St. John's University

Center for Educational Research
Leadership and Accountability

Educational Research and Data Analysis II
EDU 7211

Dr. Francesco Ianni

Distribution Analysis

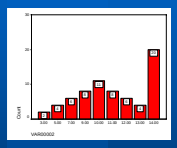
The Location of a distribution

- **Central tendency measures the center of the distribution.**
- **There are three ways to measure the center of the distribution**
  - **MEAN**
  - **MEDIAN**
  - **MODE**

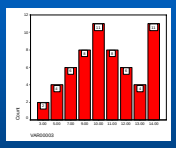
Mode

- **The Mode is the number that appears the most.**
- **The Mode is not affected by the outliers**
- **A distribution can be:**
  - **Unimodal**
  - **Bimodal**
  - **Multimodal**

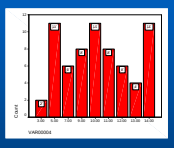
The Mode



Unimodal

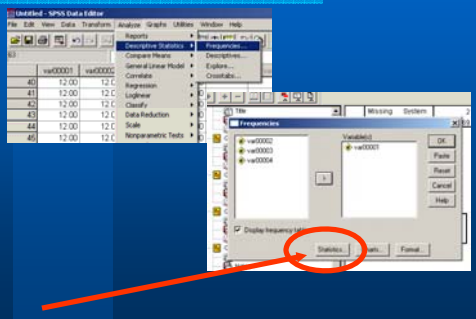


Bimodal

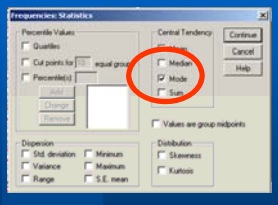


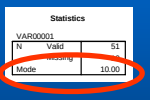
Multimodal

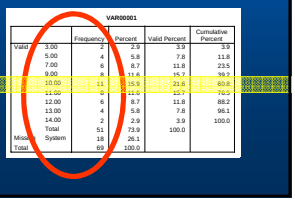
Mode on SPSS



Mode on SPSS







The Median

- The Median is the value that divides the distribution in two halves.
- The Median is the Q2 in the Box Plot
- The Median is a ***Resistant Statistic*** because it is not affected by the numerical value of the outliers

How do we calculate the Median?

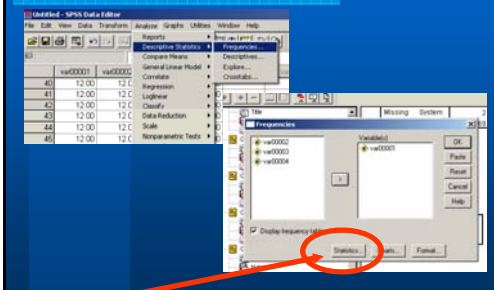
- First order the data from low – high
- If there is an odd number of values
  - Pick the middle number
- If there is an even number of values
  - Average the two middle numbers

Example

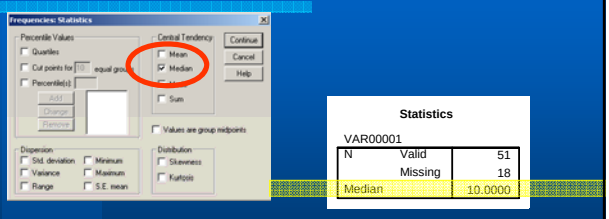
- 2,5,7,8,-3
- -3,2,5,7,8
- The Median is 5

- 3,6,7,1,3,9
- 1,3,3,6,7,9
- The median is $\frac{3+6}{2}=4.5$

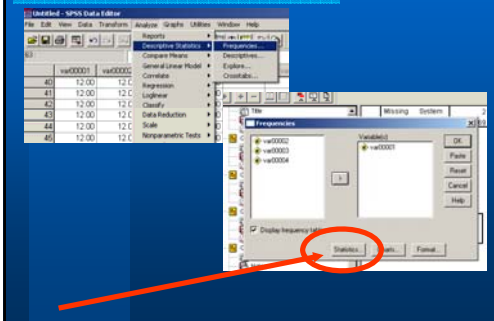
Median on SPSS



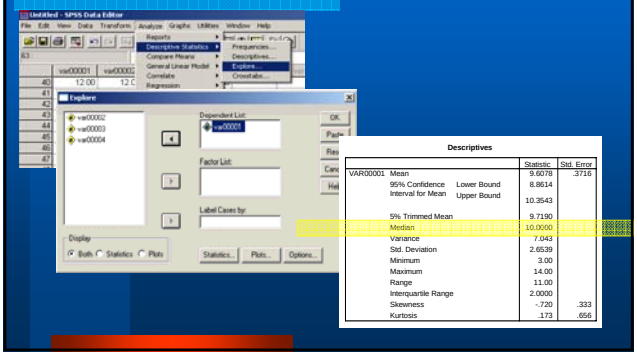
Median on SPSS



Median on SPSS



Another way to get the Median



Descriptives

			Statistic	Std. Error
VAR00001	Mean		9.6078	.3716
	95% Confidence Interval for Mean	Lower Bound	8.8614	
		Upper Bound	10.3543	
	5% Trimmed Mean		9.7190	
	Median		10.0000	
	Variance		7.043	
	Std. Deviation		2.6539	
	Minimum		3.00	
	Maximum		14.00	
	Range		11.00	
	Interquartile Range		2.0000	
	Skewness		-.720	.333
	Kurtosis		.173	.656

The Mean

- The arithmetic mean or simply mean is the most used measure of location
- Affected by the value of the outliers in the distribution
- There is only one mean in a distribution

How do we calculate the Mean?

