SAS Tutorial for STAT 350 for Lab 2

Author: Leonore Findsen, Cheng Li

1. Numerical Summaries and Graphs

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 of the times it took to start a new business among all countries in the data set.
b) Find the five-number summary of the times it took to start a new business.
c) Create a histogram of the times it took to start a new business.
d) Do you think the median is close to the mean?
e) Create a boxplot (modified) of the times it took to start the new businesses.

Solution

In your solution, please only list the complete code at the beginning of the question. In this tutorial, the code will also be repeated in each part so that it can be explained more fully. Please do NOT repeat code in your lab report unless explicitly stated. Note: An easy check for typos is to be sure that all of the colors are correct in the SAS Editor.

```
/* Reading in the data by code;
If you are using code to read the data, ALWAYS read in all of the
variables whether you use them or not. */
data TimeStart;
  infile 'W:\eg01-23time24.txt' delimiter = '09'x firstobs = 2;
  input country $ time;
run;

/* Reading in the data by GUI procedure;
File --> Import Data --> Tab Delimited File (.txt) --> Next --> browse
for the file --> Library: Work, Member: TimeStart (or an appropriate
name) --> Finish */

*a) and b);
proc univariate data = TimeStart;
  var time;
* The above command means that only the variable time will be analyzed;
run;
```

```
*c);
proc sgplot data = TimeStart;
  histogram time;
/* You may add the "binwidth" option if you want to modify the default
  value. You can play around with it until you find the plot is optimal
  for your analysis purposes.  Make sure the number of bins is neither
  too small nor too big. For small to medium data sets, the number of
  bins should be approximately equal to the square root of the number
  of data points. */

/* For example, you can use the following 'histogram' statement for a
  histogram with 14 bins:
  histogram time / binwidth = 14;
  (I changed the colors to the correct ones) */

  density time;
   /* This adds the probability density curve of a THEORETICAL normal
    distribution using the mean and standard deviation from the data */
  density time / type=kernel;
   /* This adds the smoothed kernel density curve, representing the
     overall shape of the distribution of the ACTUAL data */
run;

*e);
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 "boxplot procedure" (i.e., proc boxplot)
requires two variables: the first variable has all the quantitative
values and the second variable is which indicates which category each
of the quantitative values are in. If there is only one category, we
still need to create the second variable and arbitrarily indicate the
label of the one category as something. */

proc boxplot data = TimeStart1;
  plot time * index/boxstyle = schematic idsymbol = circle;
/* This creates a modified boxplot(s) of the response variable for each
group in the categorical 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.
Details about the options:
  boxstyle = schematic: produce a modified boxplots, that the outliers
                        are points
  idsymbol = circle: the outliers are circles.
  The diamond in the plot is the location of the mean.*/
run;
```

SAS Tutorial for STAT 350 for Lab 2

Author: Leonore Findsen, Cheng Li

Now, we will answer the questions presented above and explain the commands in more detail.

a) Find the mean and the standard deviation of the times it took to start a new business among all countries in the data set.
b) Find the five-number summary of the times it took to start a new business.

Solution

```
*a) and b);
proc univariate data = TimeStart;
  var time;
* The above command means that only the variable time will be analyzed;
run;
```

In SAS, most pre-defined functions are defined as procedures (proc). The general form for each procedure is:

```
proc ProcedureName data = DataName <other options if needed>;
  other statements;
run;
```

Remember to end each command line with a semicolon, ";."

proc univariate is concerned with the statistics of one variable. The var statement allows you to specify multiple variables, and the analysis will be conducted for each variable, respectively.

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

SAS Tutorial for STAT 350 for Lab 2

Author: Leonore Findsen, Cheng Li

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:
Min = 5, Q_1 = 7, Median = 13, Q_3 = 32, Max = 94

Note: There are other ways of computing the five-number summary. However, this method will generate the values that are obtained using the method in our textbook.

