Are statistics relevant to real life?

Frequency distributions
Central tendency (mode, median, mean)
Measures of variability

Nikos Comoutos PhD

Why Do we use statistics

- Statistics is basically about understanding how to use data

Descriptive statistics:
Methods used to describe data and their characteristics

Inferential statistics:
What we know to make estimates or predictions (inferences) about what we don't know.

STATISTICS IS ALL ABOUT WEIGHING UP THE CHANCES OF SOMETHING HAPPENING OR BEING TRUE

Graphs

- Once you have collected some data – plot a graph of how many times each score occurs
- Producing graphs enables you to learn a lot about a dataset at a glance.
- When the data are quantitative to see what the general trends in the data are
- Fitting statistical models to the data

Plotting data…

- There are graphs for discrete data, continuous data, time series data, etc.
- The vertical axis of a graph is known as the y – axis of the graph
- The horizontal axis of a graph is known as the x-axis of the graph
- When you create a graph avoid 3-D effects, pictures

Some of these methods

- Frequency distributions
- Histograms
- Bar graphs
- Frequency polygons
- Stem-and-leaf
- Pie graphs

Frequency distributions

- How many subjects were similar in the sense that, measured on the dependent variable, they ended up in the same category or had the same score.
- Different shapes and sizes
- Ideal world data would be distributed symmetrically around the center of all scores

Simple or grouped frequency distribution

TABLE 2.1

SCORES OBTAINED ON A 100-POINT EXAM

82	90	84	89	92	90	85
89	82	76	87	87	83	82
78	74	91	93	84	79	87
87	79	87	81	80	74	95
90	92	83	70	91	93	88
81	85	86	87	87	88	89

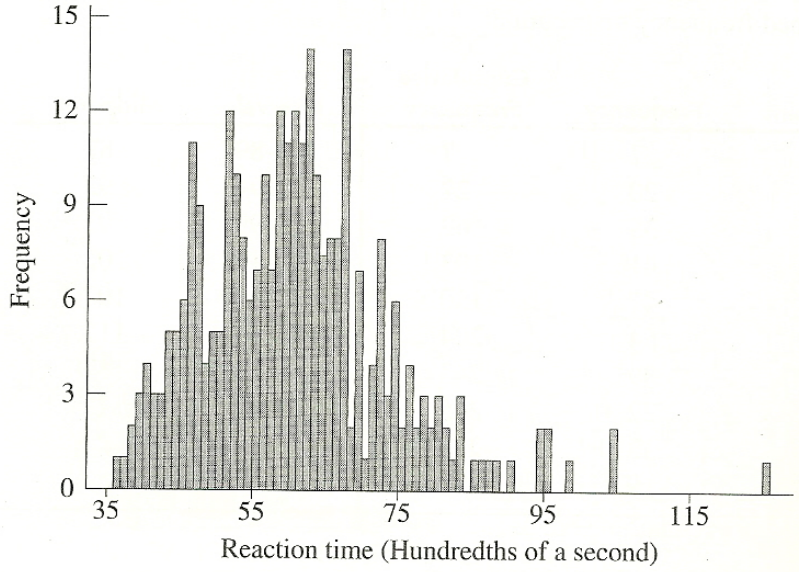


FIGURE 2.1 *Plot of reaction times against frequency*

TABLE 2.2

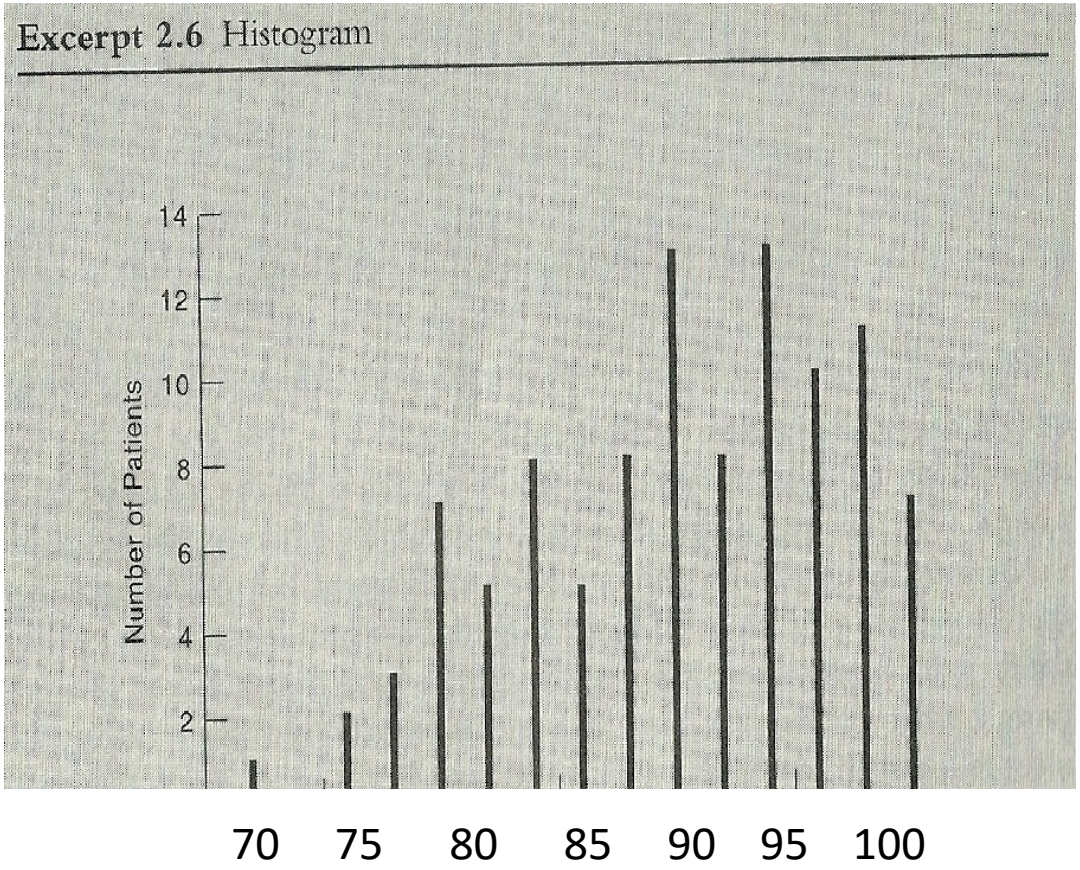
SIMPLE AND GROUPED FREQUENCY DISTRIBUTIONS

Score	*f*
95	1
94	0
93	2
92	2
91	2
90	3
89	3
88	4
87	7
86	1
85	3
84	4
83	2
82	3
81	2
80	1
79	2
78	1
77	0
76	2
75	0
74	2
73	0
72	0
71	0
70	1
$N =$	48

Interval	*f*
93–95	3
90–92	7
87–89	14
84–86	8
81–83	7
78–80	4
75–77	2
72–74	2
69–71	1
$N =$	48

