Homework #2: Descriptive Statistics: Central Tendency & Variability

Homework Hints: Please round all of your final answers to 2 decimals (e.g., 1.75, not 1.749). Round your final answers; rounding middle steps may result in inaccurate answers.

Homework 2: Questions 1, 2 (describe your answer to yourself verbally so that you know you can answer it; add your response to your notes if you desire), 3, 4, 5, 6, 7, 8

1. Here are the noon temperatures (in degrees Celsius) any particular Canadian city on Boxing Day (usually December 26) for the 10 years from 1998 through 2007: -5, -4, -1, -1, 0, -8, -5, -9, -13, and -24. Describe the typical temperature and the amount of variation to a person who has never had a course in statistics. Provide *three* ways of describing the representative temperature and *two* ways of describing its variation, explaining the differences and how you determined each.

2. Describe and explain the location of the mean, mode, and median for a normal curve.

3. Understanding measures of central tendency
   a. Describe and explain the difference between the mean, median, and mode.
   b. Generate your own example (neither from your book nor from lecture) in which the median would be the preferred measure of central tendency.

4. Understanding measures measure of dispersion
   a. Describe the concepts of the variance and standard deviation.
   b. Explain why the standard deviation is more often used as a descriptive statistic than the variance; why might the standard deviation be more useful?

5. A psychologist interested in political behavior measured the square footage of the desks in the offices of four U.S. Governors and four chief executive officers (CEOs) of major U.S. corporations. The measurements for the governors were 44, 36, 52, and 40 square feet. The measures for the CEOs were 32, 60, 48, and 36 square feet.
   a. Calculate the means and standard deviations for both sets of scores.
   b. Verbalize to yourself, in *conceptual* terms rather than computational terms, what you have just done.

6. Describe and explain the location of the mean, mode, and median of a distribution of scores that is skewed strongly negatively.

7. You calculate the variance of a distribution of scores to be -4.26. Why must this value be incorrect?

8. The sum of individual deviations from the sample mean is equal to what value?

SPSS Assignment: Complete the procedures described on the next pages (Part 1 and Part 2) and then answer the questions. Although not required, it is best to replicate the SPSS printout to be sure you are conducting you analysis properly before answering other questions with SPSS.

Part 1.

1. Open SPSS and save the file you create under the folder "Stats109" on your U:\drive. Create a folder on your McKenna Network drive, the U:\drive, and name it "**Stats109**". You can save all of your stats work in this directory. We will access the data from this directory over the course of the semester.

2. Create a data file in SPSS and enter the following data. Name your data file "HW2_Q2" and save it in your Stats109 folder so that you can locate it later if necessary.
Data: 2, 2, 0, 5, 1, 4, 1, 3, 0, 0, 1, 4, 4, 0, 1, 4, 3, 4, 2, 1, 0

3. For full credit, make up a variable name to represent the set of numbers you entered. Also enter a *variable label* for the variable.

4. Run the procedure **Frequencies**. This is found under the **Analyze** menu and then under **Descriptive Statistics**). Do not check any boxes in this frequency window yet. Press **OK** to run the Frequencies procedure. What type of frequency distribution does this procedure produce (ungrouped table, grouped table, bar chart, histogram, pie chart, etc.)?

   Use Analyze → Descriptive Statistics → Frequencies to check over your data entry.

   a. Click on **Statistics** (within the **Frequencies** option) to provide the measures of central tendency, variability, and shape (mean, median, mode, standard deviation, variance, skewness, and kurtosis).

   b. Click on **Charts** (within the **Frequencies** option) and select **histogram** so that you can identify the shape of the distribution. Is the shape unimodal, bimodal, or multimodal? Is the graph skewed? If so, describe what the skew is and how you see skewness within the graph. Is the graph kurtotic? If so, describe what the kurtosis is and how you can determine kurtosis by within the graph.

5. Print your output for these three procedures (frequencies, statistics, and charts) listed above.

Part 2.

1. Open the SPSS data file named "*Coronary artery data.sav*" from your "Stats109" folder. Whenever you need to access data files while using SPSS, said files will be located in this "Stats109" folder on your U:\ Drive for easy access in the future.

2. Looking at the *variable view* window in SPSS, what are the variable names (and variable labels, if applicable) and values (and value labels, if applicable) for both variables?

