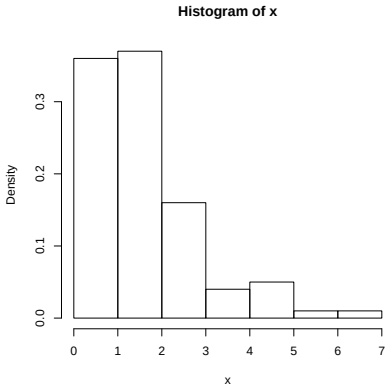

Histogram of xx
Frequency
xx
0
10
20
30
0
1
2
3
4
5
6
7


0
1
2
3
4
5
6

;

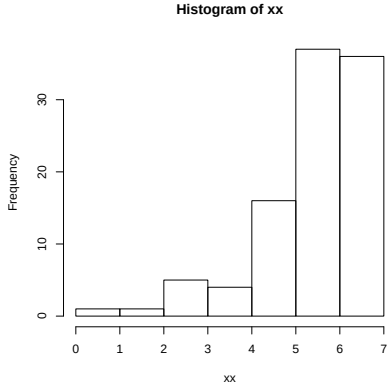
C

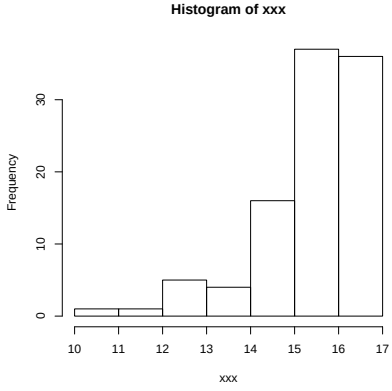

1
2
3
4
5
6
7


Histogram of x
Density
x
0.0
0.1
0.2
0.3
0
1
2
3
4
5
6
7

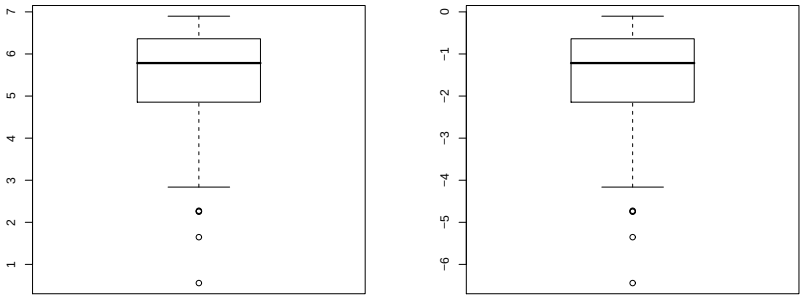
;

D

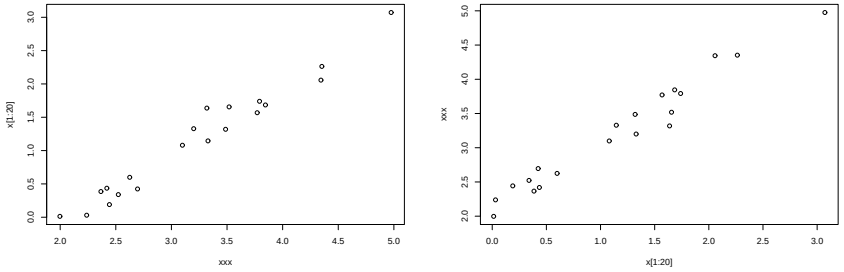

Histogram of xx
Frequency
xx
0
10
20
30
0
1
2
3
4
5
6
7

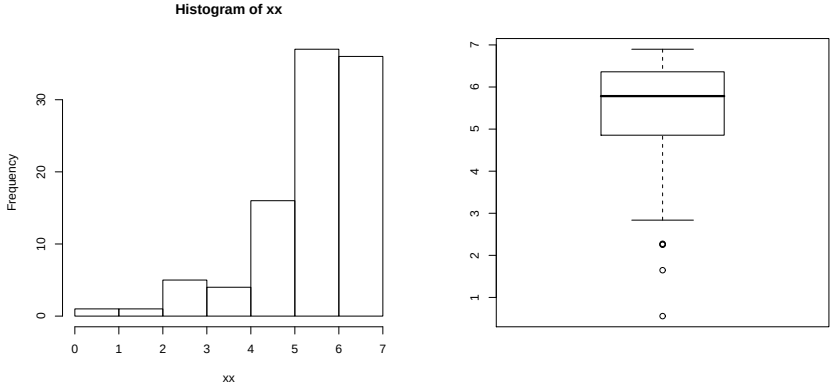

Histogram of xxx
Frequency
xxx
0
10
20
30
10
11
12
13
14
15
16
17

;