- Add all the values in the distribution and divide by the total number of values

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Let's analyze

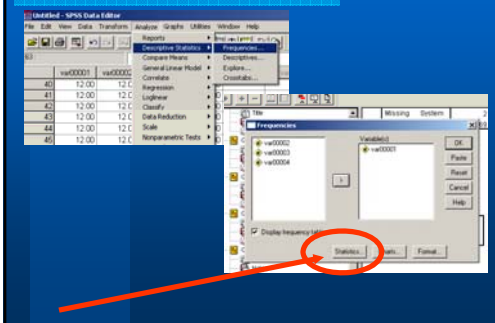
$\bar{X}$ = *The mean*

N = *The number of data values*

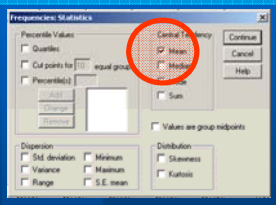
X_i = *Each individual data point for I = 1,2,3,….N*

$\sum_{i=1}^{N} X_i$ = *The sum of all the values from 1 to N*

The Mean on SPSS



The Mean on SPSS

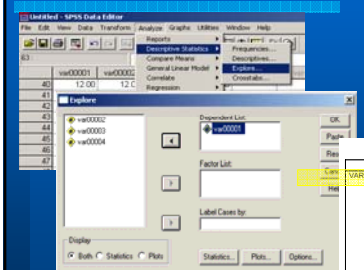


Statistics

NEGSKEW

N	Valid	271
	Missing	15
Mean		-67.9742

Another way to get the Mean



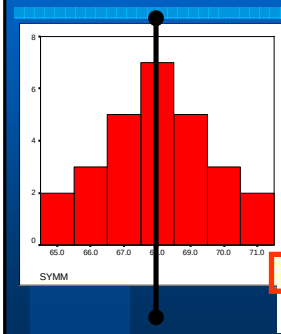
Descriptives

			Statistic	Std. Error
VAR00001	Mean		9.6678	.3718
	95% Confidence Interval for Mean	Lower Bound	8.9014	
		Upper Bound	10.3543	
	5% Trimmed Mean		9.7190	
	Median		10.0000	
	Variance		7.043	
	Std. Deviation		2.6539	
	Minimum		3.00	
	Maximum		14.00	
	Range		11.00	
	Interquartile Range		2.0000	
	Skewness		-.720	.333
	Kurtosis		.173	.656

Comparing the……

- Mean
- Median

In a Symmetric distribution

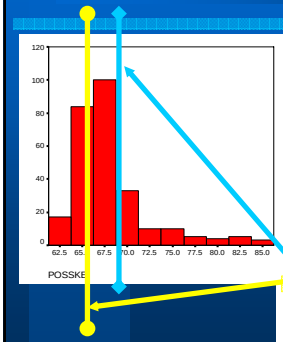


Statistics

SYMM

N	Valid	27
	Missing	259
Mean		68.0000
Median		68.0000
Skewness		.000
Std. Error of Skewness		.448

In a Positively Skewed distribution

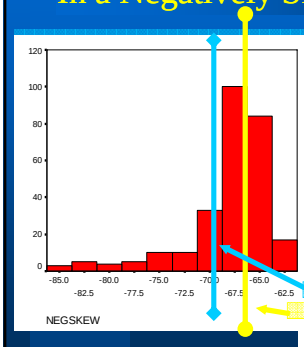


Statistics

POSSKE

N	Valid	271
	Missing	15
Mean		67.9742
Median		67.0000
Skewness		1.996
Std. Error of Skewness		.148

In a Negatively Skewed Distribution



Statistics

NEGSKEW

N	Valid	271
	Missing	15
Mean		-67.9742
Median		-67.0000
Skewness		-1.996
Std. Error of Skewness		.148

In conclusion

- **Mean** = non resistant statistic
- **Median** = resistant statistic

Practice 2

The Spread of a distribution

- Spread of the distribution refers to variability from the average or from one value to another

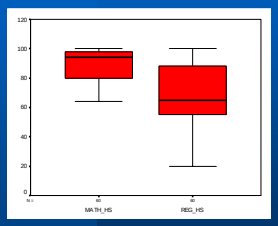
Why variability is important

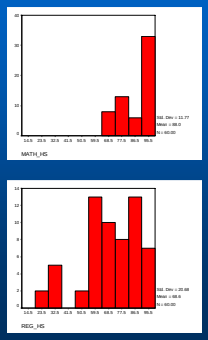
- Central tendency is only part of the story behind the numbers.
- Calculate the average of the following:
  - 8,7,4,4,2
  - 4,5,5,6,5