3. What is the scale of measurement for each of these variables? (nominal, ordinal, equal-interval, etc.). Hint, SPSS calls equal-interval variables "scale" variables.

4. Generate a histogram for ***time*** just as you did for the data in Part 1 of this assignment. How would you characterize the shape of the distribution for ***time***?

5. Print this histogram and attach it to your homework.

Homework #2:
Answers

1. **Noon Temperature**

    The average temperature, in the sense of adding up that 10 readings and dividing by 10, was -7° C. This value represents that *mean*. However, if you organize, or rank, the temperatures from highest to lowest, the middle two scores are both -5° C. The middle number represents the *median,* or the score that splits the distribution into two equal parts. The specific temperature that occurred most often represents the mode; in this case, there are two modes -1 and -5 degrees Celsius. In other words, the distribution is bimodal.

    As for the variation of scores (the amount of variability among scores in a distribution), the variance represents the average of each temperatures squared deviation from the mean temperature, which is 46.8° C. However 46.8°C, as a squared value, may not best described the variability of scores because the variance is a squared value. Therefore, by taking the square root of the variance, the standard deviation represents the average deviation from the mean of all scores. In this case, the standard deviation is 6.84°C. This means that on average daily temperatures deviate 6.84° (either higher or lower) from the average of -7° C.

2. **Understanding measures of central tendency**
    - **a.** The mean, median, and mode for a normal distribution are all located at the same midpoint of the curve. This point is also the highest point in the distribution. In other words, this midpoint represents the most peaked point of the distribution. The mean, median, and mode then also describe the center of the distribution for a normal distribution equally well, however, the mean provides the most statistical information for describing the distribution. The *mode* is the highest point in the distribution because it is the most frequently occurring score and is located in the center of the distribution. This midpoint is also the *median* value which splits the distribution into two equal parts, above and below which 50% of the distribution falls. The *mean* also falls at the same point because the normal distribution is symmetrical around the midpoint and the mean is the mathematical fulcrum of the normal distribution .

3. **Understanding measures of dispersion**
    - **a.** The mean is the average of a set of numbers; which represents the sum of the scores divided by the total number of scores. The mean is the most commonly used measure of central tendency because it accounts for every score in the distribution and does not very much from sample to sample. Specifically, the mean is less sensitive to sampling variation than are other measures of central tendency. The median represents the middle score of a distribution; the median is easy to identify when scores are organized from lowest to highest. The median is also sometimes used in lieu of the mean when outlying scores change the shape of the distribution to become skewed either positively or negatively. In such cases, the mean of the sample will not be the best measure of central tendency to describe the center of the distribution. Finally, the mode is the most commonly occurring value in a distribution; the mode is often used to analyze variables with a nominal scale of measurement.
    - **b.** The median would be the preferred measure of central tendency to represent a data set of equal-interval variables with an outlying score. A data set best represented by the median would be the following set of scores for the shoe sizes of 9-year-old boys: 5, 6, 6, 7, 7, 7, 7, 7, 8, and 15.

Variance and Standard Deviation

a. The *variance* represents the average squared deviation of scores from the mean. The variance is calculated by squaring each score's individual deviation from the mean, adding these squared values, and dividing by the total number of scores. $SD^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{N}$ *or* $\frac{SS}{N}$. By extension, the standard deviation represents the average deviation of scores from the mean and is calculated by taking the square root of the variance.

b. The *standard deviation* is more often used as a descriptive statistic than the variance because the standard deviation offers a more direct representation of the deviations from the mean because the deviations are not squared. In other words, because when people talk about data they talk about actual measured units (rather than squared units), the standard deviation is common parlance. The variance is more commonly used to calculate the standard deviation.

4. **Governor's Desks vs. CEOs' Desks**

a. The mean for governors' desks $\overline{X} = \left(44\ +\ 36\ +\ 52\ +\ 40\right)/\,4\ = 43$ square feet

The mean for CEOs' desks: $\overline{X} = \left(32\ +\ 60\ + 48\ +\ 36\right)/\,4\ =\ 44$ square feet

The standard deviation for governors' desks:

$$\left(44 - 43 = 1\right)^2 + \left(-7\right)^2 + \left(9\right)^2 + \left(-3\right)^2 = 140$$
$$140\,/\,4 = 35$$
$$\sqrt{35} = 5.92$$

The standard deviation for CEOs' desks:

$$\left(32 - 44 = -12\right)^2 + \left(16\right)^2 + \left(4\right)^2 + \left(-8\right)^2 = 480$$
$$480\,/\,4 = 120$$
$$\sqrt{120} = 10.95$$

b. Although the means for the two samples are similar (e.g., 43 vs. 44 square feet), the standard deviation for the CEOs' desks is nearly twice as much as for the governors' desks. This means that there is a broader range of desk sizes among CEOs than among governors. In other words, governors' desk sizes are more similar to each other than CEOs' desks are similar to each other. This could potentially indicate that there is a greater range of incomes among CEOs than among governors, due to larger slush funds, etc.