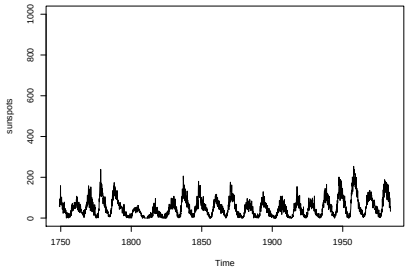
E  .

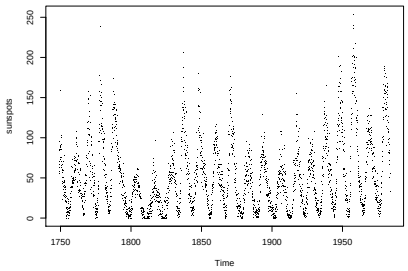
8. Of the following pairs of graphs, there is only one pair, graphs of which do not correspond to the same data set. Identify this pair

A  ;

B  ;

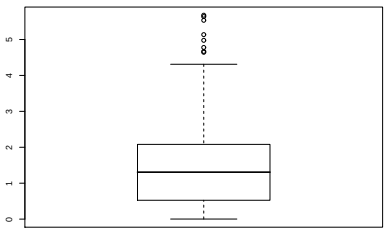
C

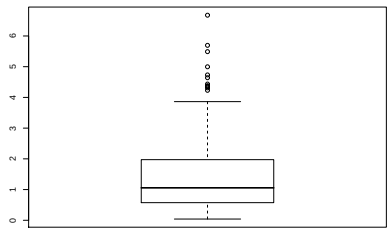




;

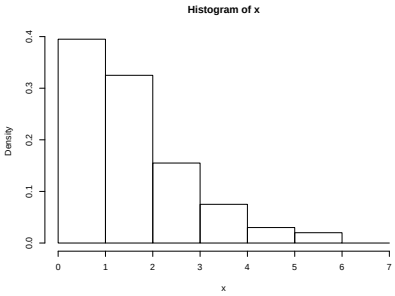
D

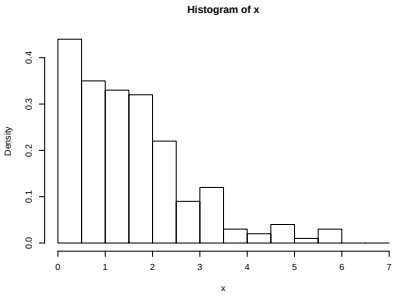




;

E





.

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.

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

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

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

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

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,

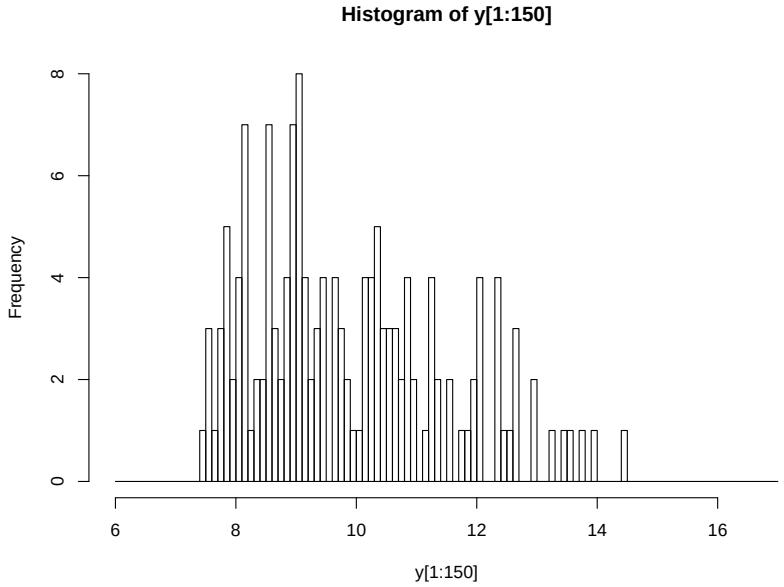
(b) create a table summarizing the results for two branches,

(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?

5. In an extensive study about the cost of running a real estate agency in the Bay Area, 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.