SAS Tutorial for STAT 350 for Lab 2

Author: Leonore Findsen, Cheng Li

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:

13	66	36	12	8	27	6	7	5	7	52	48
15	7	12	94	28	5	13	60	5	5	18	18

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' delimiter = '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.
```

SAS Tutorial for STAT 350 for Lab 2

Author: Leonore Findsen, Cheng Li

```
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 SAS System

The UNIVARIATE Procedure
Variable: time

Moments			
N	24	Sum Weights	24
Mean	23.625	Sum Observations	567
Std Deviation	23.8287596	Variance	567.809783
Skewness	1.60081155	Kurtosis	2.07559957
Uncorrected SS	26455	Corrected SS	13059.625
Coeff Variation	100.862474	Std Error Mean	4.86402518

Basic Statistical Measures			
Location		Variability	
Mean	23.62500	Std Deviation	23.82876
Median	13.00000	Variance	567.80978
Mode	5.00000	Range	89.00000
		Interquartile Range	25.00000

Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	4.857088	Pr > |t|	<.0001
Sign	M	12	Pr >= |M|	<.0001
Signed Rank	S	150	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	94
99%	94
95%	66
90%	60
75% Q3	32
50% Median	13
25% Q1	7
10%	5
5%	5
1%	5
0% Min	5

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
5	22	48	12
5	21	52	11
5	18	60	20
5	9	66	2
6	7	94	16

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.

SAS Tutorial for STAT 350 for Lab 2

Author: Leonore Findsen, Cheng Li

Mean = 23.625, Standard deviation = 23.82876
The five number summary:
Max = 94, Q_3 = 32, Median = 13, Q_1 = 7, Min = 5

2. Creating Histograms

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

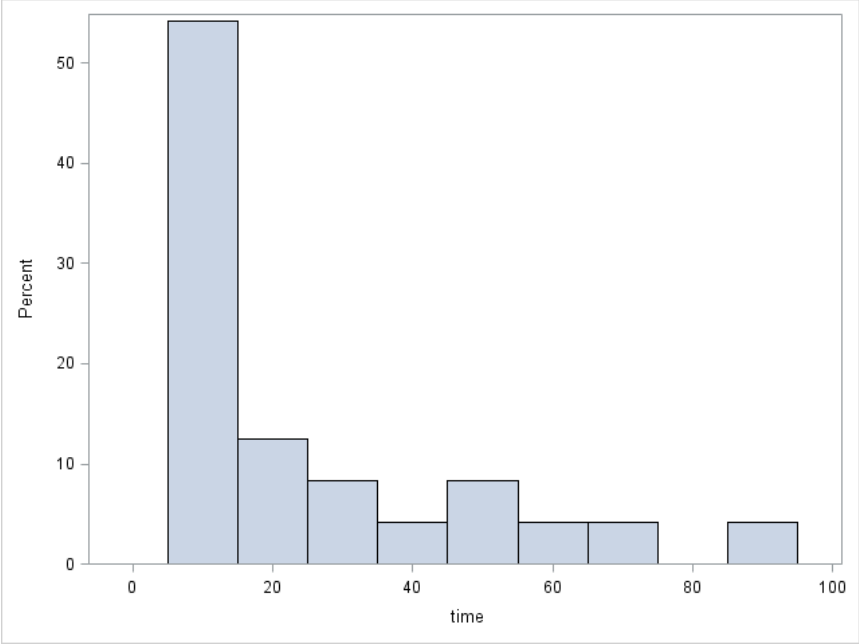
Solution

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

$$number\ of\ binx\ \approx\ \sqrt{number\ of\ data\ points}$$

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.

SAS Tutorial for STAT 350 for Lab 2

Author: Leonore Findsen, Cheng Li