SPSS and R Examples of Descriptive Statistics of Psychological Measurement Evaluations

These are the course evaluations I received from my Psychological Measurement class. Below is a list of the questions. Responses were made on a 6-point scale, from 1=strongly disagree to 6=strongly agree. For the analyses shown below, I computed an average of the 11 questions (EVALAVE).

1. The instructor uses class time effectively.
2. The instructor answered questions in a clear and understandable way.
3. The instructor was prepared for class.
4. The instructor used examples effectively in explaining difficult concepts.
5. The instructor raised challenging questions for discussion.
6. The instructor welcomed students questions or comments.
7. The instructor seemed to enjoy teaching.
8. The grading policies were clearly stated.
9. The course material was presented clearly.
10. Relevant information in addition to the textbook was presented.
11. I learned a lot in this course.

SPSS

output close *.
*The above command clears the output file (notice asterisk above refers to current output file, which differs from its use in the present line which indicates that everything until the period is ignored by SPSS).

get file='c:\jason\spsswin\uvclass\pmclass evals.sav'.

compute evalave=mean(q1 to q11).

frequencies vars=evalave
/histogram=normal.

descriptives vars=evalave
/statistics=all.

Frequencies

<table>
<thead>
<tr>
<th>Statistics</th>
<th>evalave</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>15</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>evalave</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>1.73</td>
<td>6.7</td>
<td>6.7</td>
<td>6.7</td>
</tr>
<tr>
<td>1.82</td>
<td>1</td>
<td>6.7</td>
<td>6.7</td>
<td>13.3</td>
</tr>
<tr>
<td>5.00</td>
<td>1</td>
<td>6.7</td>
<td>6.7</td>
<td>20.0</td>
</tr>
<tr>
<td>5.40</td>
<td>1</td>
<td>6.7</td>
<td>6.7</td>
<td>26.7</td>
</tr>
<tr>
<td>5.55</td>
<td>1</td>
<td>6.7</td>
<td>6.7</td>
<td>33.3</td>
</tr>
<tr>
<td>5.64</td>
<td>3</td>
<td>20.0</td>
<td>20.0</td>
<td>53.3</td>
</tr>
<tr>
<td>5.70</td>
<td>1</td>
<td>6.7</td>
<td>6.7</td>
<td>60.0</td>
</tr>
<tr>
<td>5.73</td>
<td>2</td>
<td>13.3</td>
<td>13.3</td>
<td>73.3</td>
</tr>
<tr>
<td>5.82</td>
<td>2</td>
<td>13.3</td>
<td>13.3</td>
<td>86.7</td>
</tr>
<tr>
<td>5.91</td>
<td>1</td>
<td>6.7</td>
<td>6.7</td>
<td>93.3</td>
</tr>
<tr>
<td>6.00</td>
<td>1</td>
<td>6.7</td>
<td>6.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>
Descriptives

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Sum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Std. Error</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>evalave</td>
<td>15</td>
<td>4.27</td>
<td>1.73</td>
<td>6.00</td>
<td>77.10</td>
<td>5.1400</td>
<td>.35808</td>
<td>1.38683</td>
<td>-2.276</td>
<td>.580</td>
<td>3.933</td>
<td>1.121</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>15</td>
<td>4.27</td>
<td>1.73</td>
<td>6.00</td>
<td>77.10</td>
<td>5.1400</td>
<td>.35808</td>
<td>1.38683</td>
<td>-2.276</td>
<td>.580</td>
<td>3.933</td>
<td>1.121</td>
</tr>
</tbody>
</table>

I encourage you to begin using syntax as much as possible, but the above output can also be obtained through the windows: Analyze -> Descriptive Statistics -> Frequencies, then move over the desired variable to the Variable(s) box and click "ok."

R

```r
#clear active frame from previous analyses
rm(mydata)

#note: all lines starting with hashtag/lb sign are ignored by R

> #I read an SPSS data file (.sav) in with the haven package, which must be installed
> #on the first used on a computer. It must be loaded for use after that using the library() function
> #the lessR package allows you to perform some R commands simply, including reading in SPSS data
> #install.packages("haven")
> 
> library(haven)
> mydata = read_sav("c:/jason/spsswin/uvclass/pmclass evals.sav")
> 
> #the lessR package can also be used
> #library(lessR)
> #mydata = Read("c:/jason/spsswin/uvclass/pmclass evals.sav",quiet=TRUE)
> 
> #two ways to look at the data set
> #fix(mydata) #allows you to edit the data
> #View(mydata) #data view-like appearance
> 
> #compute average of questions using the lessR package's Transform function
> #recently data=mydata required, but did not used to be
> #note that this code produces different mean and sd from SPSS because of missing data
> library(lessR)
> mydata = Transform(evalave=(q1+q2+q3+q4+q5+q6+q7+q8+q9+q10+q11)/11, data=mydata)
```

Number of variables of mydata to transform: 1
Number of cases (rows) of mydata: 15

Transformation
create new variable: evalave
> # the freq function in the summarytools package has a nice frequency table
> # I ask for a frequency table only for the evalave variable here,
> # by prefixing the variable name with the data set plus $
> # but using freq(mydata) gives tables for all of the variables
> library(summarytools)
> freq(mydata$evalave)