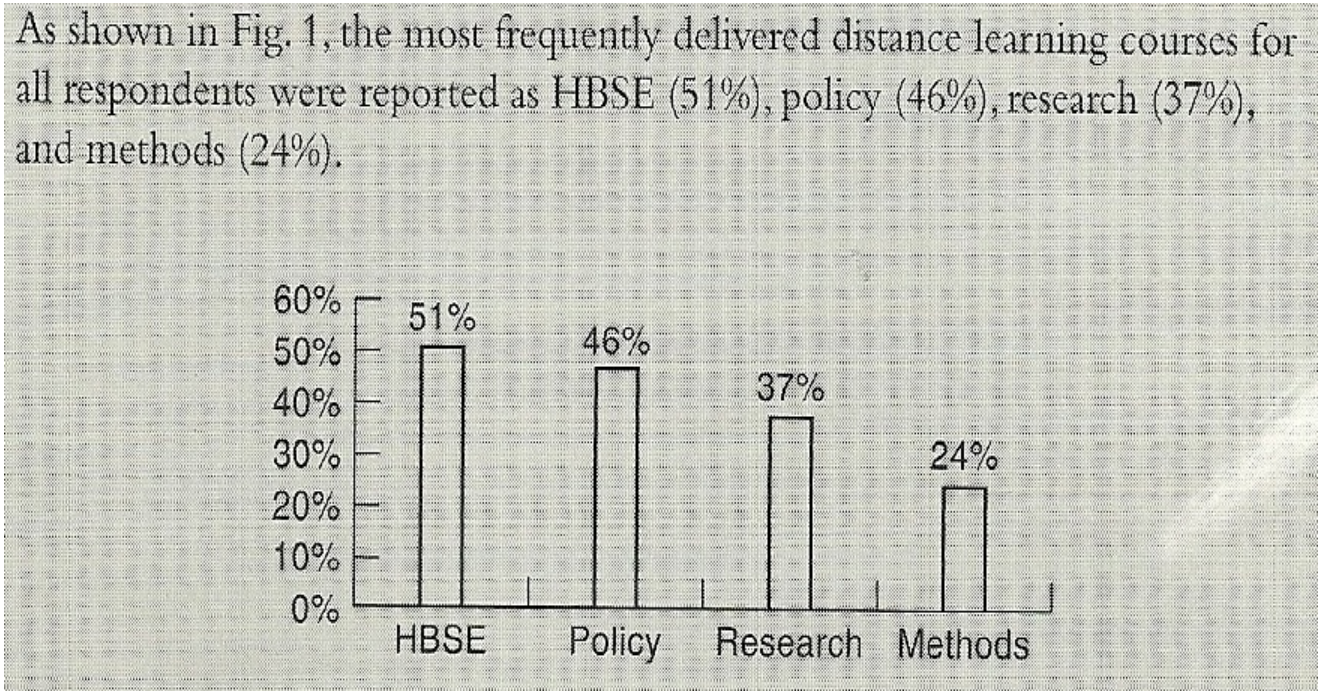
Histogram

- Vertical columns indicate how many times any given score appears in the data set. Technique the baseline (horizontal axis – x) corresponds with the observed scores on the dependent variable, while the vertical axis (y) is labeled with frequencies. The columns or lines are positioned above each baseline value to indicate how often each of these scores was observed.



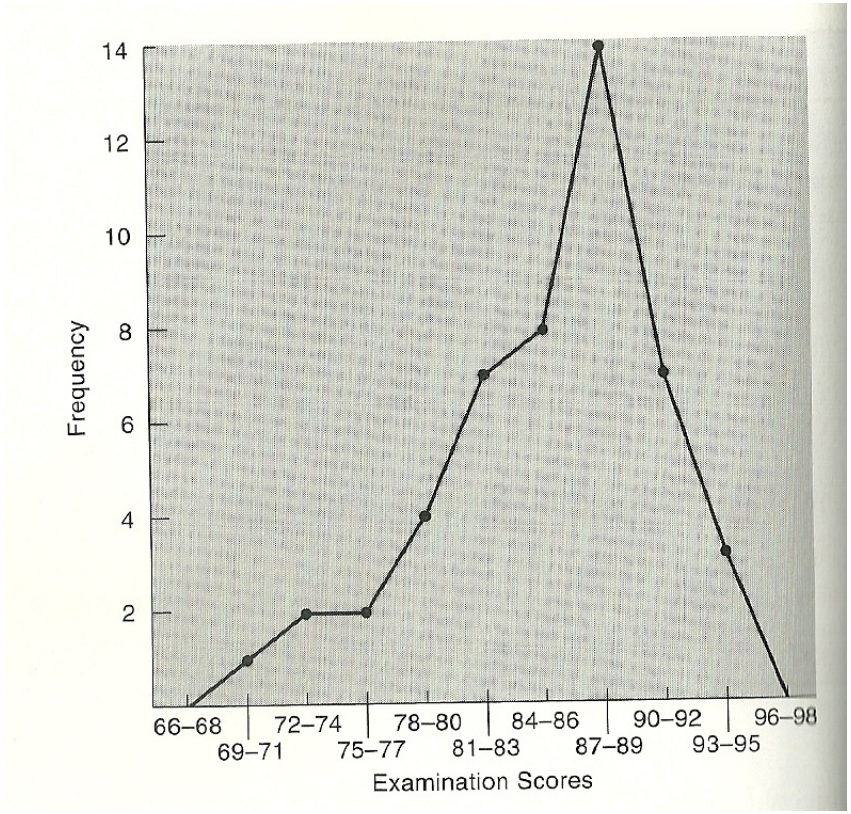
Bar graph

- Identical to histogram
- In a histogram the horizontal axis is labeled with numerical values that represent a quantitative variable
- In bar graph the horizontal axis represents different categories of a qualitative variable



Frequency Polygon

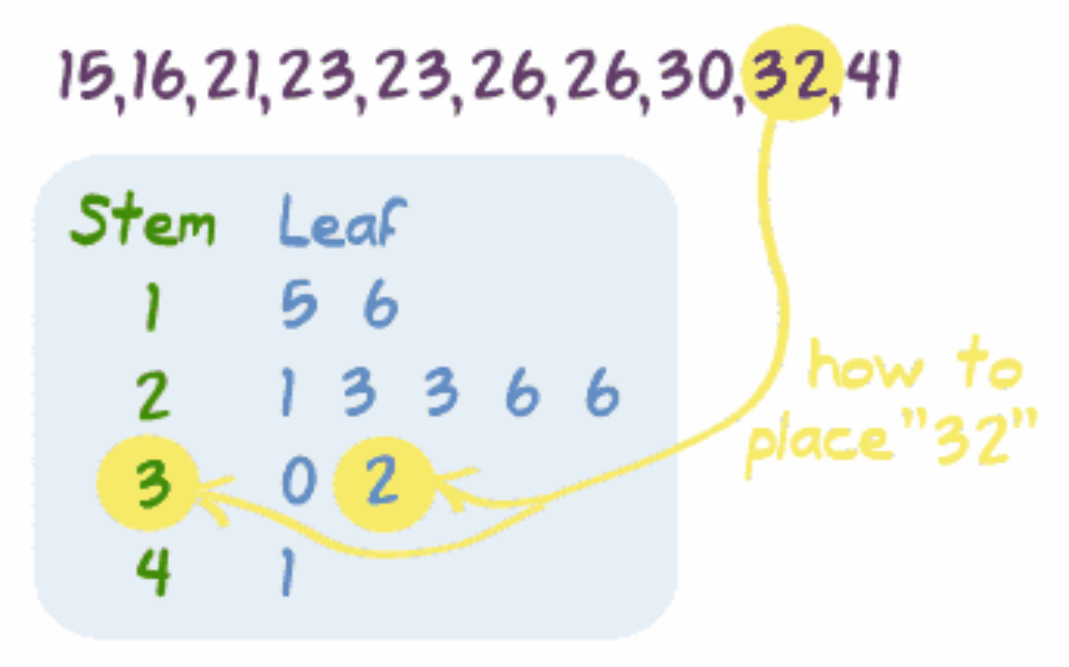
- Technical name of the line graph
- Each dot represents individual scores or score intervals, then these adjacent dots are connected with straight lines to form the final graph



Stem-and-leaf

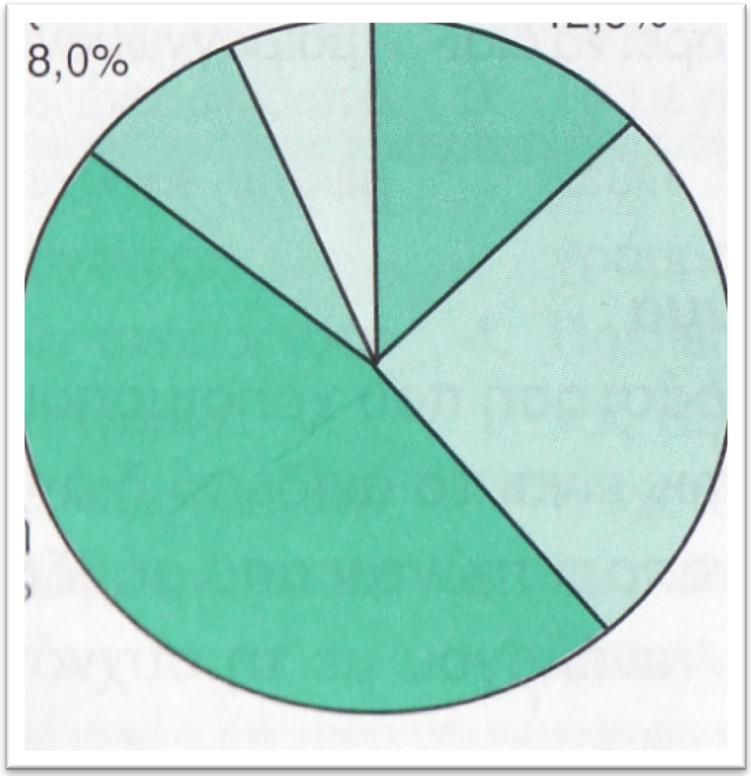
- Is like a grouped frequency distribution that contains no loss of information. First set up score intervals on the left side of a vertical line. These intervals collectively called the “stem”, and are presented in a coded fashion by showing all but the last digit of the scores falling into each interval. Then to the right of the vertical line, the final digit is given for each observed score that fell into the interval being focused upon

stem	leaf
9	0 4
8	3 4 5 7 9
7	0 2 2 9
6	5 8 9
5	3



Pie graphs

- How a full group is made up of subgroups- and also of showing the relative size of the subgroups