Why variability is important

- The average is the same, but the second set has less variability
- The next one has no variability at all
  - 5,5,5,5,5

The Spread of a distribution





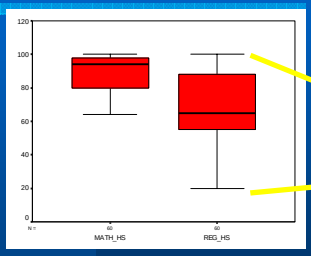
How do we measure the spread

- Range
- Interquartile Range IQR
- Variance
- Standard Deviation

Range

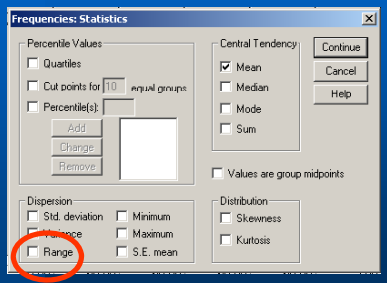
- The range is the difference between the maximum and the minimum value
- The range is a *non resistant statistic*

The Range

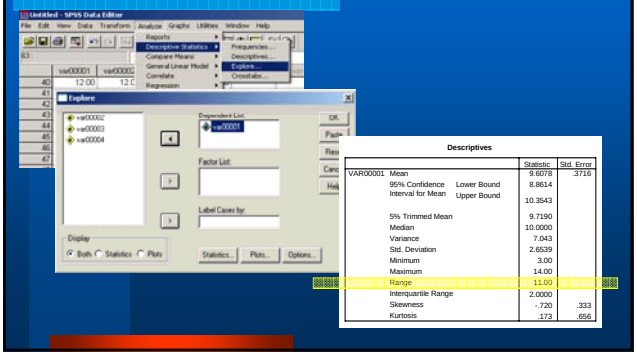


The range is equal to the difference between the two most extreme values.

The Range on SPSS



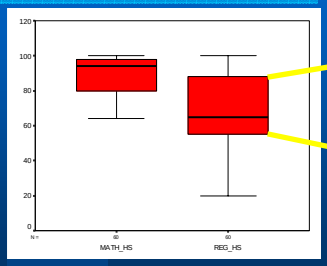
Another way to get the Range



Interquartile Range

- **The IQR is the difference between the 3rd quartile and the 1st quartile.**
- **IQR is a resistant statistic because the value of the IQR is <u>not influenced by extremes.</u>**

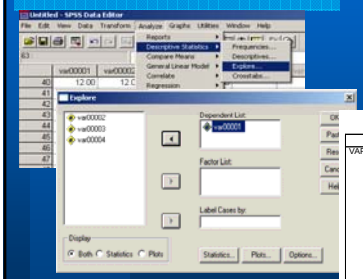
IQR



Quartile 3

Quartile 1

IQR on SPSS



Descriptives

		Statistic	Std. Error
VAR00001	Mean	9.6078	.3716
	95% Confidence Interval for Mean Lower Bound	8.8614	
	95% Confidence Interval for Mean Upper Bound	10.3543	
	5% Trimmed Mean	9.7190	
	Median	10.0000	
	Variance	7.043	
	Std. Deviation	2.6539	
	Minimum	3.00	
	Maximum	14.00	
	Range	11.00	
	Interquartile Range	2.0000	
	Skewness	-.720	.333
	Kurtosis	.173	.656

Calculating Variation

- Variation is something that is around us and we should be able to measure it.
- For example:
  - The salary of different individual in the company varies from year to year
  - The price of a car varies from year to year, state to state.
- That's the reason we are studying variance, standard deviation

Variance

- Definition: The variance is average squared deviations about the mean.
- The variance measures the spread of a distribution.
- Variance is a ***non-resistant*** *statistic*
- Let's analyze the way the formula was derived to have a better understanding.

Given the set of scores…

- 55,57,50,56,62
- The mean = 56
- How far is each value form the mean?

55 – 56 = - 1
57 – 56 = 1
50 – 56 = -6
56 – 56 = 0
62 – 56 = 6

Calculating the variance

- Now from the formula "The variance is the average squared deviations about the mean" we knew that they had to take the average of the numbers just found

-1,1,-6,0,6

- So they did -1 + 1 + -6 + 6 + 0 = 0
- And that became a big problem…….

