Chat 1
Notes 1

Descriptive and Inferential Statistics

Descriptive:

describe data, relationships among variables, and differences among groups

Inferential:

(a) Applied only to samples of data, not to population data

(b) tests which help determine whether relationships or differences found with descriptive statistics could be the result of random chance. If random chance seems unlikely, then we conclude that the relationships or differences are real, i.e., could be found in the population/census data.

Statistics vs. Parameters:

(a) Parameters apply to populations, examples include

- mean (μ) age of everyone in this class (we define this class as a population),
- standard deviation (σ) of age in this class, and
- variance (σ^2) of age in this class.
- Note use of Greek symbols for population parameters.

(b) Statistics are estimates of population parameters, for example

- mean (M or $\bar{X}$) age for a sample of students in this class (not everyone; at least 1 less than all students in this class is a sample),
- standard deviation (s or SD) of age in this class, and
- variance (s^2 or VAR) of age in this class.
- Note use of Roman symbols for sample statistics.
- In statistical inference we attempt to infer the value of a population parameter from its corresponding statistic.

Some differences:

(a) There will be no hypothesis testing – no inferential statistics – if one is working with a census (population data) – no test statistics (t-ratios, F-ratios), p-values, confidence intervals, or standard errors. Why – because inferential statistics are only applied to sample data in an effort to infer to the population parameters.

(b) Some small differences in formulas between statistics and parameters, e.g.,

Variance – divide by n-1 with sample, but by n with census

Variables

Anything that has more than one unique category; **variables** have multiple categories.

Examples:

Age in years = 1, 2, 3, 4, 5, 6, …, 48, 49, 50, etc.

Test Scores = SAT verbal ranges from 200 to 800; 200, 201, 202, 203, etc.

Sex: Biological distinction with Female (XX) and Male (XY) being most common categories of sex

Gender: APA Manual explains that gender is a psychological state often measured on scale that ranges from feminine to masculine with multiple steps between these anchor points; gender is not a biological distinction between sexes.

If only one category is present, then one has a ***constant*** rather than a variable

Example: If everyone in class is female, then sex is a constant, no sex variability

Students in a classroom (variable or constant?):

- Age = all 21 years of age
- GPA = ranges between 1.65 and 3.86
- Race = Asian, Black, Hispanic
- Transportation to class = Walk

Age = constant
GPA = variable
Race = variable
Transportation to class = constant

What are the variables in these hypotheses?

1. There is no difference in Body Mass Index (BMI) between females and males?

Two variables: Sex and BMI

Why are female and male not variables?

These are categories of the variable sex. It is easy to confuse categories with variables, so watch for this when identifying variables or writing hypotheses.

2. The higher one's level of academic self-efficacy, the lower will be one's test anxiety.

(Note that academic self-efficacy and test anxiety are measured on a 20-point scale ranging from 1 = low to 20 = high.)

Two variables: Academic Self-efficacy and Test Anxiety

Measurement

Process of assigning labels to categories of a variable.

Questionnaire Item:

What is your sex?
Female ____
Male ____
Etc.

What is your age in years? _____
What was your pretax income last year? _____

Scales of Measurement

Nominal: has only unranked categories (i.e., no inherent rank to categories)

Examples: sex, race, type of flower

Ordinal: categories with inherent rank, i.e., ranked categories (i.e., this makes it easy to sort categories from high to low, more to less, etc.)

Examples:

Questionnaire Item – rate instructor on the following dimensions

(a) The instructor's content was well organized:

Strongly disagree
Disagree
Somewhat agree
Agree
Strongly Agree

(b) The instructor presented material in a clear manner.

Use scale above

(c) The instructor was open to student questions, comments, and concerns

Use scale above

SES – socio-economic status (originally measured by three indicators: educational level, income, and occupational prestige)

High
Middle
Low

Interval: ranked categories with equal distance between measuring categories or measuring device– sticking point is lack of true zero point, few variables with equal interval have no zero point

Ratio: same as interval, but also has a true zero point (a true start or end point); allows for formation of ratios

Examples:

- time to complete a task
- counting objects in a box
- number of points scored during a game

Unique attribute of ratio variables is that the can form ratios:

- He took twice as long to complete that task as she
- 30 seconds to complete task vs. 15 seconds, 30/15 = 2 (twice as long)

Equal interval characteristics – this is a function of the measuring scale used, not of the categories themselves

Examples of measure scales that produce equal intervals:

- Ruler in millimeters or inches,
- stop watch to record in seconds,
- counts of number of items scored correctly on tests,
- percentage of items scored correctly on tests

Test 1:
Bryan = 45%
Miriam = 85%
Melinda = 100%
Lakee = 90%
Elizabeth = 80%

Ratio = 45/100 = .45*100 = 45%

Types of Variables

(a) Qualitative Variable

Nominal or categorical (i.e., no inherent rank to categories), or ordinal variable with limited number of categories (e.g., SES with three categories of low, middle, high --- treat this as qualitative or nominal for convenience of statistical analysis)

Examples

Sex
Race
Types of flowers

(b) Quantitative Variable

Ranked categories (ordinal, interval, or ratio, assuming the ordinal measure contains many ranked categories)

Examples

Number of test items answered correctly
Weight in lbs.
Number of pages read over the summer

Qualitative = nominal (categorical) variables (those with only categories that cannot be ranked)

Example = sex and race

Quantitative = any that is not nominal or any variable that has a number of ranked categories (more than 3 or 4 categories); or any interval, ratio, and some ordinal variables.

Examples

- age,
- weight,
- score on EDUR 8131 Test 1,
- tally or count of misspoken words during chat;
- one's rating to this item: "The instructor's content was well organized" with ratings ranging from 1 = "strongly disagree" to 5 = "strongly agree"

Summary
Nominal, Categorical = Qual,
Ordinal = Quan (sometimes treated as qual if categories are few)
Interval = Quan
Ratio = Quan

(c) Independent (Predictor) Variable (IV)

That which precedes the dependent variable in time and is expected to predict or influence the dependent variable

(d) Dependent (Criterion) Variable (DV)

That which follows the IV in time and is expected to be predicted or influenced by the independent variable

IV = variable that comes first in time sequence
DV = variable that follows IV in time sequence

Examples – find the IV and DV in these hypotheses and determine whether the variables are Quantitative or Qualitative

Example 1:

There will be a difference in math scores between males and females.

What are the variables, and which are IV and DV, and are they Qual or Quan variables?

Answer

IV = sex (female/male), Qual
DV = math scores, Quan
Reason = one's sex precedes math performance in time

Example 2:

Class size and student final test scores are not related.

What are the IVs and DVs, and is the IV qual or quan?

Answer

IV = class size, Quan
DV = student final test scores, Quan

Example 3:

Students whose parents are educators will earn higher scores on a test than students whose parents are not educators.