Distributional shape

- All the above methods allow us to tell whether our data are symmetrical.
- Most of the times pictures of the data sets don't appear in journal articles because they are costly to prepare and take up lots of space
- So researchers tell their readers what their data sets look like / describe the distributional shape of the their data

Normal distribution

- Most scores will clustered near the middle of the continuum of observed scores and there will be a symmetrical decrease in frequency in both directions away from the middle area of scores
- If we drew a vertical line through the center of the distribution then it should look the same on both sides- NORMAL DISTRIBUTION – BELL SHAPED CURVE
- Many naturally occurring things have this shape of distribution (e.g., most men are about 175 cm tall, some are a bit taller or shorter but most cluster around this value)
- In a normal distribution the values of skew and kurtosis is 0.

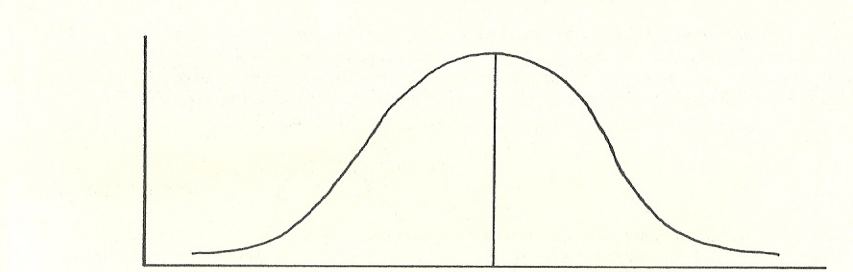


FIGURE 4.6 DATASET FOLLOWING A NORMAL DISTRIBUTION

Deviation from normality

- Two main ways in which a distribution can deviate from normality

1) Lack of symmetry (skew)
2) Lack of pointyness (kyrtosis)

Skewed distributions

- In **skewed distributions** most of the scores end up being high (or low) with a small percentage of scores strung out in one direction away from the majority
- **Positively skewed**: the frequent scores are clustered at the lower end and the tail points towards the higher or more positive scores
- **Negatively skewed**: the frequent scores are clustered at the higher end the tail points towards the lower or more negative scores

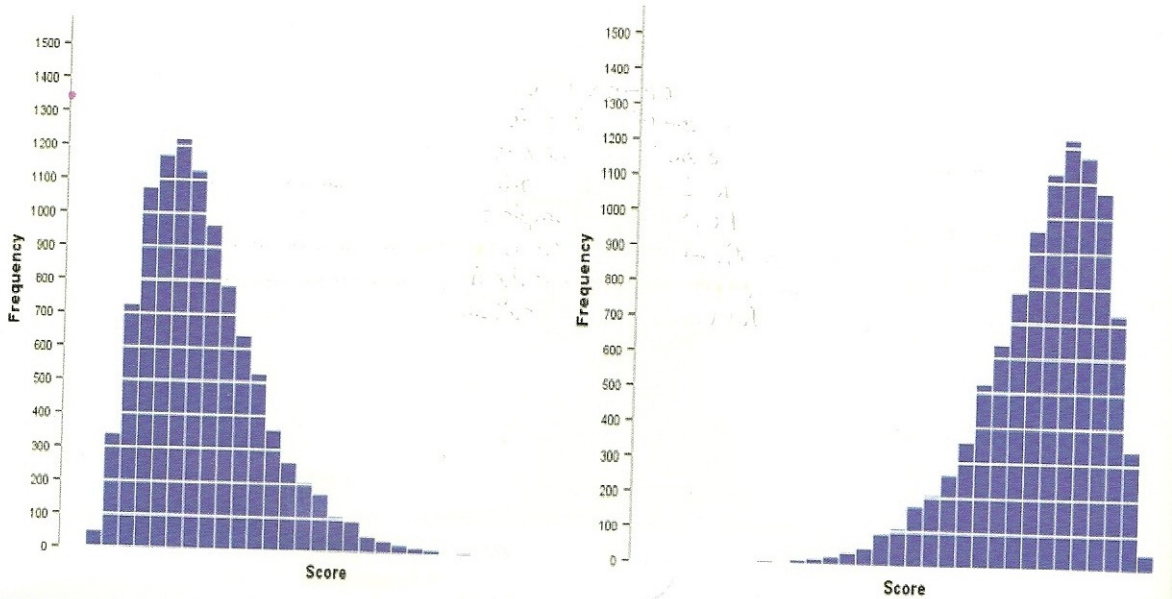


FIGURE 1.4 A positively (left figure) and negatively (right figure) skewed distribution

Kurtosis

- **Leptokurtic distribution**: A distribution with positive kurtosis has many scores in the tails (ends of the distribution) and is pointy.
- **Platykurtic distribution:** A distribution relatively thin in the tail with negative kurtosis and tends to be flatter than normal.
- **Mesokurtic distribution:** A distribution shape that is neither overly peaked nor overly flat.
- **If a distribution has values of skew and kurtosis above or below 0 then this indicates a deviation from normal.**