5. **Location of Mean, Median, and Mode on Skewed Distribution**
For a distribution of scores that is skewed strongly negatively, the mean would be much further to the left tail of the distribution than would either the median or the mode. This is because the calculation of the mean makes it sensitive to a single outlying score without having as large an impact on either the median or the mode. The mode would be located toward the right of the distribution and surrounded by most of the other scores because the mode is not affected by outliers. The median would be located between the mean and the mode; it would be located left of the mode because the median has to take account of the outlier and is affected by the outlier, but not to the extent the mean is affected by the outlier.

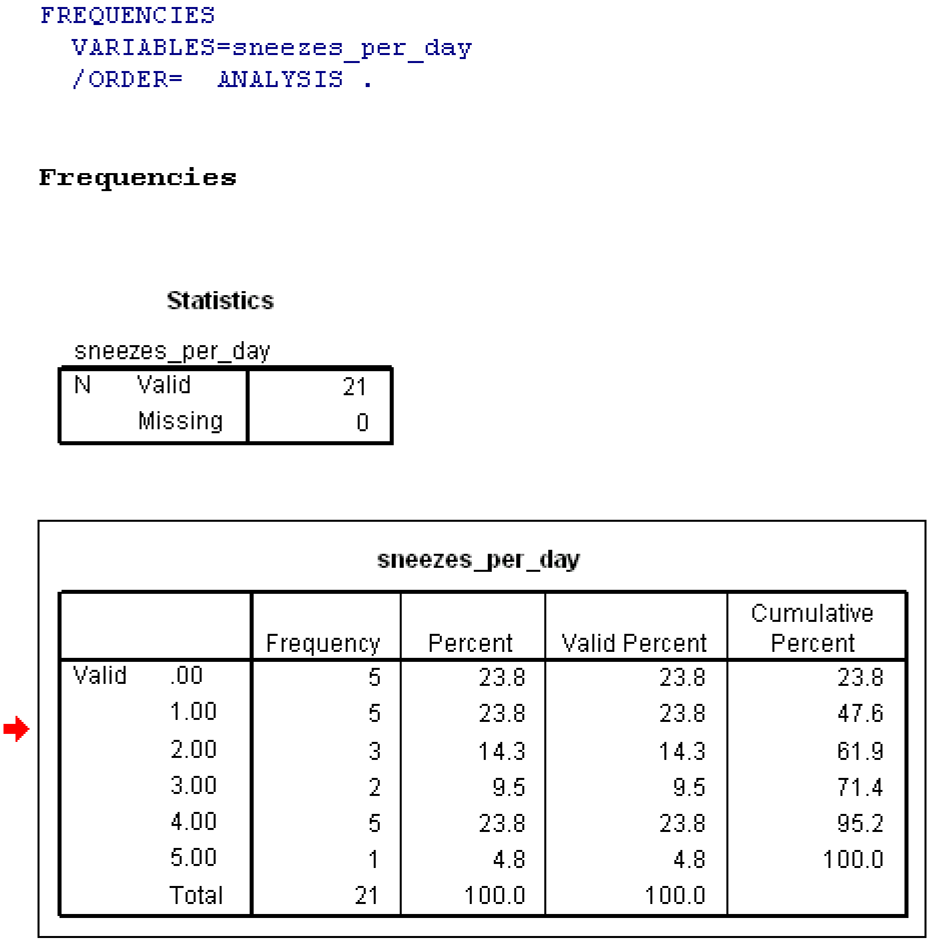
6. **Calculate a Negative Variance?**

Calculating a negative variance measure must have occurred in error because the variance represents the average SQUARED deviation from the mean. All squared values have to be positive. $SD^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{N} \; or \; \frac{SS}{N}$

7. The sum of all individual deviations from the mean is $\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0$. The mean serves as a fulcrum of the distribution that splits the distribution into two equal parts based on the variability of the scores in the sample.