What are the IVs and DVs, and is the IV qual or quan?

Answer

IV = occupation of parents, Qual
DV = test scores, Quan

Example 4:

For females in public schools, researchers found that one's mathematics attitude predicts well one's mathematics achievement.

(Note that mathematics attitude is measured on a 20-point scale ranging from 1 = low to 20 = high.)

Which are the IV and DV, and are they Qual or Quan variables?

Answer

Variables -
IV = mathematics attitude
DV = mathematics achievement

Constants –
Sex: because there is only one category, female
School Setting: only public schools included, nonpublic schools not included in hypothesis

---

Below are
(a) Additional Hypothesis Examples (Study on your own)
(b) SPSS screen shots show how to obtain (these are also covered in Chat 2)

1. **Measures Central Tendency**
2. **Measures Variability**

3. **Frequency Displays**
4. **Percentile Ranks obtained from Frequency Displays**
5. **Quartiles**
6. **Box Plots (Box-and-Whisker Plots)**
7. **Stem-and-Leaf Displays**

(a) Additional Examples - Study on your own (skip in chat)

I. The more time (measured in hours) spent studying, the greater will be one's final exam test score (note that scores range from 0% to 100%, percent correct).

IV = time spent studying
DV = final exam scores

Time spent studying, qual or quant?
Final exam test score, qual or quant?

time spent studying → (how measured – in hours) quan
final exam scores (measured in percent correct) → quan

II. The greater one's mathematics self-efficacy, the greater will be one's math test performance. Note that self-efficacy is often measured as the level of one's confidence to successfully complete a task before that task is undertaken.

IV = mathematics self-efficacy
DV = math test performance

How does one measure mathematics self-efficacy? Normally through a self-administered questionnaire that contains usually 3 to 30 items, like this:

	Strongly Disagree	Disagree	Somewhat Agree	Agree	Strongly Agree
1. I am sure that I can complete even the most difficult math problems I will see in this course.	1	2	3	4	5
2. Even when math is difficult, I will persist until I find a solution.	1	2	3	4	5
3. I enjoy working challenging mathematics problems.	1	2	3	4	5

Student A answers with 4, 5, and 5.
Student B answers with 1, 2, and 1.

What are their math self-efficacy scores?

Student A = 4 + 5 + 5 = 14 (out of max possible for 15)
Student B = 1 + 2 + 1 = 4 (out of minimum possible of 3 with maximum of 15)

Math. Self-efficacy, qual or quant?

Math test performance, qual or quant?

Math. Self-efficacy -> quant
Math test performance -> quant

III. There is a negative association between academic self-efficacy and test anxiety

What are the variables in this hypothesis?

Two: academic self-efficacy and test anxiety

Which is IV and DV?

(a) Is this possible – the more anxious one about a test, the lower will be their confidence (and hence efficacy) in doing well on the test? If yes, which is IV and DV?

Yes, so test anxiety could be the IV

(b) Is this possible – the more confidence one has about a topic, the less anxious that person will be about upcoming tests? If yes, which is the IV and which is the DV?

Yes, so academic self-efficacy could be the IV

This is an example where it is not clear which is IV and DV, so more information would be needed; we would need to know the theory driving the study or the study design in order to determine IV and DV.

(b) SPSS screen shots for obtaining results for the following topics (won't be covered during chat unless requested)

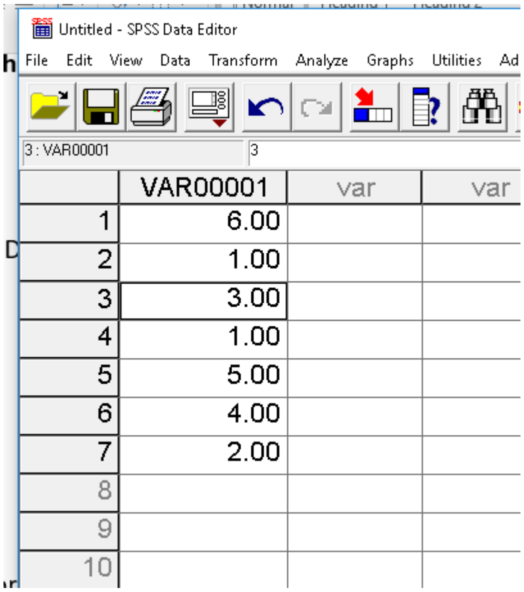
1. Measures Central Tendency
2. Measures Variability
3. Frequency Displays
4. Percentile Ranks obtained from Frequency Displays
5. Quartiles
6. Box Plots (Box-and-Whisker Plots)
7. Stem-and-Leaf Displays

Central Tendency (in SPSS)

Example: 6, 1, 3, 1, 5, 4, 2

Find mean (M), median (Md), and mode (Mo) for these scores.