2. Creating Histograms

```
*c);
proc sgplot data = TimeStart;
  histogram time;
/* You may add the "binwidth" option if you want to modify the default
  value. You can play around with it until you find the plot is optimal
  for your analysis purposes.  Make sure the number of bins is neither
  too small nor too big. For small to medium data sets, the number of
  bins should be approximately equal to the square root of the number
  of data points. */

/* For example, you can use the following 'histogram' statement for a
  histogram with 14 bins:
  histogram time / binwidth = 14;
  (I changed the colors to the correct ones) */

  density time;
   /* This adds the probability density curve of a THEORETICAL normal
    distribution using the mean and standard deviation from the data */
  density time / type=kernel;
   /* This adds the smoothed kernel density curve, representing the
     overall shape of the distribution of the ACTUAL data */
run;
```

proc sgplot is a plotting procedure. SG stands for Statistical Graphics.

The statement to create a histogram is `histogram VariableName;`

SAS is usually pretty good at getting the correct binwidth; however, the code to change the binwidth is provided. I have put in the correct colors so if you use the option you will know if it is correct.

The `density` statement adds in additional curves which will be used in later labs.

SAS Tutorial for STAT 350 for Lab 2

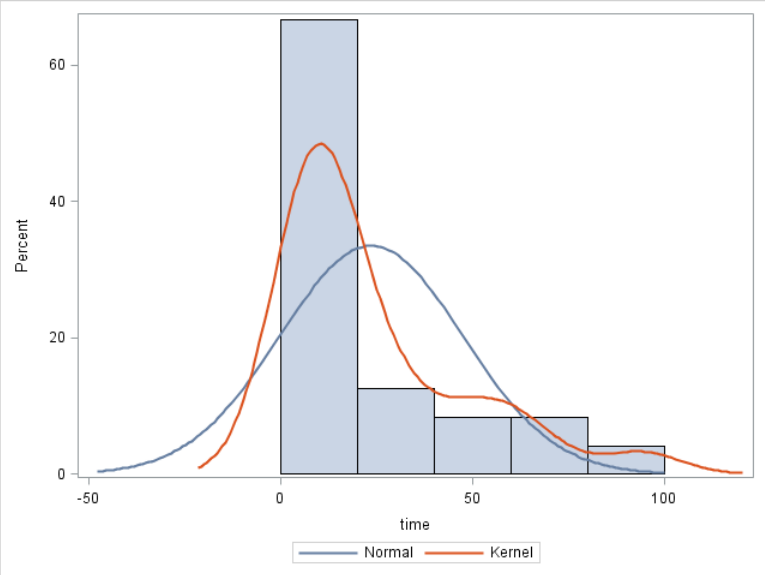
Author: Leonore Findsen, Cheng Li

c) Create a histogram of the times it took to start a new business.

Solution

Remember that the histogram should have the appropriate number of bins.
$number\ of\ bins \approx \sqrt{number\ of\ data\ points}$

The following shows the resulting histogram using the default option. Note that since there are 24 rows, we should have $\sqrt{24} = 4.899 \approx 5$ bins.



The red line is a smoothed curve representing the histogram which is called a Kernel. The blue line is the normal approximation to the histogram. We will revisit these lines in future labs.

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

d) Do you think the median is close to the mean?

Solution

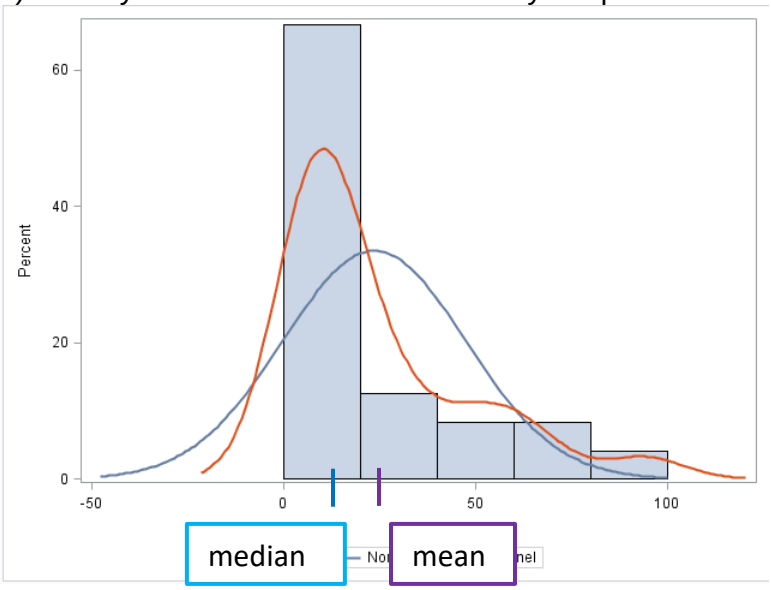
There are a number of different ways that this can be accomplished.