SPSS Part 1

For the following scores (2, 2, 0, 5, 1, 4, 1, 3, 0, 0, 1, 4, 4, 0, 1, 4, 3, 4, 2, 1, 0



```
FREQUENCIES
  VARIABLES=sneezes_per_day
  /ORDER=  ANALYSIS .
```

Frequencies

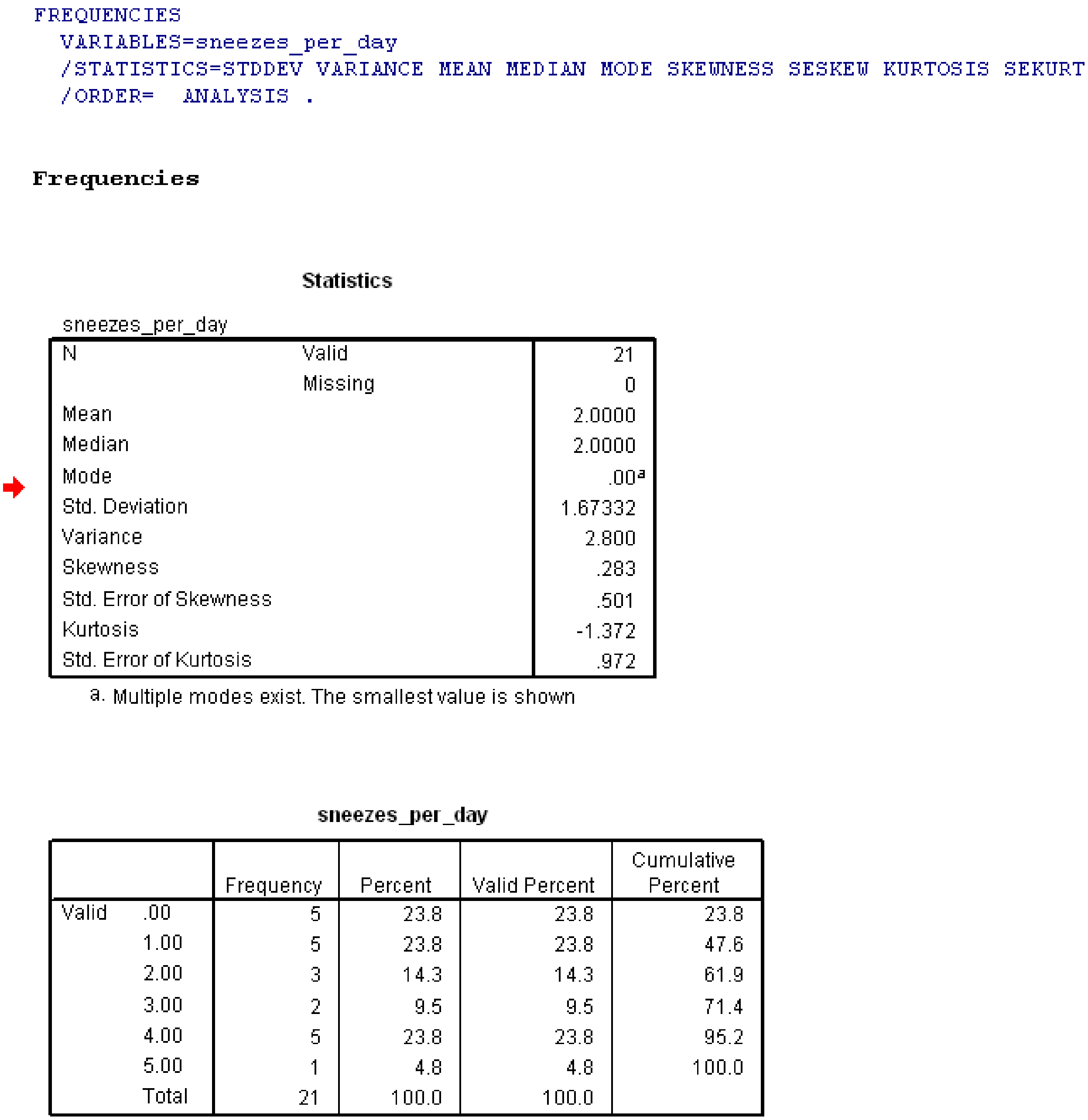
Statistics

sneezes_per_day

N	Valid	21
	Missing	0

sneezes_per_day

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	5	23.8	23.8	23.8
	1.00	5	23.8	23.8	47.6
	2.00	3	14.3	14.3	61.9
	3.00	2	9.5	9.5	71.4
	4.00	5	23.8	23.8	95.2