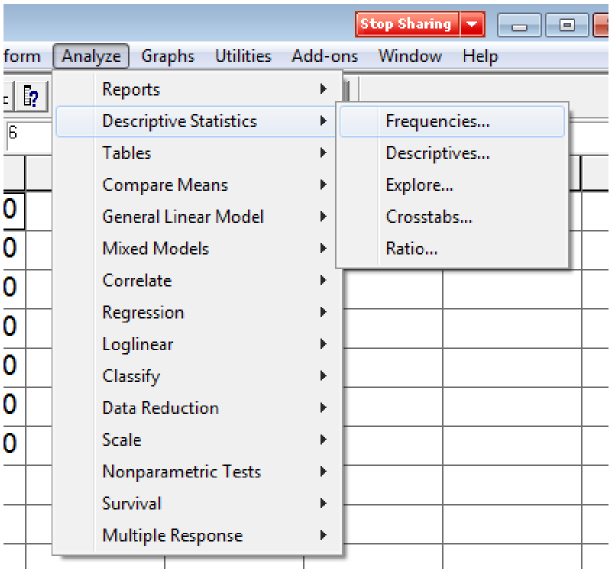
SPSS Data Entry

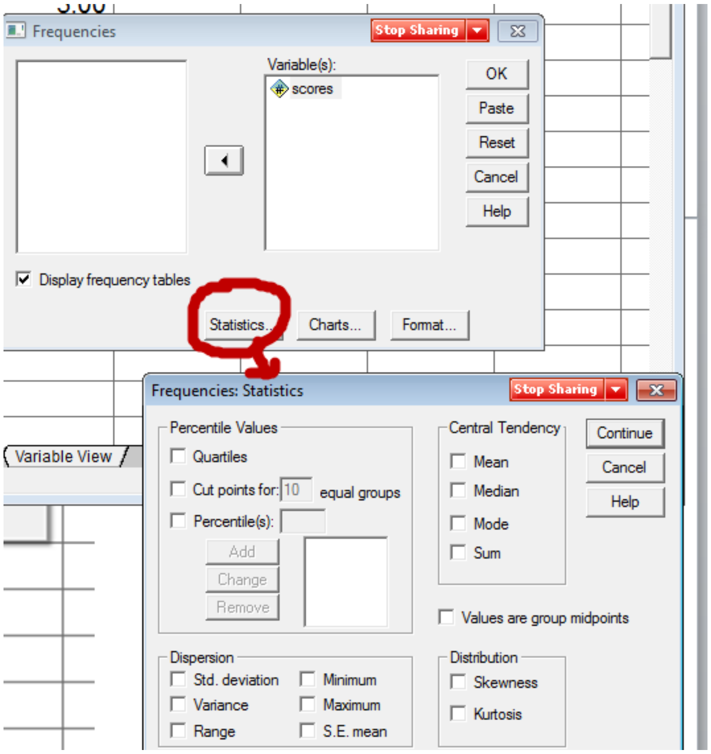

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Ad

3 : VAR00001 3

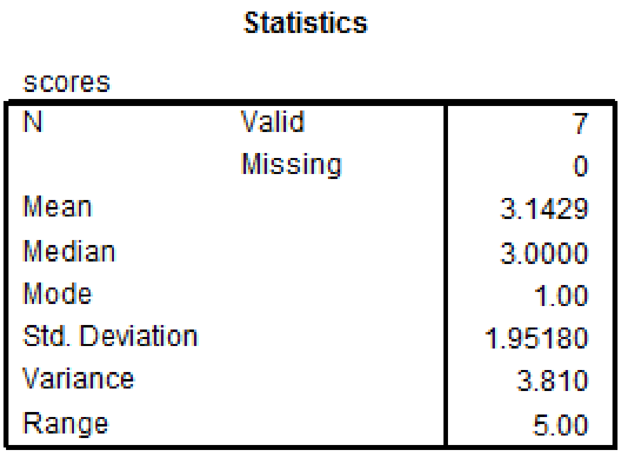
	VAR00001	var	var
1	6.00		
2	1.00		
3	3.00		
4	1.00		
5	5.00		
6	4.00		
7	2.00		
8			
9			
10			

SPSS Frequency Command





From SPSS



Statistics

scores

N	Valid	7
	Missing	0
Mean		3.1429
Median		3.0000
Mode		1.00
Std. Deviation		1.95180
Variance		3.810
Range		5.00

Mean = 3.1429

Median = 1 1 2 **3** 4 5 6

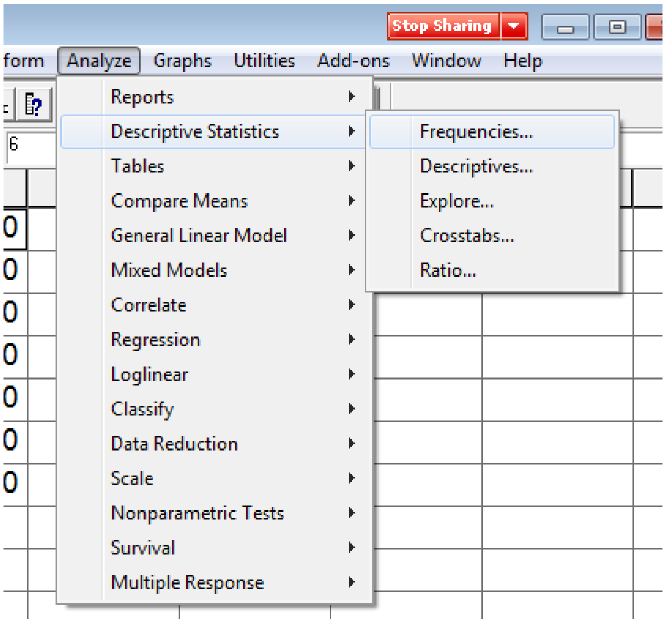
Mode = 1

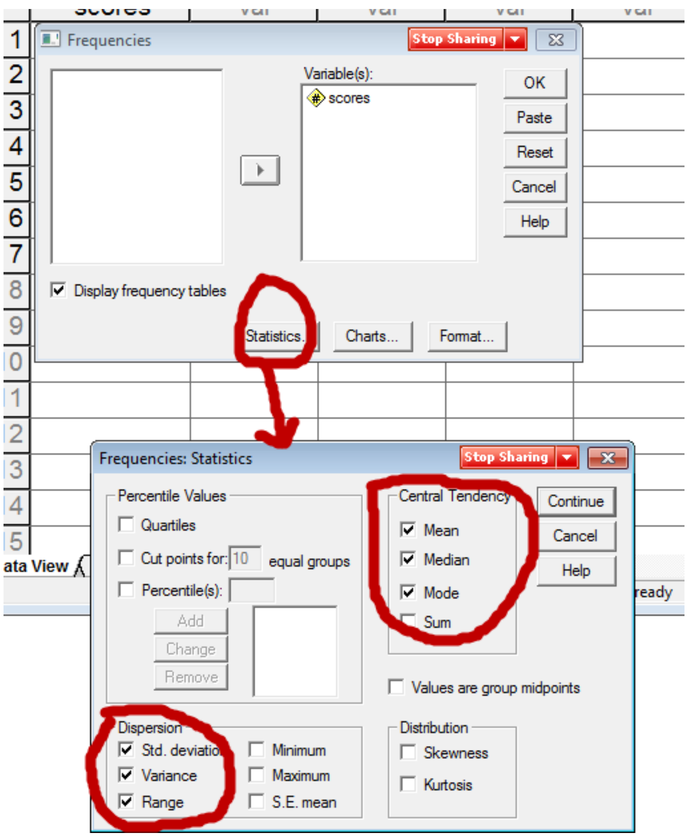
Variability, Dispersion (in SPSS)

Example: 6, 1, 3, 1, 5, 4, 2

Find range (R), variance (VAR), and standard deviation (SD) for the above scores.