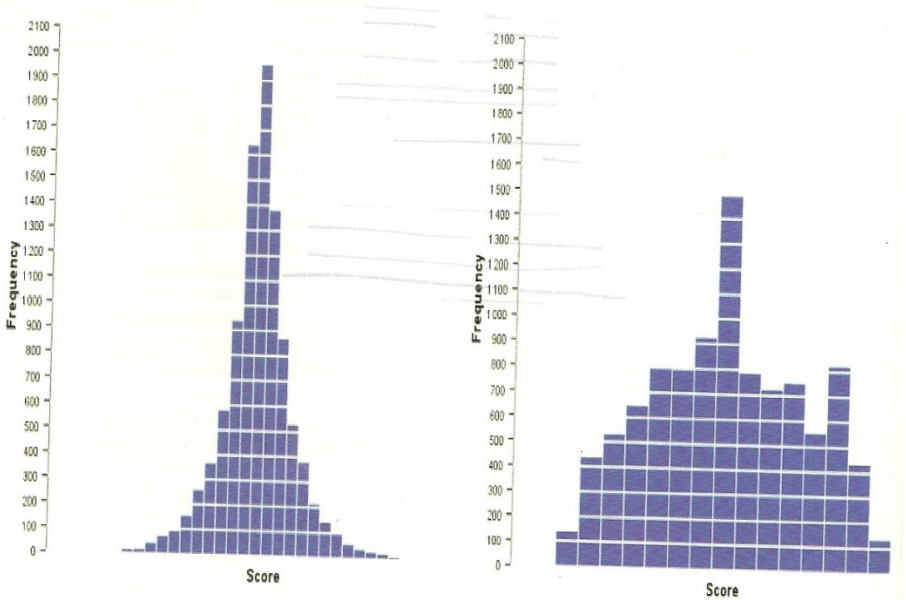
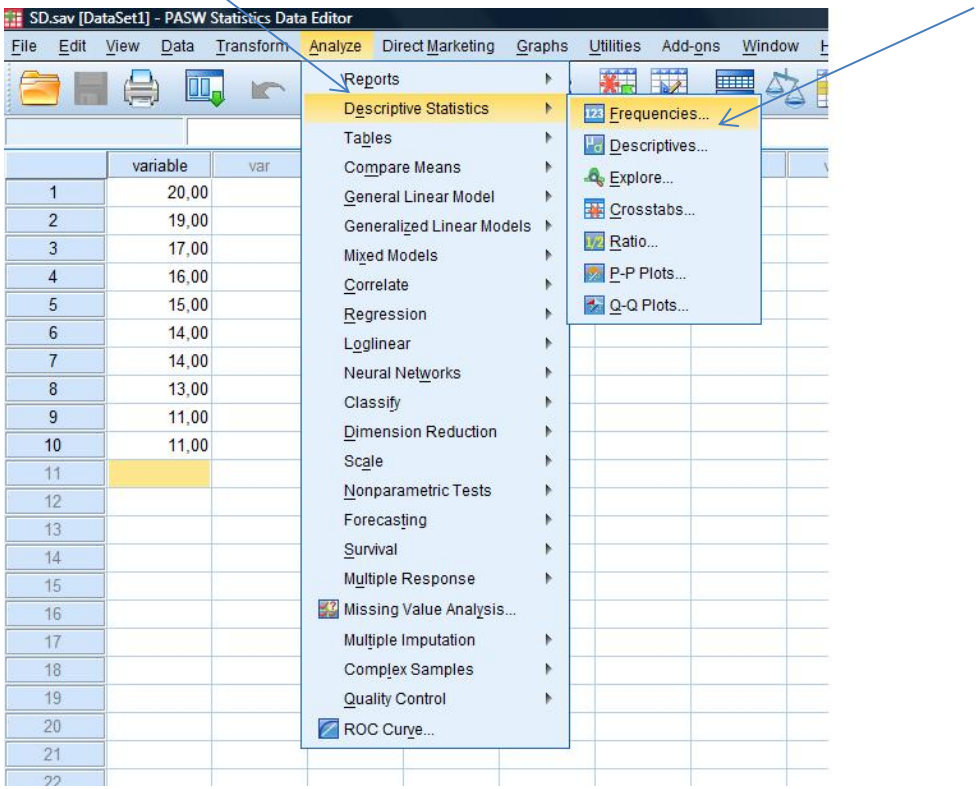
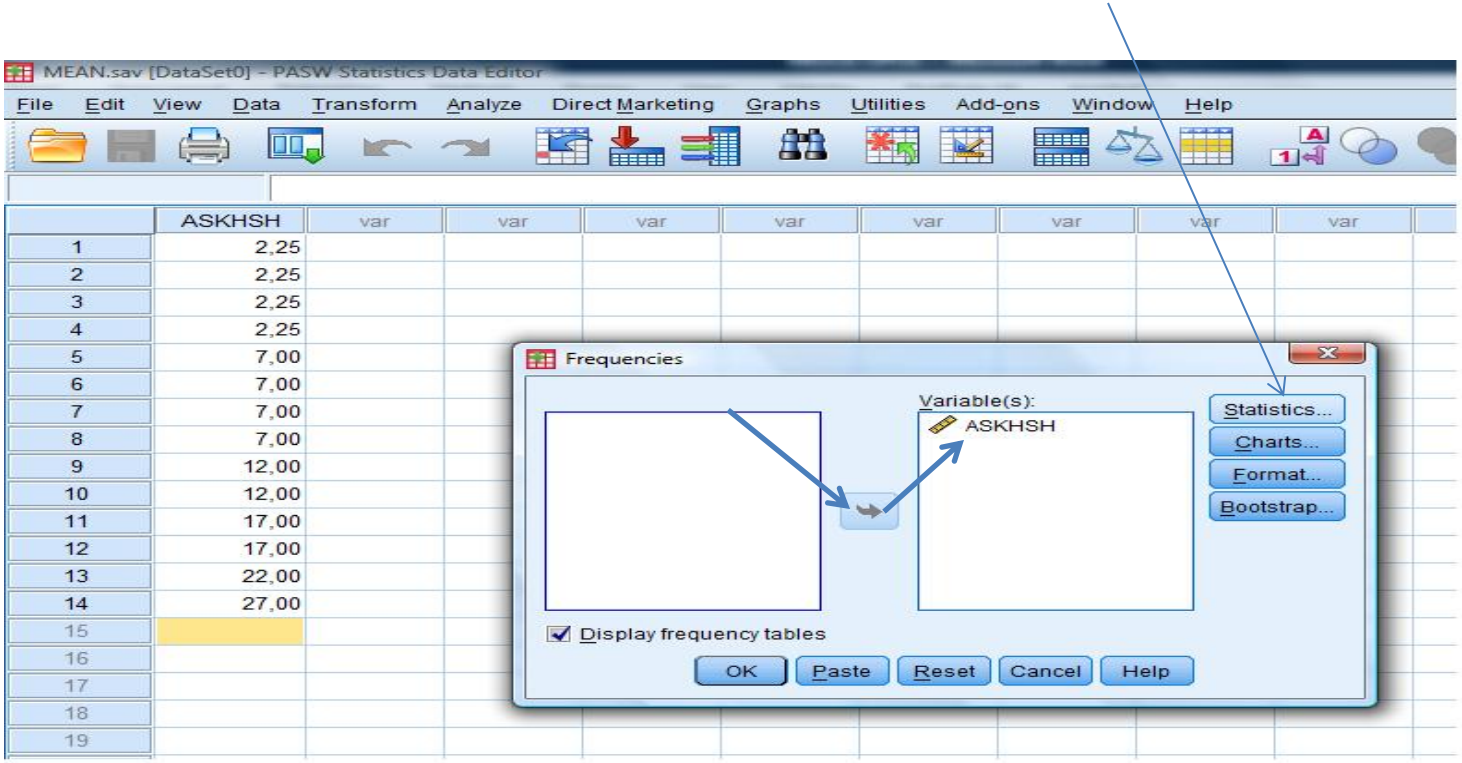
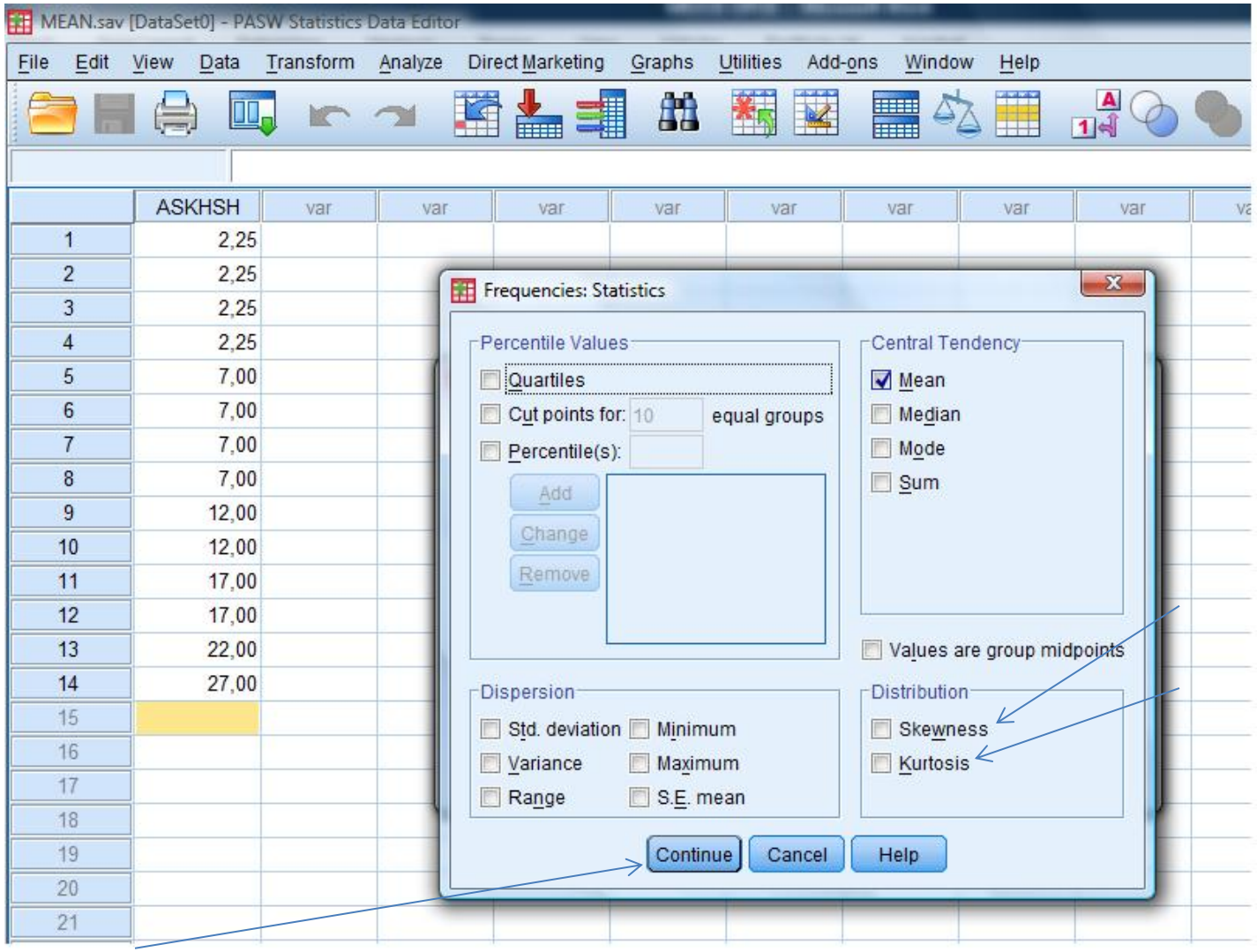