Calculating the variance

- But they decided to take each of the values

55 – 56 = - 1
57 – 56 = 1
50 – 56 = -6
56 – 56 = 0
62 – 56 = 6

And square them, so the sum will not be zero

Calculating the Variance

- So the values squared gave them a sum of 74 and not zero anymore [1+1+36+0+36 = 74]
- Now we can calculate the average of the squared deviations.

$$s^2 = \frac{\sum (x-\bar{x})^2}{n}$$

Summarizing

- Calculate the mean of the distribution
- Take the difference between each value and the mean
- Square the result
- Sum all the numbers
- Divide by the number of values

$$\frac{\sum (x-\bar{x})^2}{n}$$

Let's try ……

a. 2,3,3,3,3,5,6,7,8,10

b. 5,5,5,5,5,5,5,5,5,5

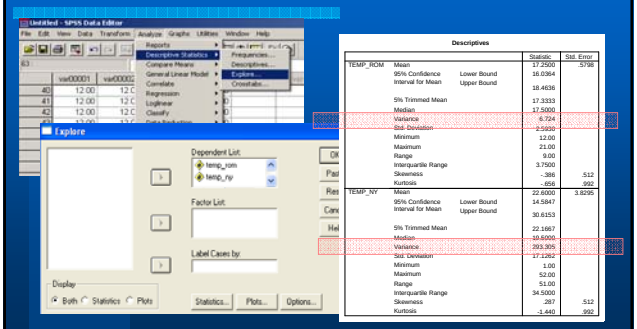
Exercise A

x	$x-\bar{x}$	$(x-\bar{x})^2$
2	-3	9
3	-2	4
3	-2	4
3	-2	4
3	-2	4
5	0	0
6	1	1
7	2	4
8	3	9
10	5	25
Average is =5		the sum is =64

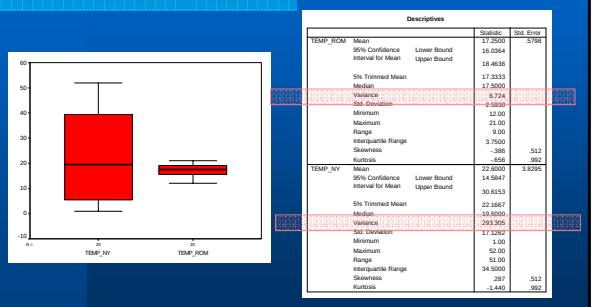
The variance is 6.4

Practice 3

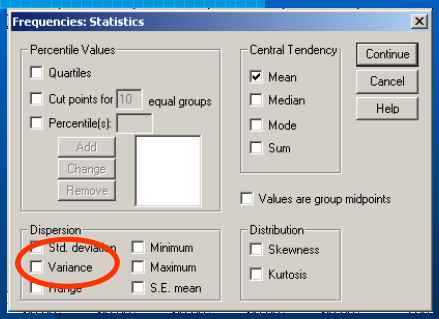
Variance on SPSS



The Variance and the Box Plot



Another way to find the Variance



The Standard Deviation

- **The standard deviation is the positive square root of the variance**

$$sd = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

The Standard Deviation

- The standard deviation is the positive square root of the variance.
- In general, a large SD means that the values are spread out from one another.

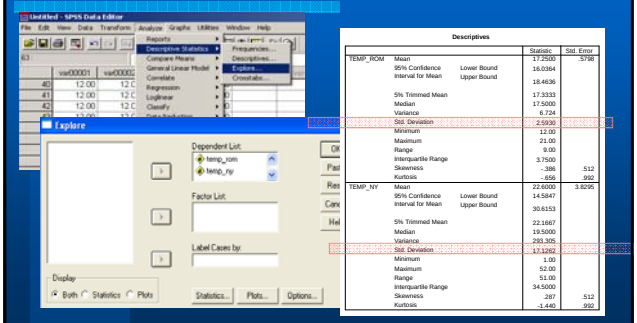
Standard Deviation facts

- It is always positive
- Variance is in square units, SD is not (same units as original data)
- ~68% of the values in a sample will be within 1 sd of the mean
- ~95% of the values in a sample will be within 2 sd of the mean
- ~99.7% of the values in a sample will be within 3 sd of the mean

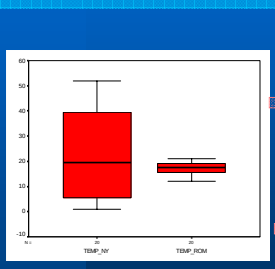
Standard Deviation Facts

- The smallest value for SD is 0 which means that all the values are exactly the same (No Deviation)
- The SD is affected by outliers [NON RESISTANT]

Standard Deviation on SPSS



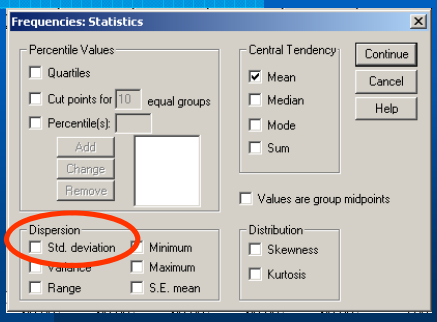
The Standard Deviation and the Box Plot



Descriptives

			Statistic	Std. Error
TEMP_ROM	Mean		17.2500	.5798
	95% Confidence Interval for Mean	Lower Bound	16.0364	
		Upper Bound	18.4636	
	5% Trimmed Mean		17.3333	
	Median		17.5000	
	Variance		6.724	
	Std. Deviation		2.5930	
	Minimum		12.00	
	Maximum		21.00	
	Range		9.00	
	Interquartile Range		3.7500	
	Skewness		-.386	.512
	Kurtosis		-.656	.992
TEMP_NY	Mean		22.6000	3.8295
	95% Confidence Interval for Mean	Lower Bound	14.5847	
		Upper Bound	30.6153	
	5% Trimmed Mean		22.1667	
	Median		19.5000	
	Variance		293.305	
	Std. Deviation		17.1262	
	Minimum		1.00	
	Maximum		52.00	
	Range		51.00	
	Interquartile Range		34.5000	
	Skewness		.287	.512
	Kurtosis		-1.440	.992