SPSS Commands
SPSS Frequency Command





Variable(s): scores

☑ Display frequency tables

Percentile Values

☐ Quartiles

☐ Cut points for: 10 equal groups

☐ Percentile(s):

Central Tendency

☑ Mean

☑ Median

☑ Mode

☐ Sum

☐ Values are group midpoints

Dispersion

☑ Std. deviation

☑ Variance

☑ Range

☐ Minimum

☐ Maximum

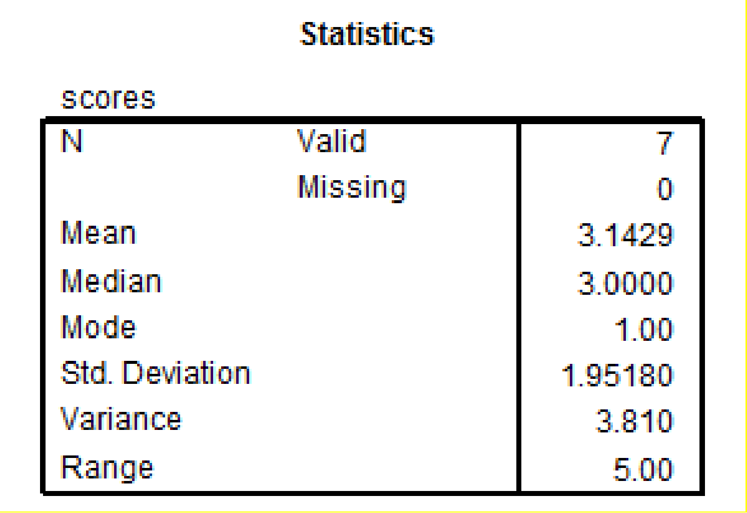
☐ S.E. mean

Distribution

☐ Skewness

☐ Kurtosis

From SPSS



Statistics

scores

N	Valid	7
	Missing	0
Mean		3.1429
Median		3.0000
Mode		1.00
Std. Deviation		1.95180
Variance		3.810
Range		5.00

Range = 5.00
Variance = 3.81
Standard Deviation = 1.9518

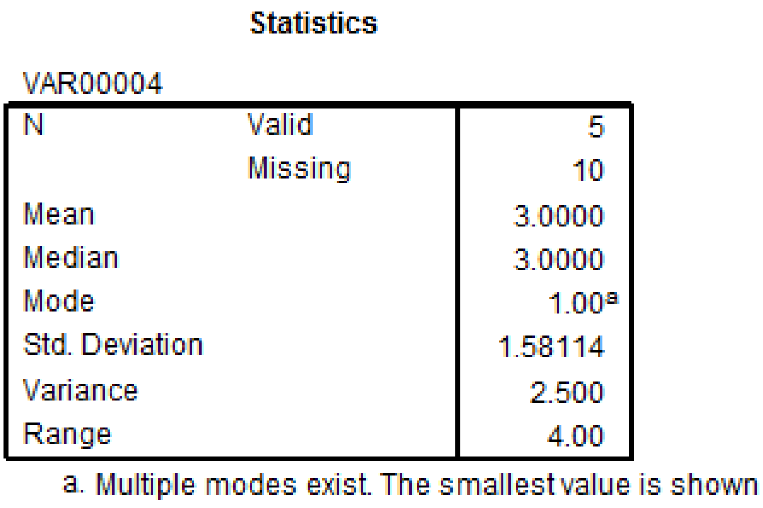
What does the SD represent?

Approximate mean (or average) of how far raw scores (X) deviate from the mean (M).

Example of deviation scores, SD is rough approximation to deviation score average (mean)

Scores	Mean	Deviations Scores (DS)	Squared DS
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
	Sum	0	10 = sums of squares (SS)
		VAR = SS/(n-1)	= 10 / 4 = 2.5
		SD = Square root of VAR	= SQRT(2.5) =1.581

VAR and SD from SPSS shown below



Statistics

VAR00004

N	Valid	5
	Missing	10
Mean		3.0000
Median		3.0000
Mode		1.00[a]
Std. Deviation		1.58114
Variance		2.500
Range		4.00

[a] Multiple modes exist. The smallest value is shown

Frequencies and Percentile Ranks

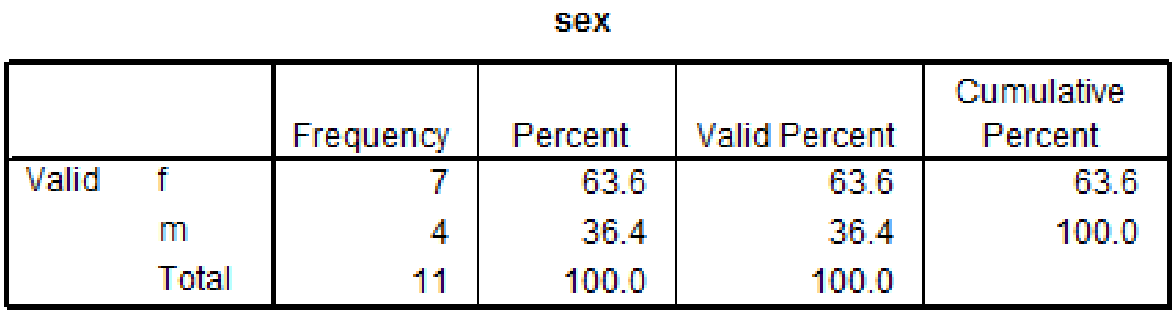
Variable is Sex of students in class. Below is a sample of students in class:

M, M, F, F, F, F, M, F, M, F, F

Sex	Freq	Relative Freq.
Females	7	.6363 (64%)
Males	4	.3636 (36%)

N = 11 students,
7/11 = .6363,
4 / 11 = .3636

Example of sex with SPSS



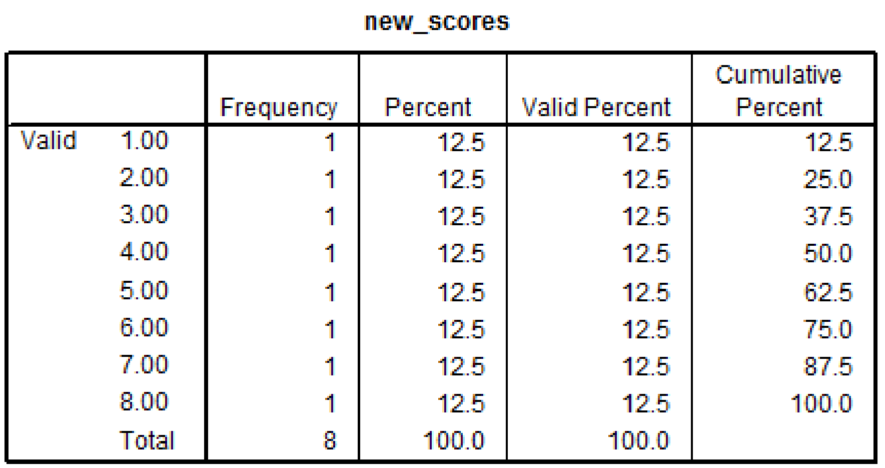
sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	f	7	63.6	63.6	63.6
	m	4	36.4	36.4	100.0
	Total	11	100.0	100.0	