	5.00	1	4.8	4.8	100.0
	Total	21	100.0	100.0	



```
FREQUENCIES
  VARIABLES=sneezes_per_day
  /STATISTICS=STDDEV VARIANCE MEAN MEDIAN MODE SKEWNESS SESKEW KURTOSIS SEKURT
  /ORDER=  ANALYSIS .
```

Frequencies

Statistics

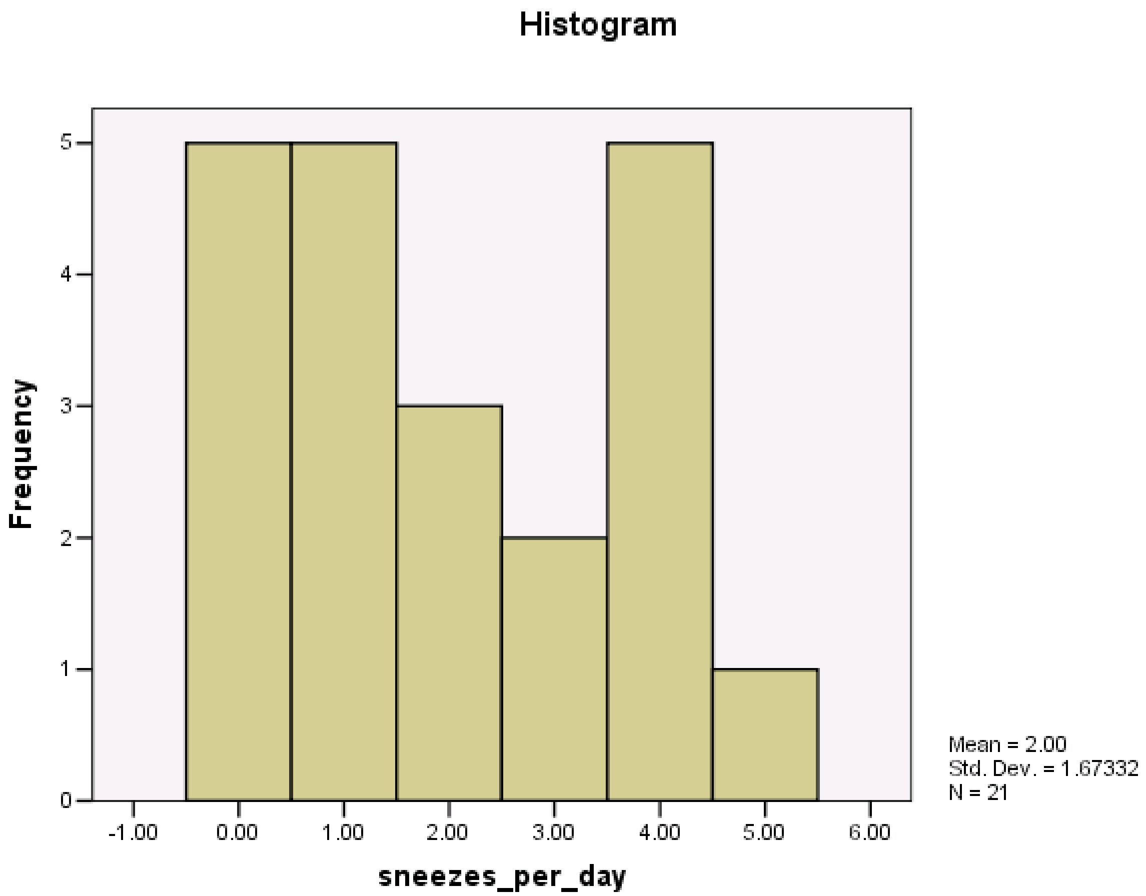
sneezes_per_day

N	Valid	21
	Missing	0
Mean		2.0000
Median		2.0000
Mode		.00[a]
Std. Deviation		1.67332
Variance		2.800
Skewness		.283
Std. Error of Skewness		.501
Kurtosis		-1.372
Std. Error of Kurtosis		.972