Frequencies
mydata$evalave
Type: Numeric

<table>
<thead>
<tr>
<th>Freq</th>
<th>% Valid</th>
<th>% Valid Cum.</th>
<th>% Total</th>
<th>% Total Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.72727272727273</td>
<td>1</td>
<td>7.69</td>
<td>6.67</td>
<td>6.67</td>
</tr>
<tr>
<td>1.81818181818182</td>
<td>1</td>
<td>7.69</td>
<td>15.38</td>
<td>13.33</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>7.69</td>
<td>23.08</td>
<td>20.00</td>
</tr>
<tr>
<td>5.54545454545455</td>
<td>1</td>
<td>7.69</td>
<td>30.77</td>
<td>26.67</td>
</tr>
<tr>
<td>5.63636363636364</td>
<td>3</td>
<td>23.08</td>
<td>53.85</td>
<td>46.67</td>
</tr>
<tr>
<td>5.72727272727273</td>
<td>2</td>
<td>15.38</td>
<td>69.23</td>
<td>60.00</td>
</tr>
<tr>
<td>5.81818181818182</td>
<td>2</td>
<td>15.38</td>
<td>84.62</td>
<td>73.33</td>
</tr>
<tr>
<td>5.90909090909091</td>
<td>1</td>
<td>7.69</td>
<td>92.31</td>
<td>80.00</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>7.69</td>
<td>100.00</td>
<td>86.67</td>
</tr>
<tr>
<td>&lt;NA&gt;</td>
<td>2</td>
<td></td>
<td>13.33</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

> # obtain frequency histogram from lessR package of just evalave variable
> Histogram(evalave)

Histogram

Density plot

> Density(evalave)
Let's see what the distribution looks like after eliminating the two outliers. Ordinarily, you would not want to remove outliers from a data set unless you have a good reason to do so (e.g., there is an error, the participant did not belong in the study to begin with—did not meet study inclusion criteria). This time, we will add a boxplot and stem-and-leaf plot.

**SPSS**

*the following statement selects out the two outliers.*

```plaintext
select if evalave gt 4.

frequencies vars=evalave  
    /histogram=normal.

descriptives vars=evalave  
    /statistics=all.

examine variables=evalave  
    /plot=boxplot stemleaf  
    /statistics=none.
```

**evalave Stem-and-Leaf Plot**

<table>
<thead>
<tr>
<th>Evaluations</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.00</td>
<td>1</td>
<td>7.7</td>
<td>7.7</td>
<td>7.7</td>
</tr>
<tr>
<td>5.40</td>
<td>1</td>
<td>7.7</td>
<td>7.7</td>
<td>15.4</td>
</tr>
<tr>
<td>5.55</td>
<td>1</td>
<td>7.7</td>
<td>7.7</td>
<td>23.1</td>
</tr>
<tr>
<td>5.64</td>
<td>3</td>
<td>23.1</td>
<td>23.1</td>
<td>46.2</td>
</tr>
<tr>
<td>5.70</td>
<td>1</td>
<td>7.7</td>
<td>7.7</td>
<td>53.8</td>
</tr>
<tr>
<td>5.73</td>
<td>2</td>
<td>15.4</td>
<td>15.4</td>
<td>69.2</td>
</tr>
<tr>
<td>5.82</td>
<td>2</td>
<td>15.4</td>
<td>15.4</td>
<td>84.6</td>
</tr>
<tr>
<td>5.91</td>
<td>1</td>
<td>7.7</td>
<td>7.7</td>
<td>92.3</td>
</tr>
<tr>
<td>6.00</td>
<td>1</td>
<td>7.7</td>
<td>7.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

**Histogram**

**Descriptive Statistics**

<table>
<thead>
<tr>
<th>Evaluations</th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Sum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid N (listwise)</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**evalave Stem-and-Leaf Plot**

Frequency Stem & Leaf

1.00 Extremes (=<5.0)

1.00 5 . 4
10.00 5 . 5666777889
1.00 6 . 0

Stem width: 1.00
Each leaf: 1 case(s)
The R commands are the same as I illustrated above. To save space I just show how to remove cases and the Boxplot and stem function.

```r
#select non-extreme cases and obtain a boxplot and stem-and-leaf using the lessR subset function
> mydata = Subset(evalave > 4, data=mydata)

#BoxPlot is a lessR function
> BoxPlot(evalave)

--- evalave ---

Present: 11
Missing: 0
Total : 11

Minimum : 5.000000
Lower Whisker: 5.5454545
Lower Hinge : 5.6363636
Median     : 5.7272727
Upper Hinge : 5.8181818
Upper Whisker: 6.0000000
Maximum    : 6.0000000

1st Quartile : 5.636364
3rd Quartile : 5.818182
IQR          : 0.1818182
(Box plot) Outliers: 1
Small: 5
Large: none
```
Sample Write-up [based on second analysis]

Descriptive statistics were obtained for the course evaluation measure. Results indicated high ratings of the instructor overall ($M = 5.66$ and $SD = .25$). The range was equal to 1.00, with minimum score of 5.00 and a maximum score of 6.00. The distribution was skewed to the left and leptokurtic. The skewness value was equal to -1.49 and the kurtosis value was equal to 3.40. A frequency histogram, using a normal curve overlay and a boxplot, were obtained. Both plots suggested an outlier with a value of approximately 5.00. The boxplot depicted that the median was approximately 5.7, the range was between approximately 5.50 and 6.0 (without the outlier), and 50% of the scores occurred between approximately 5.65 and 5.85.$^1$

[Note that descriptive statistics are often relegated to a table and/or are very briefly given in the description of the sample, where the descriptive statistics are given just for sociodemographic variables. Percents are usually used for categorical variables (e.g., 53% 11th graders, 47% 12th graders) and mean and standard deviations for continuous variables (e.g., instructor ratings as shown above) and there is little or no discussion of the shape of the distribution].

---

$^1$ Note: the R output gives the exact values for the median and the 25 and 75 percentile, which could be reported instead. Remember that SPSS uses hinges and rounded median, and R may compute these more exactly for the box plot. The median is not available from the Descriptives command in SPSS, but you can obtain it from the Frequencies command with an added subcommand:

```
frequencies vars=evalave
/histogram=normal
/statistics=median.
```