Another way to find the Variance



In General

- If the distribution is Skewed
- If the Distribution is NOT Skewed

Practice 4

Practice Exam 1

Transformations

We will learn….

- Linear Transformation
- Other Transformation

- A new graph: Scatterplot
- How to measure a correlation
  - Non Nominal vs. Non Nominal
  - Nominal vs. Non Nominal
  - Nominal vs. Nominal

Linear Transformation

- Addition, Subtraction, Multiplication and Division
  - The effect on the shape of a distribution
  - The effect on summary statistics
  - Common Linear Transformation
  - Standard Scores
  - Z Scores

Multiplication and Division

Multiplication or Division

1. Mean, Median and Mode = OLD */ K
2. SD, IQR, Range = OLD *or / abs(K)
3. Variance = OLD * or / K^2
4. Skewness = OLD if K is positive,
   OPPOSITE if K is negative

Addition and Subtraction

Addition or Subtraction

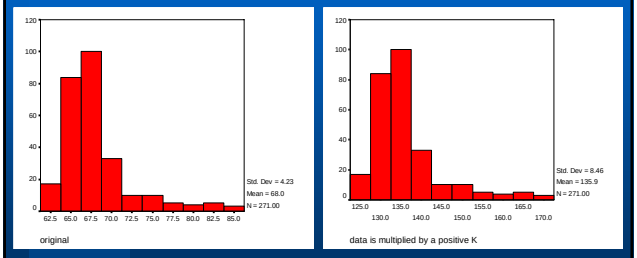
5. Mean, Median and Mode = OLD + or - K
6. SD, IQR, Range = OLD
7. Variance = OLD
8. Skewness = OLD

Let's compute on SPSS

Now let's look at the shape

- If we take a distribution and multiply every value by a positive number K
- What is the shape of the new distribution if the original was positively skewed?

Let's see..

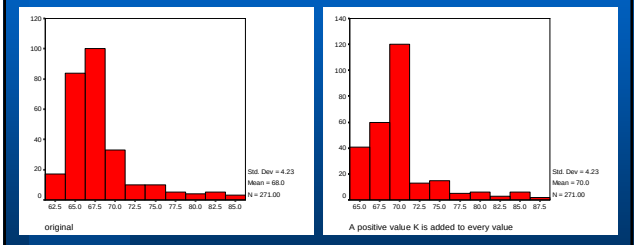


The skewness didn't change

Now let's look at the shape

- **Let's see if the skewness changes when we add a positive number K to every value in the distribution**

Let's see

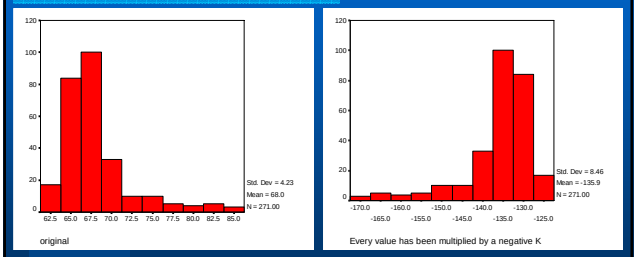


The skewness didn't change

Now let's look at the shape

- **If we take a distribution and multiply every value by a negative number K**
- **What is the shape of the new distribution if the original was positively skewed?**

Let's see



The skewness changed from positive to negative

Please Note

The Shape does not change

- **Except when we multiply or divide by a negative number**

To determine number of SD from Mean

$$z = \frac{x - \mu}{\sigma}$$

Z - Score

- **z – scores represent the number of SD a score is away from the mean.**

$$z = \frac{x - \mu}{\sigma}$$

The relationship between variables

Overview

- **Overview**
- **A new graph: Scatterplot**
- **How to measure a correlation**
  - **Non Nominal vs. Non Nominal**
  - **Nominal vs. Non Nominal**
  - **Nominal vs. Nominal**