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

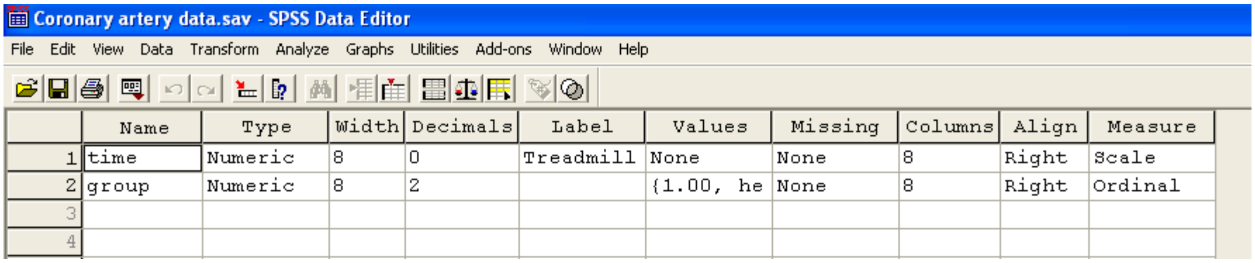
sneezes_per_day

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	5	23.8	23.8	23.8
	1.00	5	23.8	23.8	47.6
	2.00	3	14.3	14.3	61.9
	3.00	2	9.5	9.5	71.4
	4.00	5	23.8	23.8	95.2
	5.00	1	4.8	4.8	100.0
	Total	21	100.0	100.0	


Histogram
Frequency
0
1
2
3
4
5
-1.00
0.00
1.00
2.00
3.00
4.00
5.00
6.00
sneezes_per_day
Mean = 2.00
Std. Dev. = 1.67332
N = 21

SPSS Part 2

Variable View

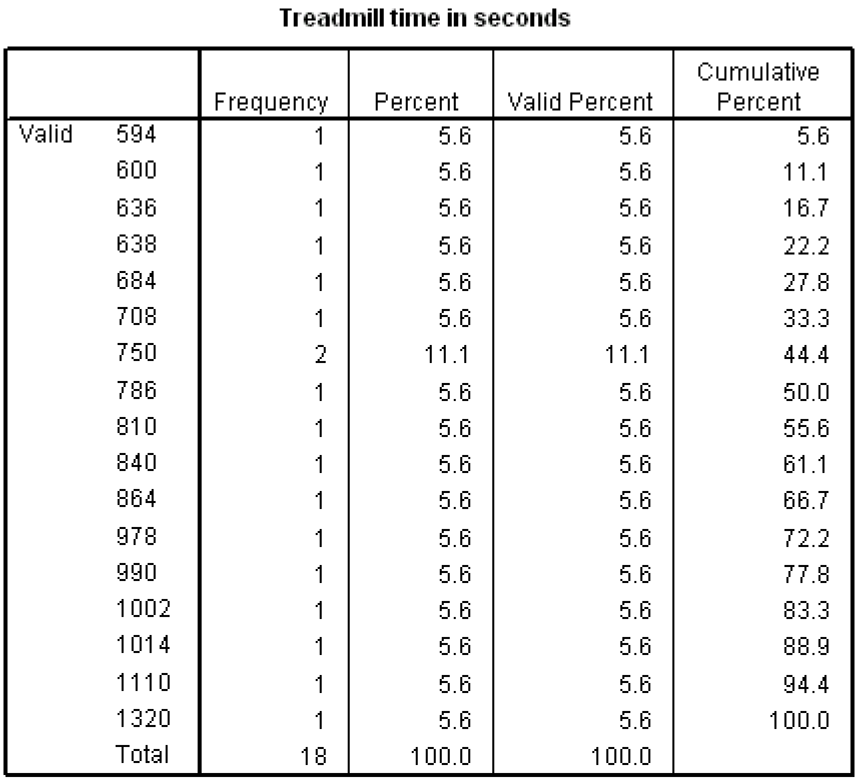


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	time	Numeric	8	0	Treadmill	None	None	8	Right	Scale
2	group	Numeric	8	2		{1.00, he	None	8	Right	Ordinal
3										
4										