FIGURE 1.5 Distributions with positive kurtosis (leptokurtic, left figure) and negative kurtosis (platykurtic, right figure)

Check Skew and Kurtosis, Mdn, Mo in SPSS

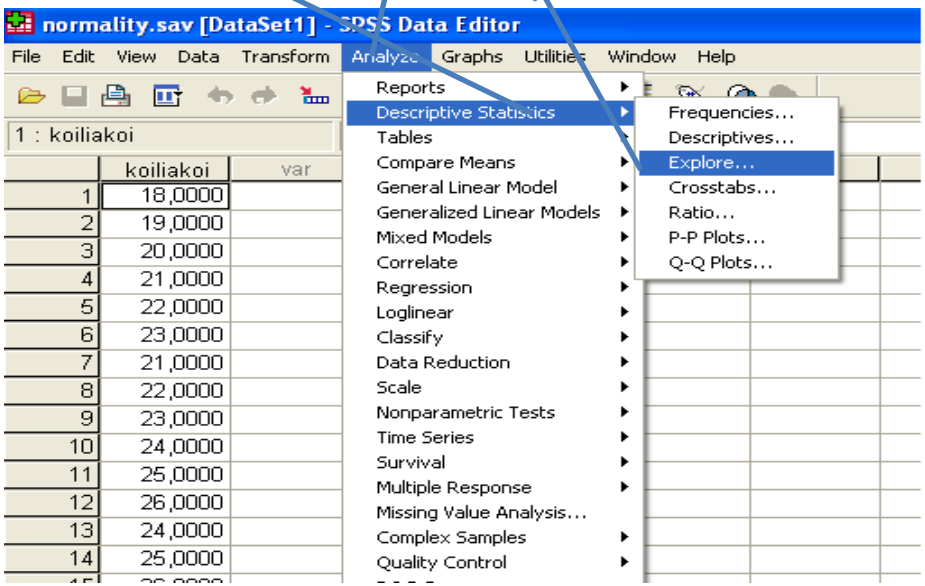


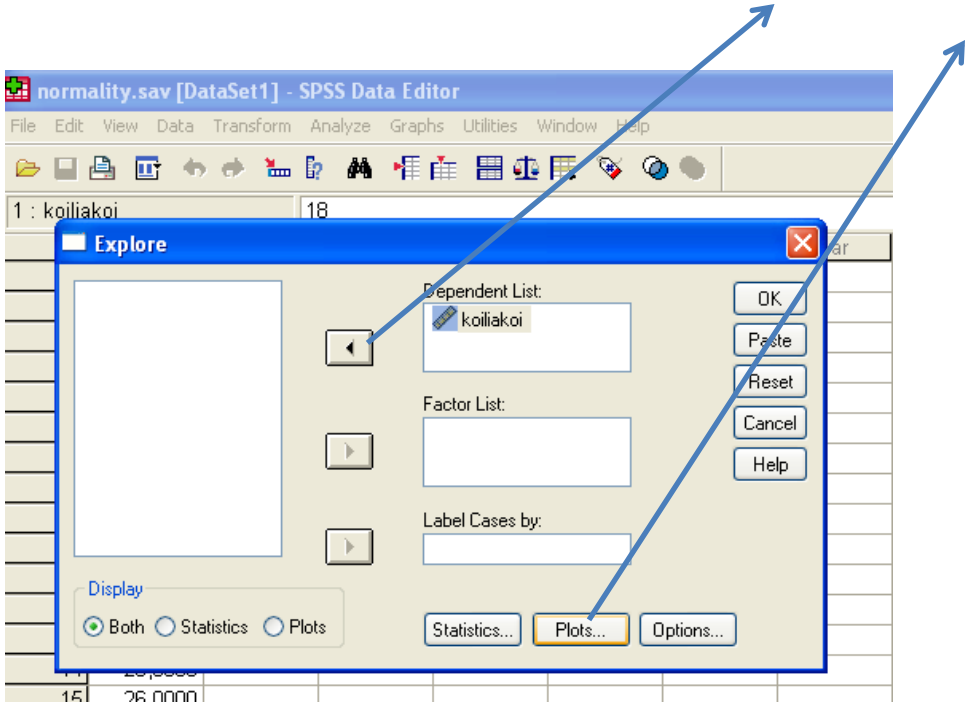

MEAN.sav [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
ASKHSH
1 2,25
2 2,25
3 2,25
4 2,25
5 7,00
6 7,00
7 7,00
8 7,00
9 12,00
10 12,00
11 17,00
12 17,00
13 22,00
14 27,00
15
16
17
18
19
Frequencies
Variable(s):
ASKHSH
Statistics...
Charts...
Format...
Bootstrap...
Display frequency tables
OK
Paste
Reset
Cancel
Help

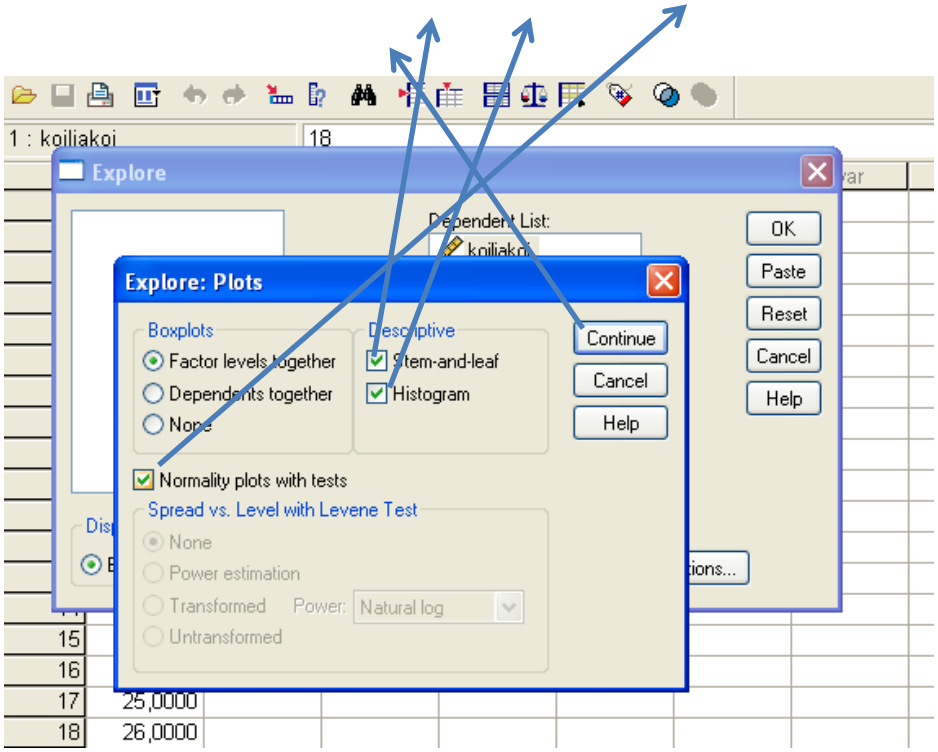

MEAN.sav [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
ASKHSH
1 2,25
2 2,25
3 2,25
4 2,25
5 7,00
6 7,00
7 7,00
8 7,00
9 12,00
10 12,00
11 17,00
12 17,00
13 22,00
14 27,00
Frequencies: Statistics
Percentile Values
Quartiles
Cut points for: 10 equal groups
Percentile(s):
Add
Change
Remove
Central Tendency
Mean
Median
Mode
Sum
Values are group midpoints
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Distribution
Skewness
Kurtosis
Continue
Cancel
Help