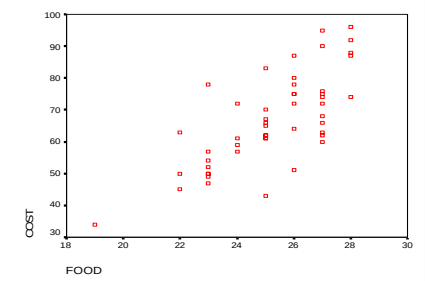
Overview

It is often interesting to know if there is a relationship between two variables

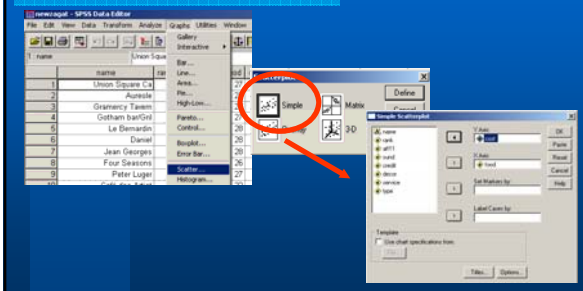
Overview

- **Is the variable hrs of homework associated with test scores?**
- **Is there a relationship between gender and 8th grade self concept?**

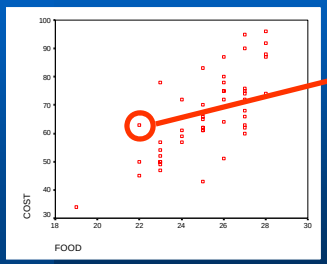
Scatterplot



Scatterplot on SPSS



Scatterplot



Each dot represents……

How to measure a correlation…

- **Both Non Nominal**
  - Income and Age
  - Correlations are described in terms of:
    - Shape, Direction and Strength
  - Correlation statistic
- **One Nominal and One Non Nominal**
  - Box Plots - Scatterplots
    - Gender and 8th grade Self Concept
- **Both Nominal**
  - Cross Tabulations
    - Gender and Ethnicity

Non Nominal vs. Non Nominal

- Describing the relationship
- Correlation Statistic
  - Pearson Correlation Coefficient
- Significance Test

Describing the relationship

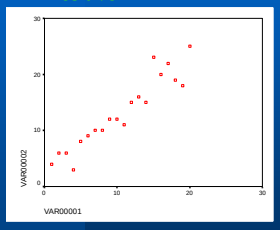
- Relationships are described in terms of:
  - Direction
    - Positive, Negative
  - Strength
    - Strong, Weak, Moderate, None
  - Shape
    - Linear, Non Linear (Quadratic, exponential…)

Direction

- Positive Direction
  - Slope is positive
  - High values of x are associated with high values of y
- Negative Direction
  - Slope is negative
  - High values of x are associated with low values of y

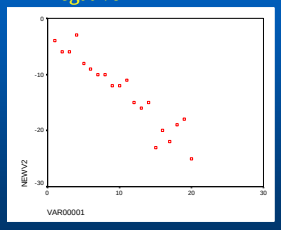
Direction

Positive



High values of x are associated with high values of y

Negative

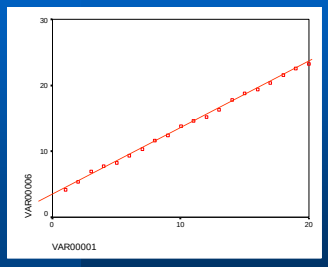


High values of x are associated with low values of y

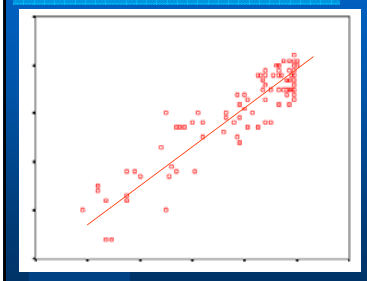
Strength

- **How far the points are from an *imaginary straight line.***
  - ***We will* study the line in the next chapter**
  - **Strong**
  - **Moderate**
  - **Weak**
  - **None**

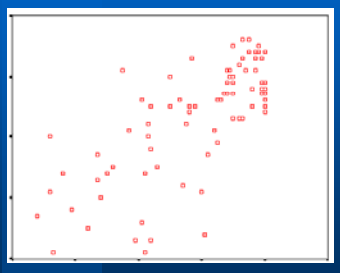
Strength - Strong



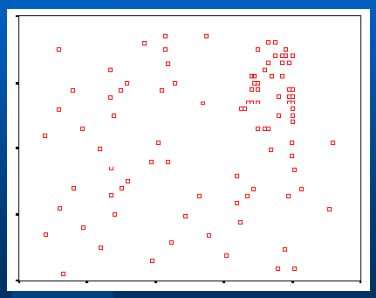
Strength - Moderate



Strength - Weak



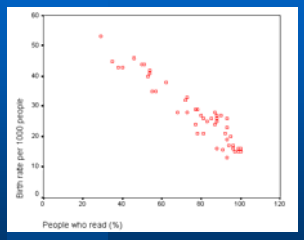
Strength - None

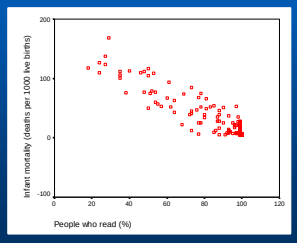


Shape

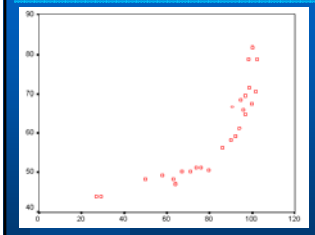
- **Linear**
  - **All the points form an imaginary straight line.**
- **Non Linear**
  - **The points will not form a straight line**
  - **Relationships could be Quadratic, Exponential etc.**