Histogram

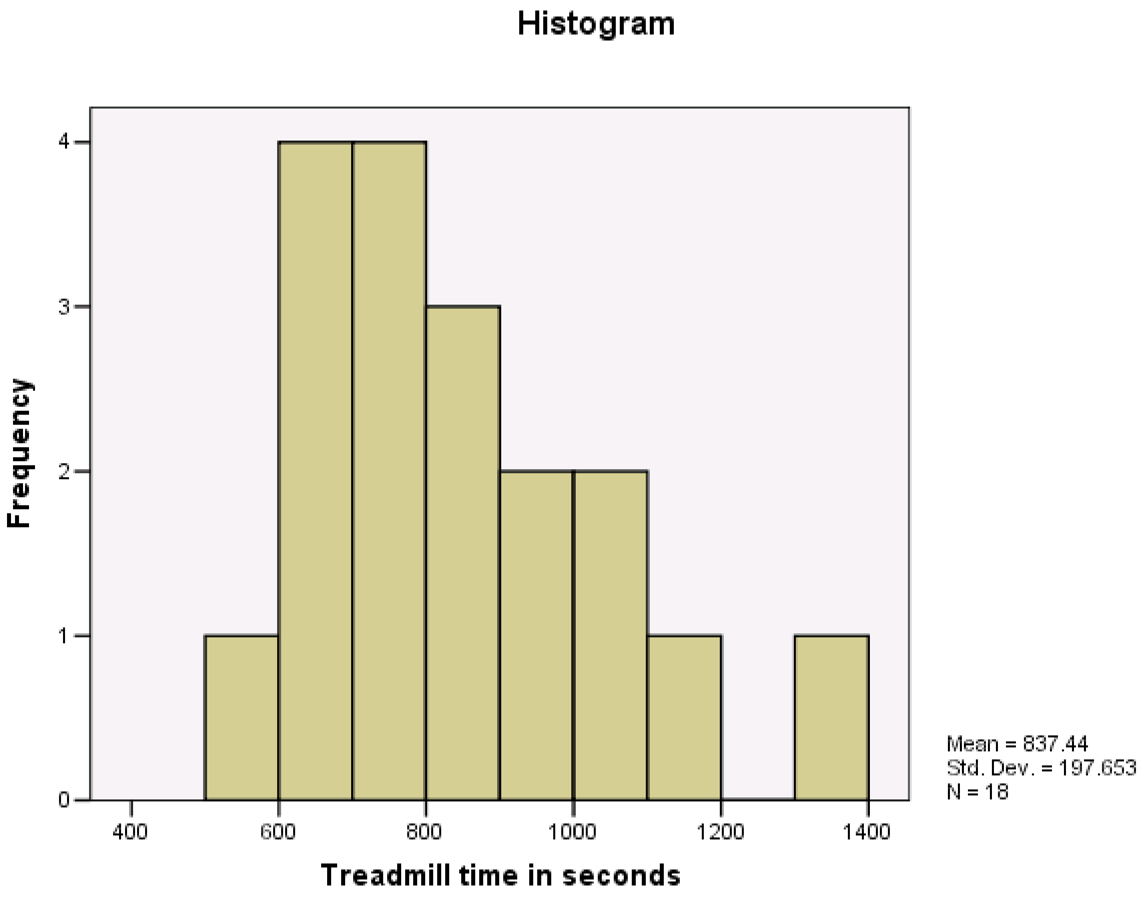
```
FREQUENCIES
  VARIABLES=time
  /HISTOGRAM
  /ORDER=  ANALYSIS .
```

Frequencies



Treadmill time in seconds

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	594	1	5.6	5.6	5.6
	600	1	5.6	5.6	11.1
	636	1	5.6	5.6	16.7
	638	1	5.6	5.6	22.2
	684	1	5.6	5.6	27.8
	708	1	5.6	5.6	33.3
	750	2	11.1	11.1	44.4
	786	1	5.6	5.6	50.0
	810	1	5.6	5.6	55.6
	840	1	5.6	5.6	61.1
	864	1	5.6	5.6	66.7
	978	1	5.6	5.6	72.2
	990	1	5.6	5.6	77.8
	1002	1	5.6	5.6	83.3
	1014	1	5.6	5.6	88.9
	1110	1	5.6	5.6	94.4
	1320	1	5.6	5.6	100.0
	Total	18	100.0	100.0	


Histogram
Frequency
0
1
2
3
4
400
600
800
1000
1200
1400
Treadmill time in seconds
Mean = 837.44
Std. Dev. = 197.653
N = 18