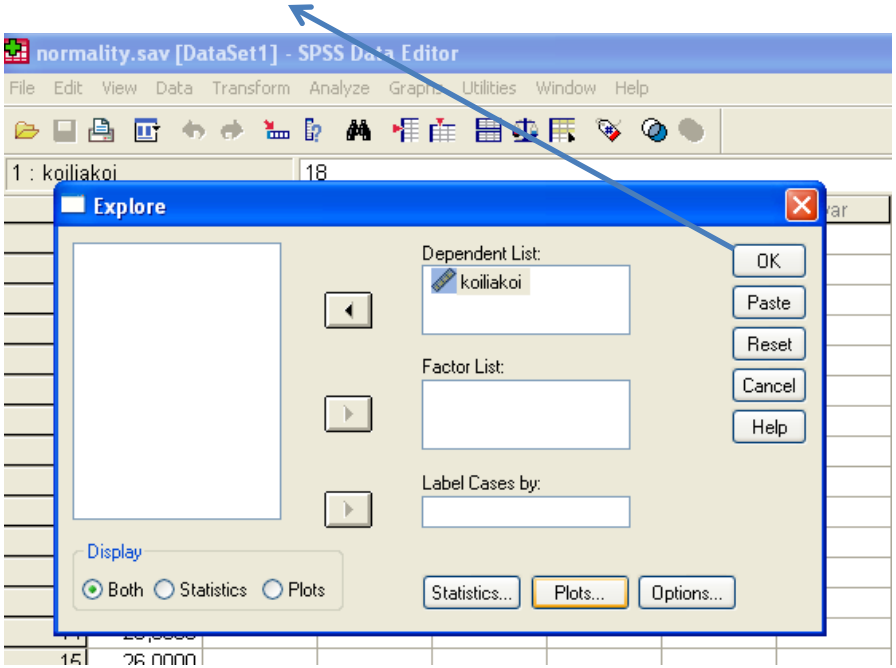
Check normality in SPSS

1) Click ***Analyze*** then ***Descriptive Statistics*** and then ***Explore***




normality.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : koiliakoi
18
Explore
Dependent List:
koiliakoi
Factor List:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Display
Both
Statistics
Plots
Statistics...
Plots...
Options...


1 : koiliakoi
18
Explore
Dependent List:
koiliakoi
OK
Paste
Reset
Cancel
Help
Explore: Plots
Boxplots
Factor levels together
Dependents together
None
Descriptive
Stem-and-leaf
Histogram
Continue
Cancel
Help
Normality plots with tests
Spread vs. Level with Levene Test
None
Power estimation
Transformed
Power:
Natural log
Untransformed
15
16
17
25,0000
18
26,0000


normality.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : koiliakoi
18
Explore
Dependent List:
koiliakoi
Factor List:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Display
Both
Statistics
Plots
Statistics...
Plots...
Options...

Measures of central tendency

- Researchers almost always say something about the typical or representative score in the group.
- There are three measures commonly used
- MODE
- MEDIAN
- MEAN

Provide a numerical index of the average score in the distribution

Mode

- The **mode** is simply the most frequently occurring score.
- Example 6,2,5,1,2,9,3,6,2 mode = 2
- To calculate the mode, simply place the data in ascending order, count how many times each score occurs, and the score that occurs the most is the mode!

Median

- Another way to quantify the center of a distribution is to look for the middle score when scores are ranked in order of magnitude: median.
- Example: 15 observations