Second Example: 6, 1, 3, 7, 5, 4, 2, 8

Find frequencies and relative frequencies for the above scores in SPSS

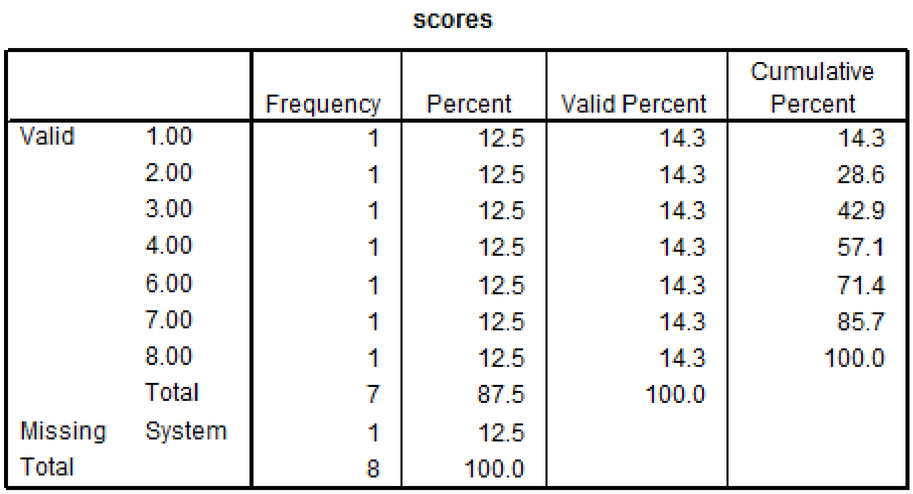
From SPSS



new_scores

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	12.5	12.5	12.5
	2.00	1	12.5	12.5	25.0
	3.00	1	12.5	12.5	37.5
	4.00	1	12.5	12.5	50.0
	5.00	1	12.5	12.5	62.5
	6.00	1	12.5	12.5	75.0
	7.00	1	12.5	12.5	87.5
	8.00	1	12.5	12.5	100.0
	Total	8	100.0	100.0	

Assume response 5 was omitted from the 8 observations – note difference in percent vs. valid percent. Valid percent is the column that typically should be used since it is based upon obtained data.



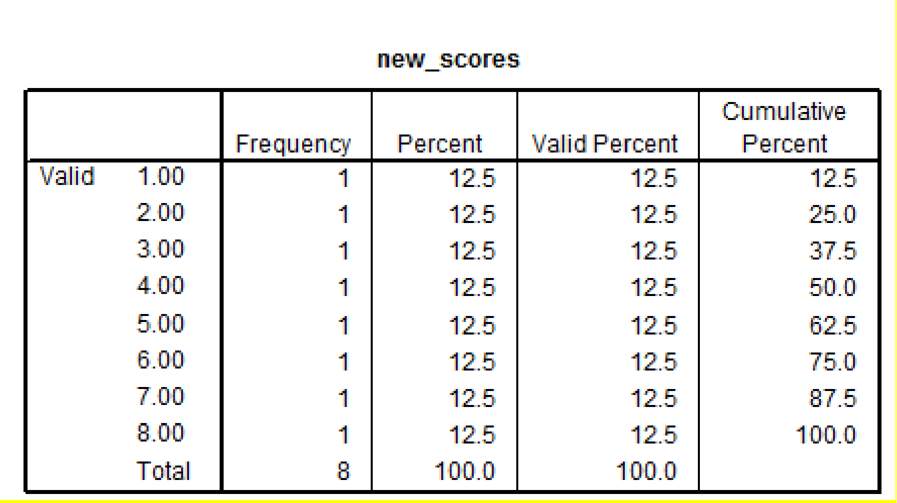
scores

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	12.5	14.3	14.3
	2.00	1	12.5	14.3	28.6
	3.00	1	12.5	14.3	42.9
	4.00	1	12.5	14.3	57.1
	6.00	1	12.5	14.3	71.4
	7.00	1	12.5	14.3	85.7
	8.00	1	12.5	14.3	100.0
	Total	7	87.5	100.0	
Missing	System	1	12.5		
Total		8	100.0		

Percentile Rank

Column cumulative percent = percentile rank for raw data.

What is a percentile rank?



new_scores

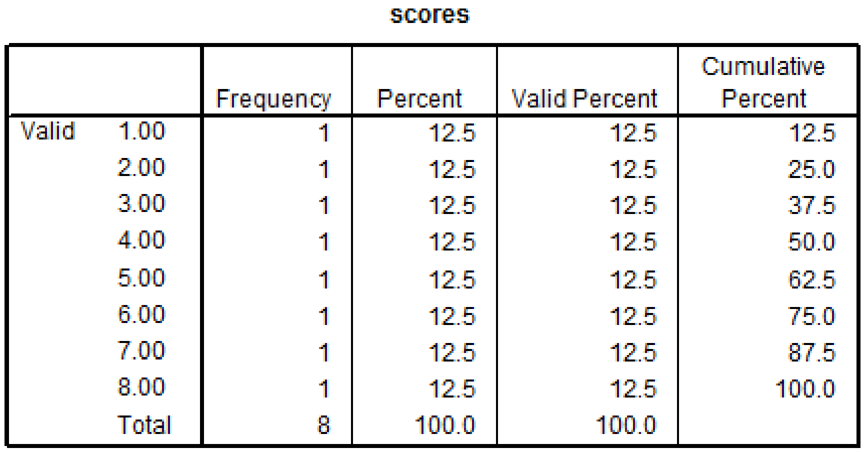
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	12.5	12.5	12.5
	2.00	1	12.5	12.5	25.0
	3.00	1	12.5	12.5	37.5
	4.00	1	12.5	12.5	50.0
	5.00	1	12.5	12.5	62.5
	6.00	1	12.5	12.5	75.0
	7.00	1	12.5	12.5	87.5
	8.00	1	12.5	12.5	100.0
	Total	8	100.0	100.0	

Most common definition and the one we will use:

PR = percentage (or proportion) of scores **at or** below a given score

Less common (and we won't use this one):

PR = proportion (or percentage) of scores below a given score.

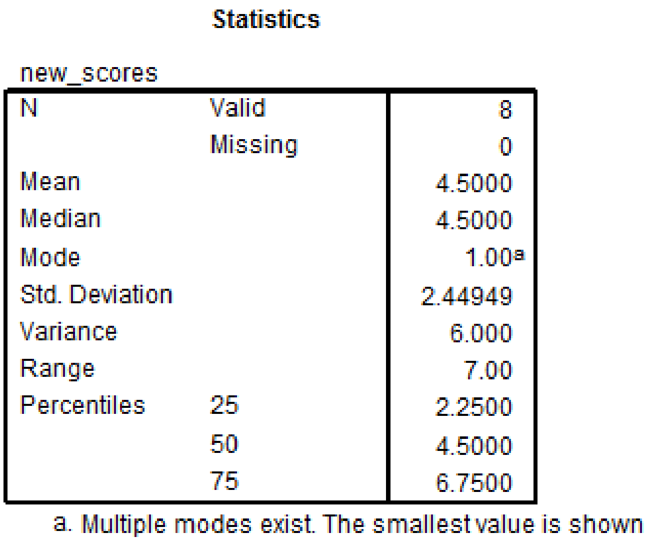


scores

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	12.5	12.5	12.5
	2.00	1	12.5	12.5	25.0
	3.00	1	12.5	12.5	37.5
	4.00	1	12.5	12.5	50.0
	5.00	1	12.5	12.5	62.5
	6.00	1	12.5	12.5	75.0
	7.00	1	12.5	12.5	87.5
	8.00	1	12.5	12.5	100.0
	Total	8	100.0	100.0	

Example – for score of 5, 62.5 is the PR which means 62.5% of sample scored 5 or less.