1) Compare the numbers

From parts a) and b), Mean = 23.625, Median = 13.
I would say that these numbers are not close in this data set because their absolute difference is 10.625 which is large compare to the range of data (approximately 100).

Author: Leonore Findsen, Cheng Li

2) Visually look at the numbers when they are plotted on the histogram.



To me, these numbers do not look close to each other.

3) Compare the difference of the numbers to the total spread of the numbers

$$\frac{mean - median}{maximum - minimum} = \frac{23.625 - 13}{94 - 5} = 0.119$$

This is ~12% which is fairly large.

4) Compare the difference of the numbers to the standard deviation

$$\frac{mean - median}{standard\ deviation} = \frac{23.625 - 13}{23.82876} = 0.446$$

Therefore the difference is about half the standard deviation which is fairly large.

In this case, I would say that the two numbers are not close to each other. However, these four methods will not always provide the same answer. When answering this question, you need to look at all of the information to make a decision. In some cases, it is possible for different people to have different answers.

3. Boxplots

The following is the procedure for generating a modified boxplot. In a modified boxplot, the outliers are explicitly plotted.

Author: Leonore Findsen, Cheng Li

```
*e);
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 "boxplot procedure" (i.e., proc boxplot)
requires two variables: the first variable has all the quantitative
values and the second variable is which indicates which category each
of the quantitative values are in. If there is only one category, we
still need to create the second variable and arbitrarily indicate the
label of the one category as something. */

proc boxplot data = TimeStart1;
  plot time * index/boxstyle = schematic idsymbol = circle;
/* This creates a modified boxplot(s) of the response variable for each
group in the categorical 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.
Details about the options:
  boxstyle = schematic: produce a modified boxplots, that the outliers
                        are points
  idsymbol = circle: the outliers are circles.
  The diamond in the plot is the location of the mean.*/
run;
```

The **data** statement can be used for more things than just reading in variables. It is also used when you want to add variables to the data set and do simple arithmetic to the data set. I always change the name of the new dataset so that I don't overwrite or get confused with the old one. When you are doing that, remember to always be sure which data set is used for the procedures that follow. The `set` statement loads in a data set. In this case, we need to add a new categorical variable. Please read the comments above for why SAS requires that we do that.

The procedure to makes boxplots is called **boxplot**. To make the boxplot, we use the `plot` statement. The order of the variables is `numeric * qualitative`. The explanations of the options are above.

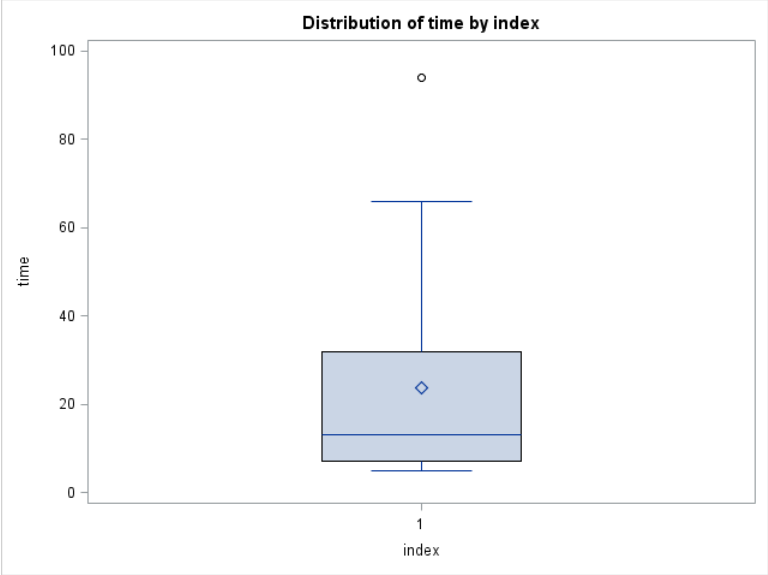
e) Create a boxplot (modified) of the times it took to start the new businesses.

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



The diamond on the boxplot is the location of the mean.

Note: Please resize your graphics so that they fit on the page especially when you have more than one plot.