35,36,37,40,42,42,42,45,46,47,4,9,49,51,51

Median 8th

35,36,37,40,40,42,42,42,45,46,47,4,9,49,51,51

8th 9th

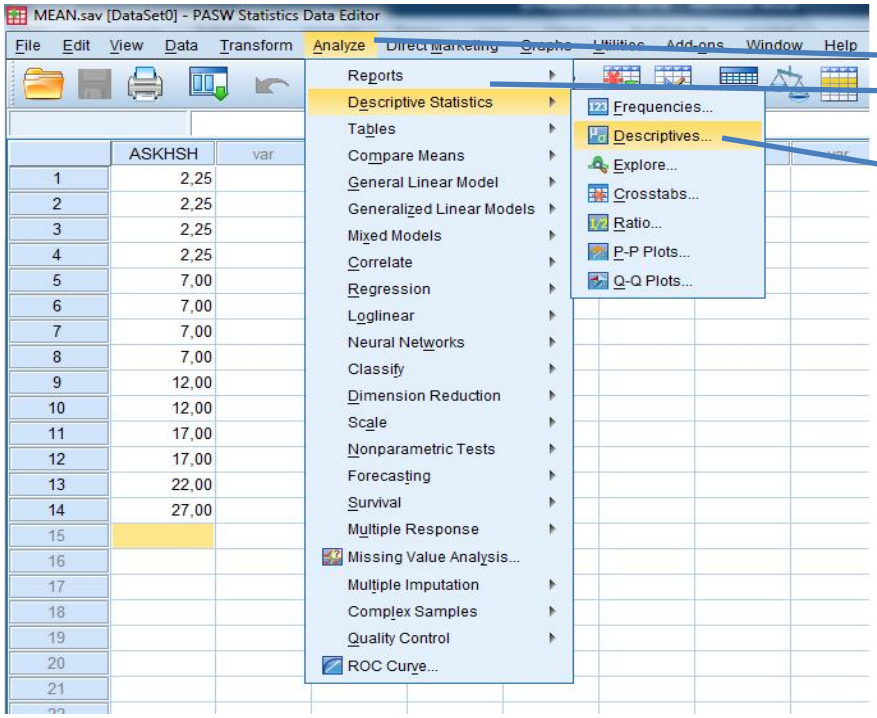
Median = 42+45/2 = 43.5

Formula for Median = n+1/2 th observation

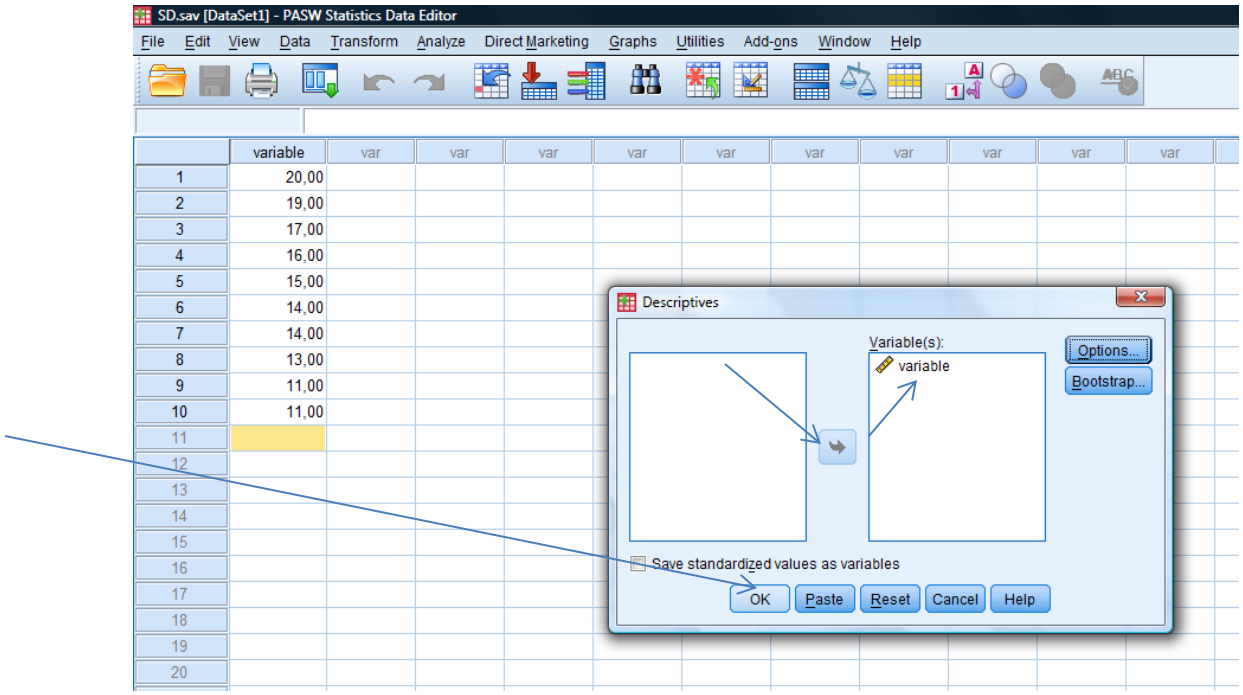
Mean

- Most commonly known measure of the average.
- To calculate the mean we simply add up all of the scores and then divide by the total number if scores we have.
- $M = \Sigma_{\chi} / n$, Σ = sigma and means sum of
- χ = observations, n = the number of observations
- One disadvantage of the mean is that it can be influenced by extreme scores

Check M & SD in SPSS



1) ***Analyze***
Descriptive Statistics
Descriptives


SD.sav [DataSet1] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
variable
1 20,00
2 19,00
3 17,00
4 16,00
5 15,00
6 14,00
7 14,00
8 13,00
9 11,00
10 11,00
Descriptives
Variable(s):
variable
Options...
Bootstrap...
Save standardized values as variables
OK
Paste
Reset
Cancel
Help

Measures of variability

- Although terms (e.g., roughly “normal”) and number (e.g., M = 5.51) help, they are not sufficient. To get a true feel for the data that have been collected, we also need to be told something about the variability among the scores.

Meaning of variability

- Most groups of scores possess some degree of variability. That is, at least some of the scores vary from one another.
- A **measure of variability** simply indicates the degree of this dispersion among the set of scores.
- If the scores are very similar, they are **homogeneous** (low variability
- If the scores are dissimilar, they are **heterogeneous** (high variability)

Variability

- Even though a measure of central tendency provides a numerical index of the average score in a group, we need to know the variability of the scores to better understand the entire group of scores.
- Example
- Group 1: IQ scores= 102,99,103,96, M = 100, homogeneous, low variability
- Group 2: IQ scores = 128,78,93,101, M= 100,

Heterogeneous, high variability

Range

- The easiest way to look at dispersion is to take the largest score and subtract from it the smallest score.
- Example: 22,40,53,57,93,98,103, 108, 116,121,252. Highest = 252, Lowest = 22; 252-22 = 230.
- Range can be affected dramatically by extreme scores.
- One way around this problem is to calculate the range when we exclude values at the extremes of the distribution. One convention is to cut off the top 25% of scores and calculate the range of the middle 50% of scores- **interquartile range.** First we need to calculate what are called **quartiles.**

Quartiles

Example: 22,40,53,57,93,98,103,108, 116,121,252

lower **second** **upper**

Interquartile range= difference between upper and lower = 116 – 53= 63

The advantage is that it isn't affected by extreme scores, but we lose a lot of data

- **Quartiles** are the three values that split the sorted data into four equal parts.
- First we calculate the median, which is called the **second quartile,** which splits our data into two equal parts. Median is 98. The **lower quartile** is the median of the lower half of the data and the **upper quartile** is the median of the upper half of the data. One rule of thumb is that the median is not included in the two halves.

Standard deviation and variance

- The standard deviation and variance are usually better indices of dispersion than the previous measures of variability. They are used when the data are not too skewed or when the mean being used as a measure of the average.
- Both of them are based upon all of the scores in a group and not the high and low scores
- The standard deviation (SD) is determined by 1) figuring how much each score deviates from the mean and 2) putting these deviation scores into a computational formula. In other words, it tells us something about the size of the residuals. A residual is the difference between a particular observation and the mean. The largest is the SD, the greater is the spread of the data
- The **variance** (s^2 or σ^2) is found by squaring the value of the standard deviation

Standard deviation

- SD = $\sqrt{\Sigma d^2 /(N-1)}$, Calculate the residual for each observation (d), square each residual, add all the squared residuals, divide your answer by n-1, square root the whole thing

	X	.d	d^2
1	20	5	25
2	19	4	16
3	17	2	4
4	16	1	1
5	15	0	0
6	14	-1	1
7	14	-1	1
8	13	-2	4
9	11	-4	16
10	11	-4	16
	$\Sigma x = 150$ M=150/10 = 15		$\Sigma d^2 = 84$

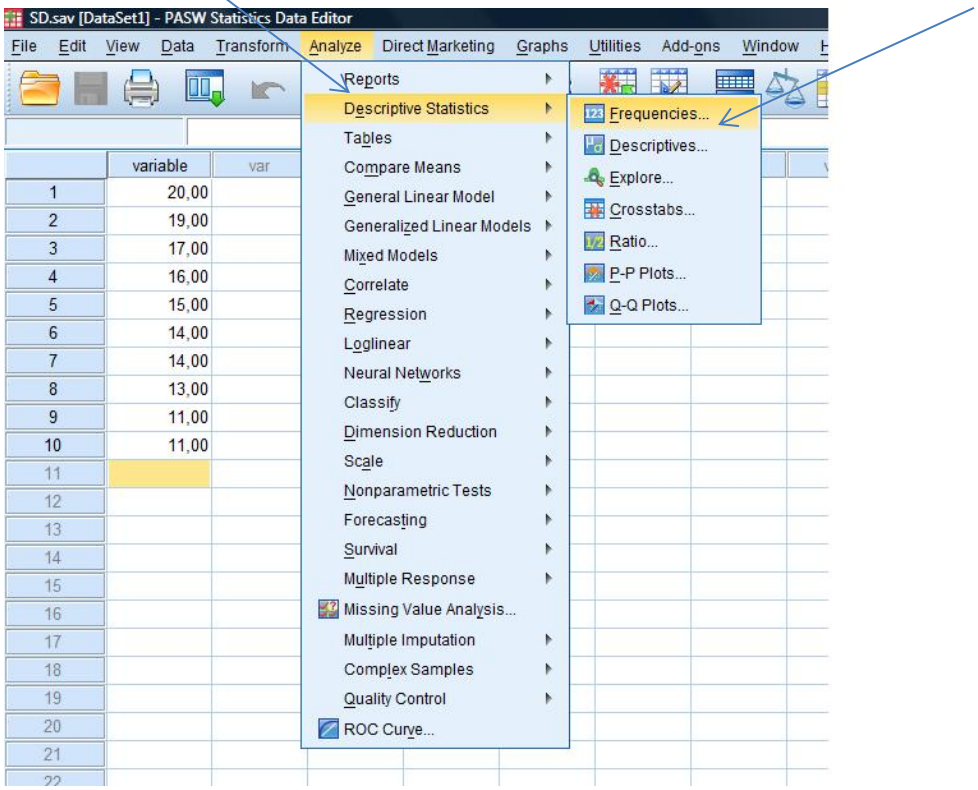
- d = difference between each score and the mean