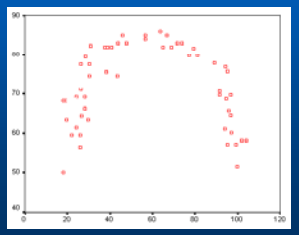
Shape – Linear





Shape – Non Linear





Correlation Statistic

- Pearson's Correlation Coefficient. (r)
- Pearson's correlation measures the strength and the direction of the linear relationship between two variables.
- Pearson's correlation value doesn't describe the shape and cannot determine whether a relationship is linear.

Correlation Statistic

- The Pearson's Correlation Coefficient takes continuous values from – 1 to 1:
  - 1 = Perfect Positive Correlation
  - 0 = No Correlation
  - -1 = Perfect Negative Correlation
  - + or - .5 = Strong
  - + or - .3 = Moderate
  - + or - .1 = Weak
    - The above table is based on Cohen's Scale

Correlation Statistic

- Correlation is not Causation - Correlation is ONLY relation (Association)
- Correlation measures the degree of relationship between two variables.
- The Pearson Correlation does not measures LINEARITY.

Correlation Statistic

- To describe how accurately one variable predicts the other you must square the correlation r.
- Example
  - r = .8 then r square = .64 which means that 64% of the variability in Y scores can be predicted from the relationship with X
- *R square* is called the coefficient of determination because it measures the proportion of variability in one variable that can be determined from the relationship with the other variable.

Practice 5

Significance Test

- **Our main goal is to know if the observed association between variables is the result of chance.**
- **Significance tests help statisticians determine if the association or pattern between variables can be treated as real or as a by product of chance occurrence.**

The Significance Levels

- **SL are estimates of the probability that indicates the degree to which chance is a an explanation for observed association between variables.**

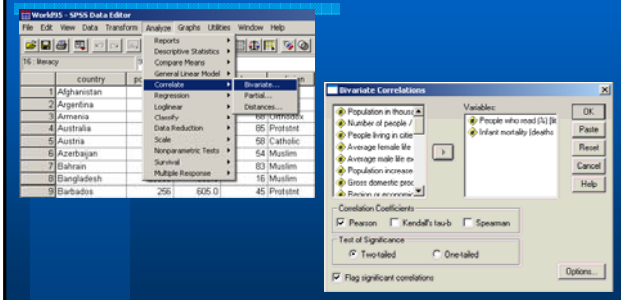
High SL vs. Low SL

- **High Significance Level**
  - **Indicates a strong possibility that chance could explain a pattern.**
  - **This means that there is no relationship between the variables.**
- **Low Significance Level**
  - **Indicates that chance is not the reason to explain the pattern. There is a relationship between the variables.**
  - **In this case, the relationship is considered to be STATISTICALLY SIGNIFICANT**

Statistically Significant

- **The threshold is typically set at .05 and .01. on SPSS = Sig.**
- **.05 = there is a 5 out of 100 possibility that the association happened by chance**
- **.01 = there is a 1 out of 100 possibility that the relationship happened by chance**

Pearson Correlation on SPSS



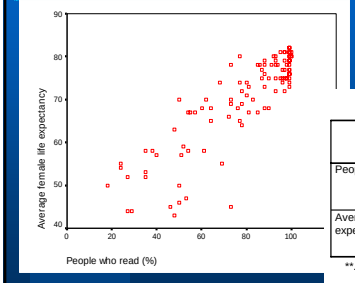
Pearson Correlation on SPSS

Correlations

		People who read (%)	Infant mortality (deaths per 1000 live births)
People who read (%)	Pearson Correlation	1.000	-.900**
	Sig. (2-tailed)	.	.000
	N	107	107
Infant mortality (deaths per 1000 live births)	Pearson Correlation	-.900**	1.000
	Sig. (2-tailed)	.000	.
	N	107	109

**. Correlation is significant at the 0.01 level (2-tailed).

Some examples

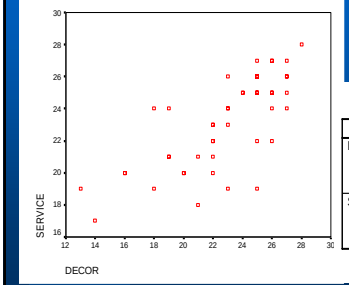


Correlations

		People who read (%)	Average female life expectancy
People who read	Pearson Correla	1.000	.865*
	Sig. (2-tailed)	.	.000
	N	107	107
Average female l expectancy	Pearson Correla	.865*	1.000
	Sig. (2-tailed)	.000	.
	N	107	109

