1. Numerical Summaries

Example 1. Time to Start a Business (Data: eg01-23time24.txt)

An entrepreneur faces many bureaucratic and legal hurdles when starting a new business. The World Bank collects information about starting businesses throughout the world. It has determined the time, in days, to complete all of the procedures required to start a business. Data for 195 countries are included in the data set. For this section we will examine data for a sample of 24 of these countries. Here are the data:

<table>
<thead>
<tr>
<th>Time (Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 66 36 12 8 27 6 7 5 7 52 48</td>
</tr>
<tr>
<td>15 7 12 94 28 5 13 60 5 5 18 18</td>
</tr>
</tbody>
</table>

a) Find the mean and the standard deviation for the time to start for the new businesses.
b) Find the five number summary for the time to start for the new businesses.
c) Create a histogram for the length of all of the flowers.
d) Create a boxplot (modified) for the new businesses.

Solution

* reading in the data;
  data TimeStart;
    infile 'W:\PC-text\eg01-23time24.txt' delimeter = '09'x firstobs = 2;
    input country $ time;
  run;

*a) and b);
  proc univariate data = TimeStart;
    var time; /*this means that only the variable time will be analyzed */
    run;

*c);
  proc sgplot data = TimeStart;
    histogram time / binwidth = 10 ;
    /*The binwidth is the width of each box in the histogram
     For large data sets, there should be no more than 20 classes
     You will need to change the binwidth to get this number based on
     the range. See the information from proc univariate*/
  run ;

*d);
  data TimeStart1;
    set TimeStart;
    index = 1;
  run;

/*Explanation of what was done above:
We created another data set called 'TimeStart1'.
We set the original data set 'TimeStart' as the input file.
This means that all variables in 'TimeStart' remain in the new data set.*/
Another variable called 'index' was created and added to 'TimeStart1'. According to the code, every observation of 'index' has value '1'. This was done because the proc boxplot requires two variables, one variable has all of the quantitative values and the other categorical variable that indicates which category each of the values are in. If there is only one category, we need to create the second variable and arbitrarily indicate the label of the one category. */

proc boxplot data = TimeStart1;
  plot time * index/boxstyle=schematic idsymbol=circle;
  /* This creates a modified boxplot(s) of the score variable for each group in the group variable.
   Note, if there is only one group it will produce a single boxplot (the group variable is still required), if there are multiple groups it will create side-by-side boxplots
   boxstyle=schematic: modified boxplots, that is outliers are points
   idsymbol = circle: the outliers are circles.
   The diamond in the plot is the location of the mean.*/
run;

a) Find the mean and the standard deviation for the time to start for the new businesses.
b) Find the five number summary for the time to start for the new businesses.

Solution

The table below shows the descriptive statistics for the time to start for the new businesses:

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>23.6500</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>23.8286</td>
</tr>
<tr>
<td>Variance</td>
<td>567.8973</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2.075595</td>
</tr>
<tr>
<td>Uncorrected SS</td>
<td>2645.6</td>
</tr>
<tr>
<td>Corrected SS</td>
<td>1309.62</td>
</tr>
<tr>
<td>Coeff Variation</td>
<td>100.86247</td>
</tr>
<tr>
<td>Std Error Mean</td>
<td>4.864025</td>
</tr>
</tbody>
</table>

Though I provided all of the tables in this tutorial, you will lose points if you provide more tables than are required to explain your answer.
Mean = 23.625, Standard deviation = 23.82876
The five number summary:
Max = 94, Q₃ = 32, Median = 13, Q₁ = 7, Min = 5

2. Creating Histograms

c) Create a histogram for the length of all of the flowers.

Solution

Remember that the histogram should have the appropriate number of bins for small data sets.

\[
\text{number of bin} \times x \approx \sqrt{\text{number of data points}}
\]

For large data sets, the best number is around 10 – 20 or what 'looks good'.

The following shows the resulting histogram:

I strongly recommend that you change the size of the graph so that it fits better on the page.

3. Boxplots

d) Create a boxplot (modified) for the new businesses.

Solution:

Remember that you need a grouping variable even if there is no categorical variable in the problem.