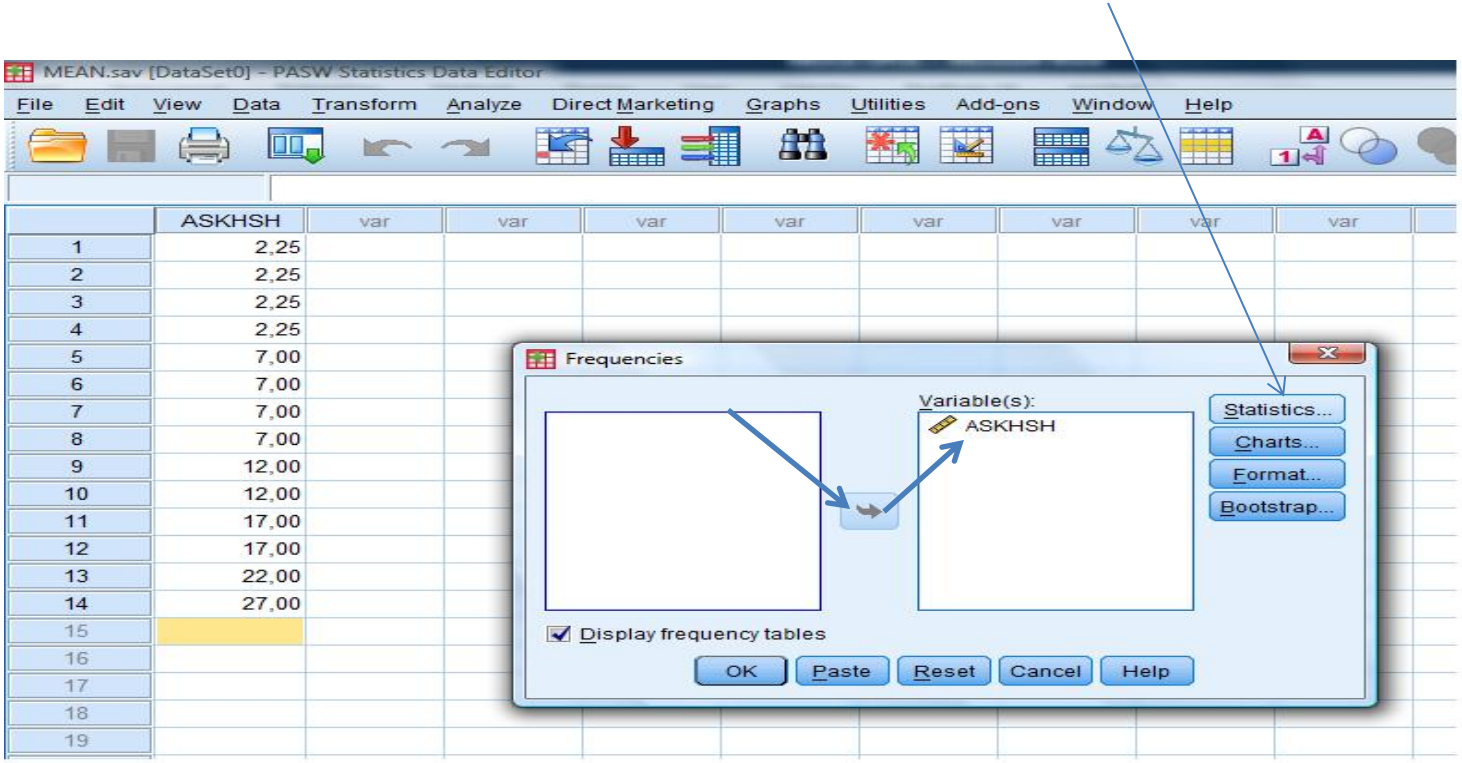
SD = $\sqrt{84/9}$ = $\sqrt{9.33}$ = **3.055**

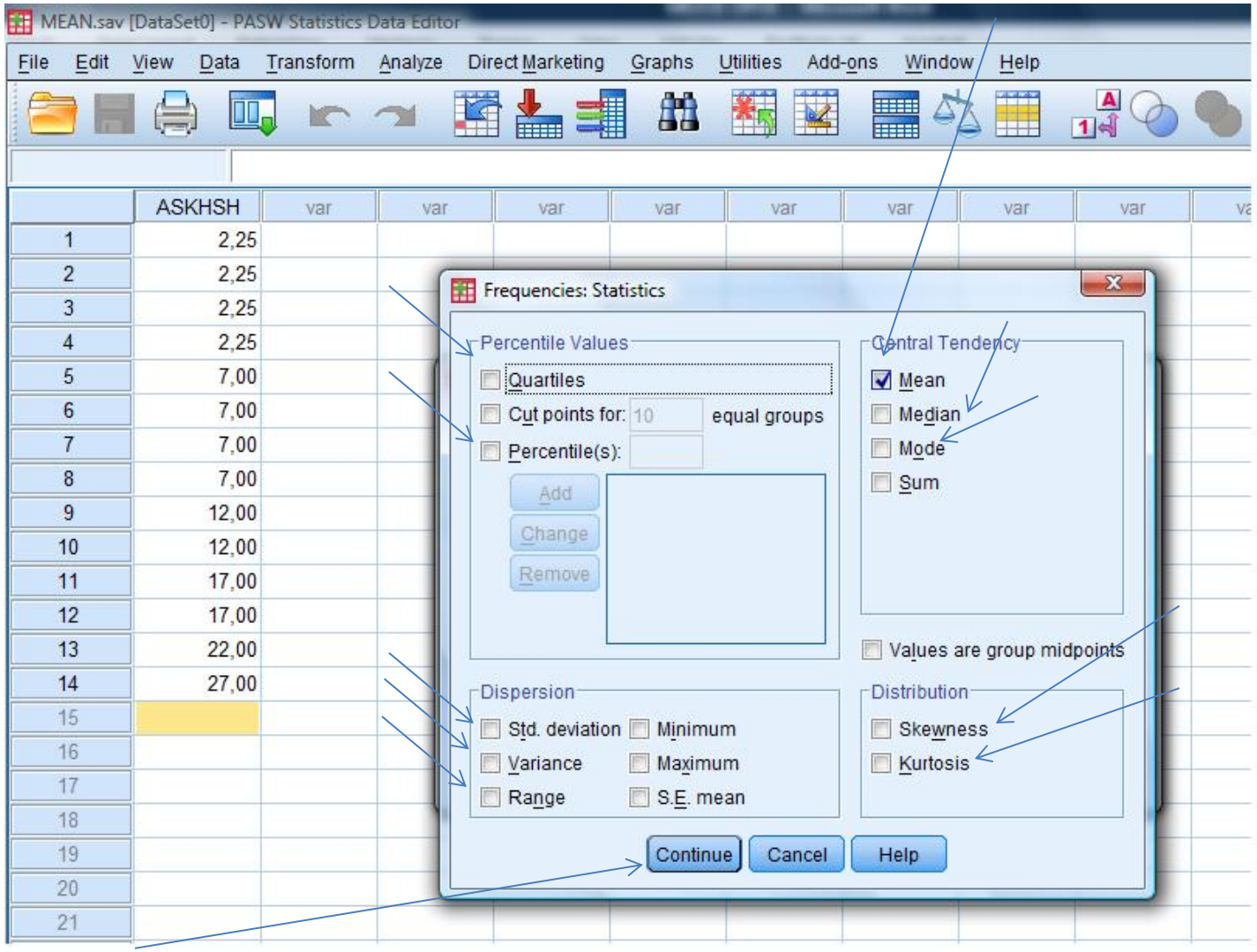
Variance

- Omitting the last stage of the standard deviation formula where the value is square rooted will give the variance instead of the standard deviation.

In SPSS


SD.sav [DataSet1] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window
variable
var
1 20,00
2 19,00
3 17,00
4 16,00
5 15,00
6 14,00
7 14,00
8 13,00
9 11,00
10 11,00
11
12
13
14
15
16
17
18
19
20
21
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...


MEAN.sav [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
ASKHSH
1 2,25
2 2,25
3 2,25
4 2,25
5 7,00
6 7,00
7 7,00
8 7,00
9 12,00
10 12,00
11 17,00
12 17,00
13 22,00
14 27,00
15
16
17
18
19
Frequencies
Variable(s):
ASKHSH
Statistics...
Charts...
Format...
Bootstrap...
Display frequency tables
OK
Paste
Reset
Cancel
Help


MEAN.sav [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
ASKHSH
1 2,25
2 2,25
3 2,25
4 2,25
5 7,00
6 7,00
7 7,00
8 7,00
9 12,00
10 12,00
11 17,00
12 17,00
13 22,00
14 27,00
Frequencies: Statistics
Percentile Values
Quartiles
Cut points for: 10 equal groups
Percentile(s):
Add
Change
Remove
Central Tendency
Mean
Median
Mode
Sum
Values are group midpoints
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Distribution
Skewness
Kurtosis
Continue
Cancel
Help