PR = 50 = median



Statistics

new_scores

N	Valid	8
	Missing	0
Mean		4.5000
Median		4.5000
Mode		1.00[a]
Std. Deviation		2.44949
Variance		6.000
Range		7.00
Percentiles	25	2.2500
	50	4.5000
	75	6.7500

a. Multiple modes exist. The smallest value is shown

Note that 2.25, 4.50, and 6.75 do not appear in our data. These represent calculated percentile scores for the ranks of 25, 50, and 75. They differ from the values provided by the Cumulative Percent column for the percentiles, and this discrepancy is common for small data files, and the even number of values creates the problem of the median of 4.00 vs. 4.50.

Quartiles

Quartiles are formed by the 25th, 50th, and 75th percentiles. Four sections with equal numbers of sampled units in each section.

Divide distribution into 4 sections, with 25% of scores in each section based, upon percentile ranks generally, but specifically using this these formulas:

1st quartile –median between lowest score and overall median of distribution
2nd quartile –median of distribution
3rd quartile –median between highest score and overall median of distribution

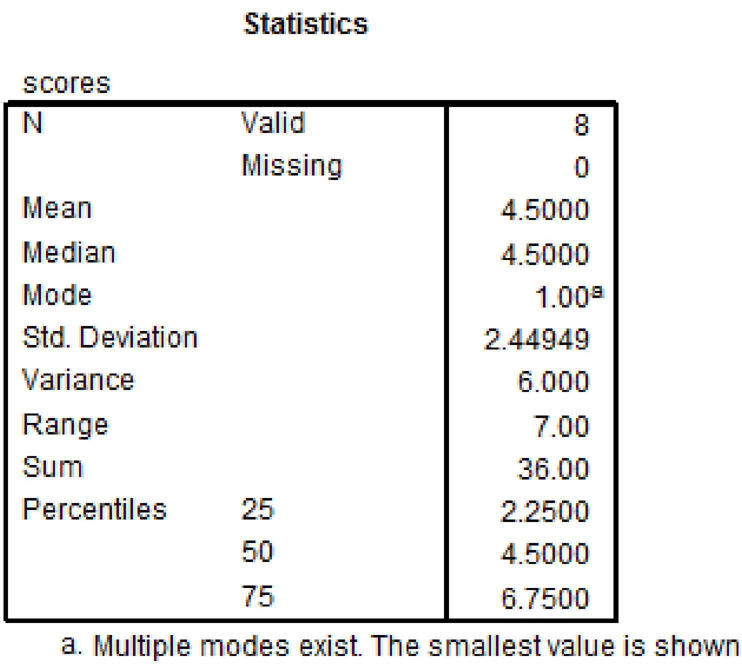
Also

1st quartile – 25th percentile
2nd quartile – 50th percentile (median)
3rd quartile – 75% percentile

i.e.,:

Scores 1 2 3 4 5 6 7 8
↑ ↑ ↑
Quartiles = 1st =2.5 2nd=4.5 3rd=6.5
Percentiles = 25 50 75

SPSS reports different values for quartiles: 2.25, 4.50, and 6.75



Statistics

scores

N	Valid	8
	Missing	0
Mean		4.5000
Median		4.5000
Mode		1.00[a]
Std. Deviation		2.44949
Variance		6.000
Range		7.00
Sum		36.00
Percentiles	25	2.2500
	50	4.5000
	75	6.7500

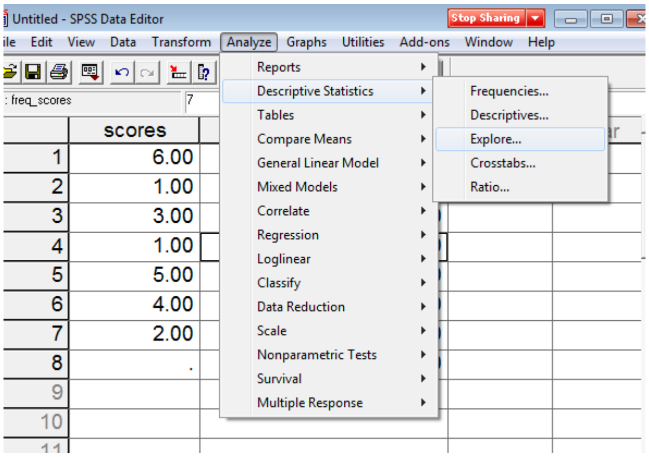
a. Multiple modes exist. The smallest value is shown

There are slight differences in quartile calculation, so if you do it by hand use the formula above and if you rely on software report whatever values they provide because all formulas for quartiles (and percentiles) provide close estimates.

Boxplot or Box and Whisker Plot

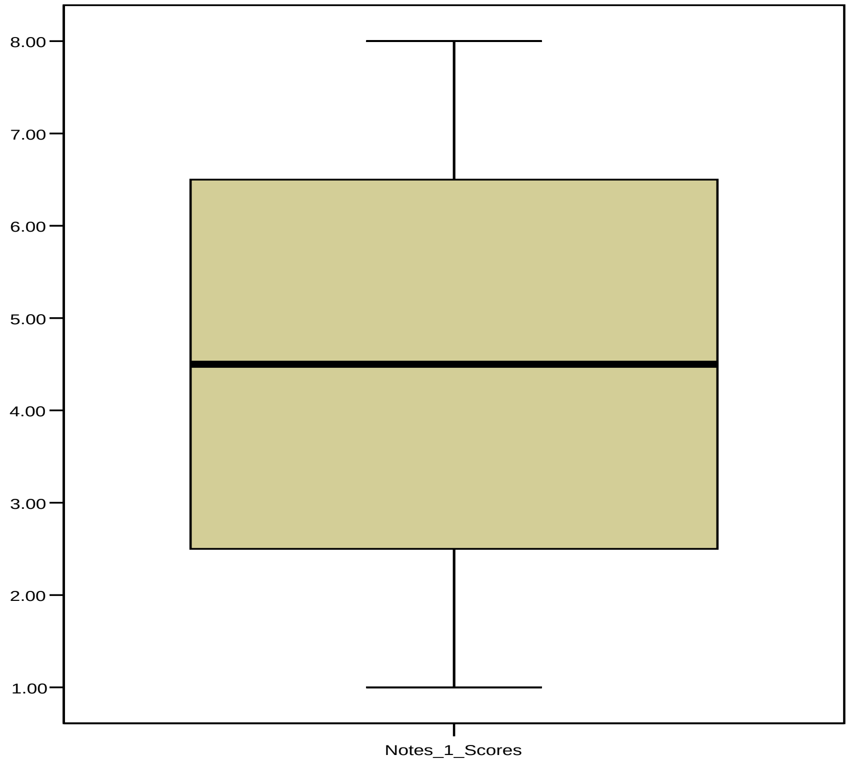
Graphical means of displaying central tendency and spread of scores.

SPSS Explore command produces boxplot of example data



: freq_scores 7

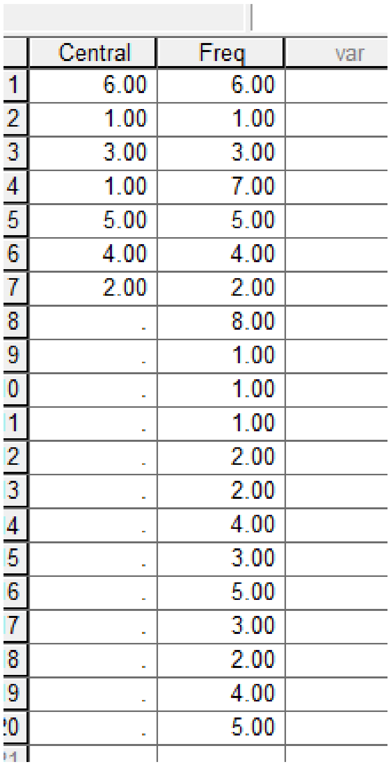
	scores
1	6.00
2	1.00
3	3.00
4	1.00
5	5.00
6	4.00
7	2.00
8	.
9	
10	



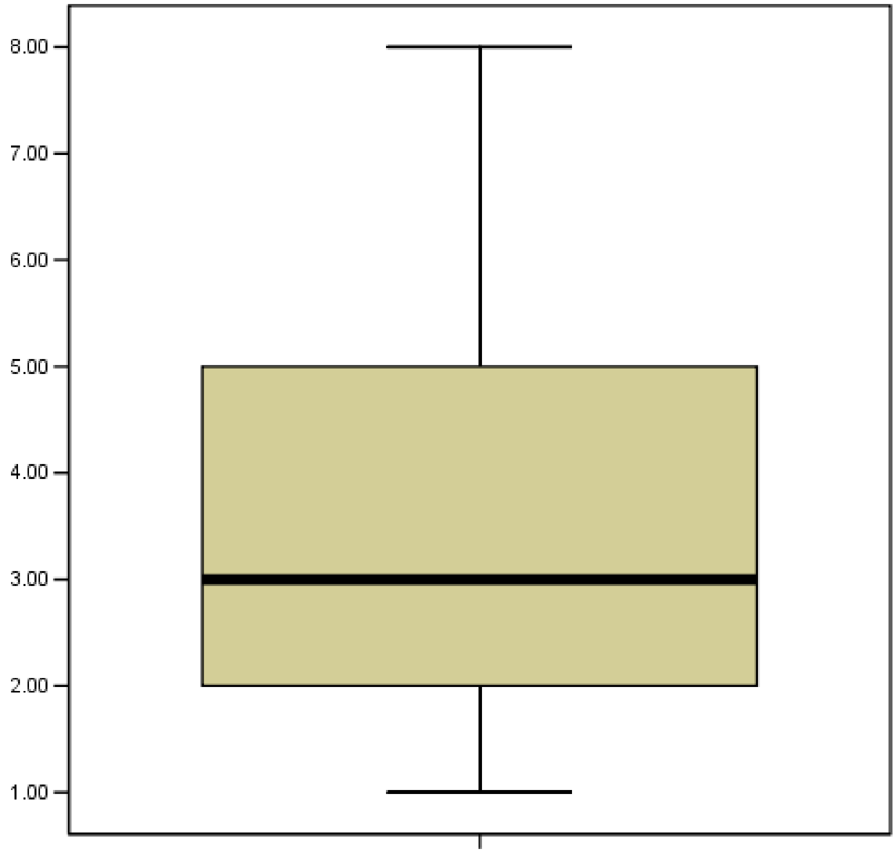
Note that the boxplot uses values of 2.5 and 6.5 for the 25th and 75th percentiles, which is inconsistent with the Frequencies command result.

Note that Box and Whisker plots need not show equal quartile sizes.

Example:



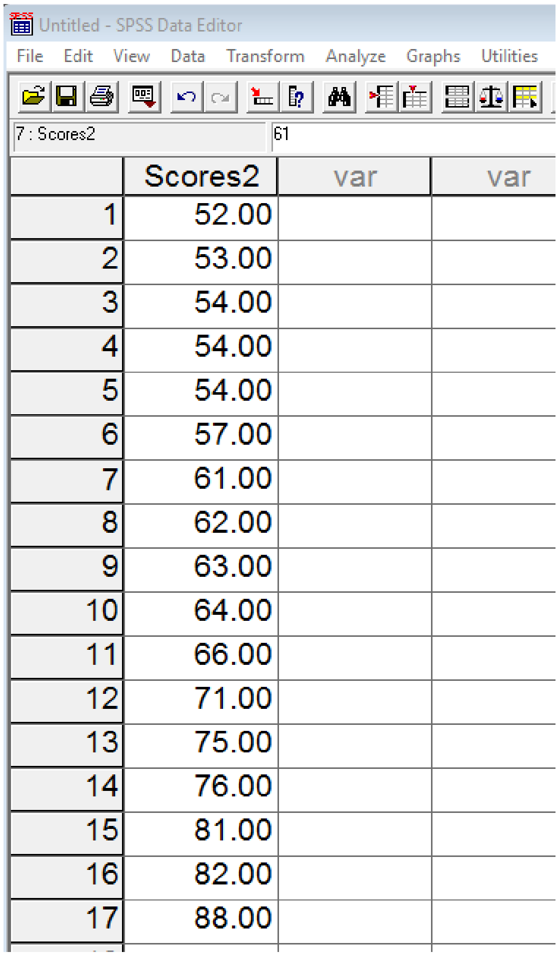
	Central	Freq	var
1	6.00	6.00	
2	1.00	1.00	
3	3.00	3.00	
4	1.00	7.00	
5	5.00	5.00	
6	4.00	4.00	
7	2.00	2.00	
8	.	8.00	
9	.	1.00	
0	.	1.00	
1	.	1.00	
2	.	2.00	
3	.	2.00	
4	.	4.00	
5	.	3.00	
6	.	5.00	
7	.	3.00	
8	.	2.00	
9	.	4.00	
20	.	5.00	



Stem-and-leaf Display

For this example the following data will be used:

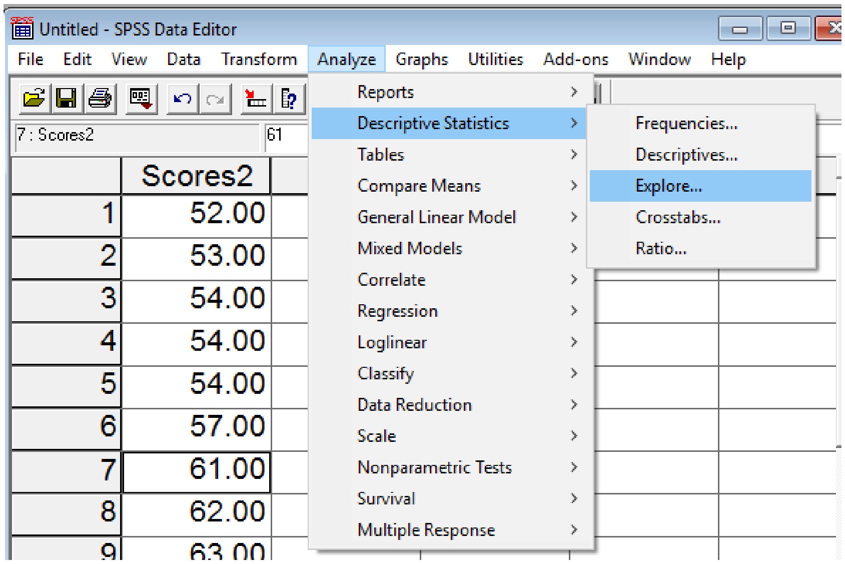
52, 53, 54, 54, 54, 57, 61, 62, 63, 64, 66, 71, 75, 76, 81, 82, 88



7 : Scores2 61

	Scores2	var	var
1	52.00		
2	53.00		
3	54.00		
4	54.00		
5	54.00		
6	57.00		
7	61.00		
8	62.00		
9	63.00		
10	64.00		
11	66.00		
12	71.00		
13	75.00		
14	76.00		
15	81.00		
16	82.00		
17	88.00		

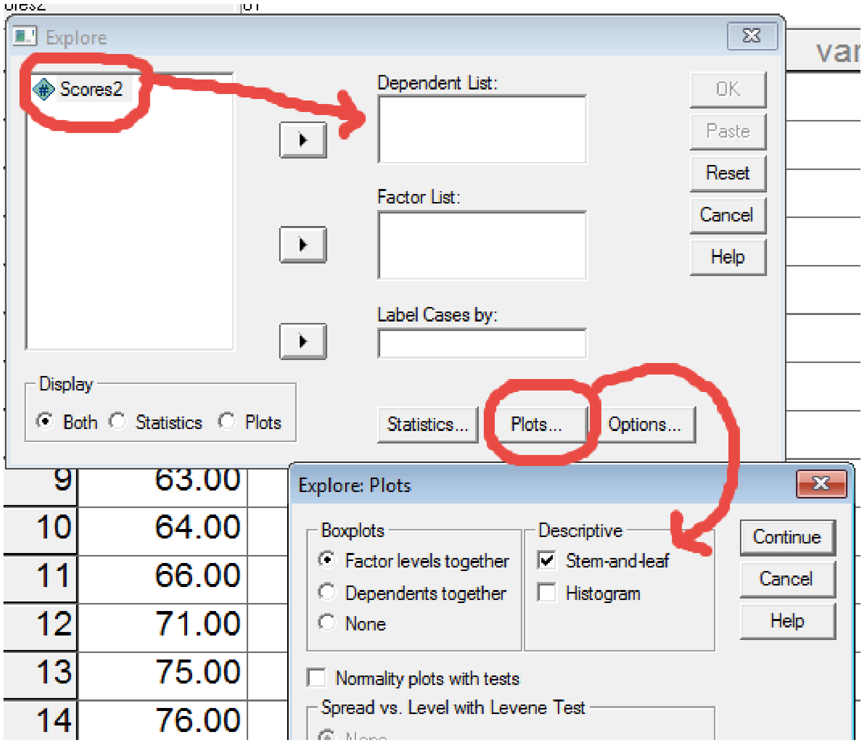
In SPSS, select the following commands:
Analyze -> Descriptive Statistics -> Explore



7 : Scores2 61

	Scores2
1	52.00
2	53.00
3	54.00
4	54.00
5	54.00
6	57.00
7	61.00
8	62.00
9	63.00

Then move the variable of interest into the Dependent List box, then select Plots and check Stem-and-leaf



Display
● Both ○ Statistics ○ Plots

Boxplots
● Factor levels together
○ Dependents together
○ None

Descriptive
☑ Stem-and-leaf
□ Histogram

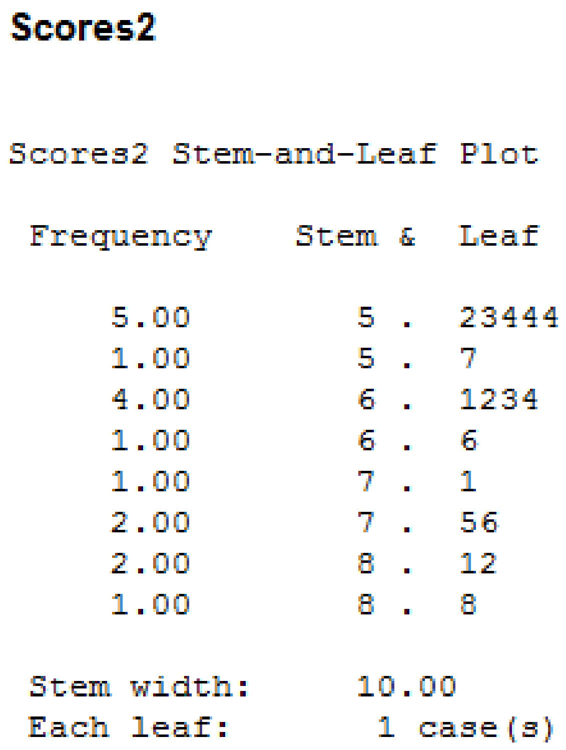
□ Normality plots with tests

Spread vs. Level with Levene Test
● None

9	63.00
10	64.00
11	66.00
12	71.00
13	75.00
14	76.00

Then click OK to run.

From SPSS



Scores2

```
Scores2 Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     5.00        5 .  23444
     1.00        5 .  7
     4.00        6 .  1234
     1.00        6 .  6
     1.00        7 .  1
     2.00        7 .  56
     2.00        8 .  12
     1.00        8 .  8

 Stem width:     10.00
 Each leaf:       1 case(s)
```

The stem are the scores on the left of the period, and the leaf on the right of the period. The first row looks like this:

5. 23444

This shows the scores 52, 53, 54, 54, and 54.

See instruction video under Notes 1 for other graphical displays.