**.Correlation is significant at the 0.01 level (2-tailed).

Some examples

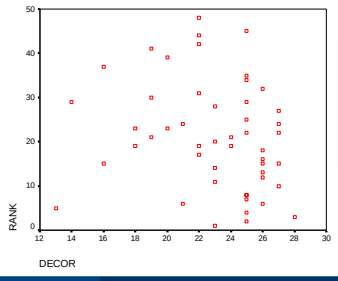


Correlations

		DECOR	SERVICE
DECOR	Pearson Correla	1.000	.743**
	Sig. (2-tailed)	.	.000
	N	50	50
SERVIC	Pearson Correla	.743**	1.000
	Sig. (2-tailed)	.000	.
	N	50	50

**.Correlation is significant at the 0.01 leve

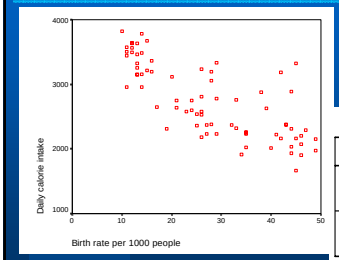
Some examples



Correlations

		DECOR	RANK
DECOF	Pearson Correla	1.000	-.214
	Sig. (2-tailed)	.	.136
	N	50	50
RANK	Pearson Correla	-.214	1.000
	Sig. (2-tailed)	.136	.
	N	50	50

Some examples

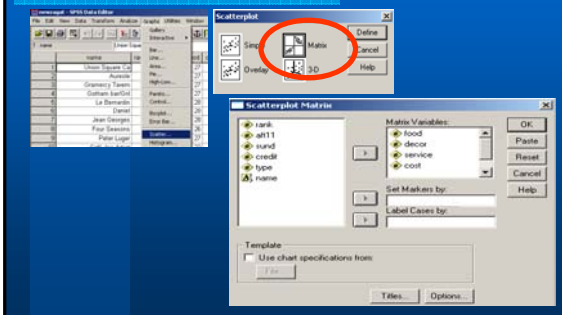


Correlations

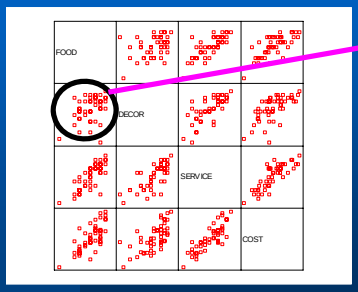
		Daily calorie intake	Birth rate per 1000 people
Daily calorie intake	Pearson Corre	1.000	-.762**
	Sig. (2-tailed)	.	.000
	N	75	75
Birth rate per 1000	Pearson Corre	-.762**	1.000
	Sig. (2-tailed)	.000	.
	N	75	109

**.Correlation is significant at the 0.01 level (2-tailed).

More than one variable



More than one variable



We interpret them the same way.

This scatter plot represents the correlation between

Food and Decor

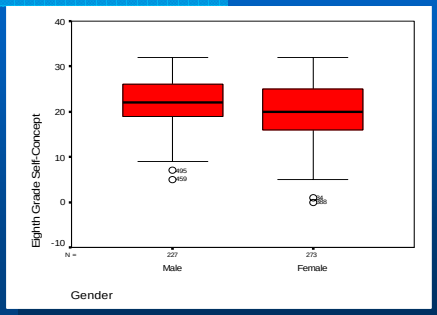
More than one variable

Correlations

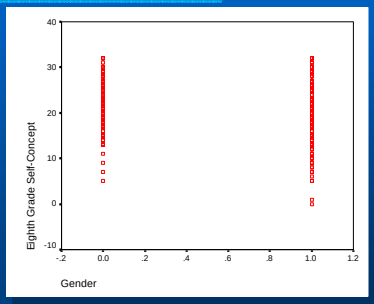
		FOOD	DECOR	SERVICE	COST
FOOD	Pearson Correlation	1.000	.415**	.792**	.734**
	Sig. (2-tailed)	.	.003	.000	.000
	N	50	50	50	50
DECOR	Pearson Correlation	.415**	1.000	.743**	.657**
	Sig. (2-tailed)	.003	.	.000	.000
	N	50	50	50	50
SERVICE	Pearson Correlation	.792**	.743**	1.000	.832**
	Sig. (2-tailed)	.000	.000	.	.000
	N	50	50	50	50
COST	Pearson Correlation	.734**	.657**	.832**	1.000
	Sig. (2-tailed)	.000	.000	.000	.
	N	50	50	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

Nominal vs. Non Nominal



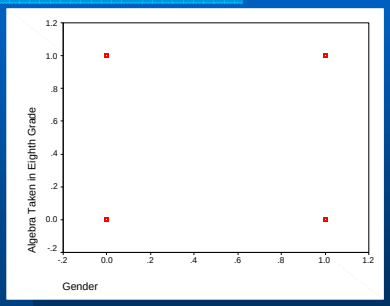
Nominal vs. Non Nominal



Nominal vs. Nominal

- **A problem using the old way**
- **Crosstabs**
- **Creating Crosstabs on SPSS**
- **Clustered Bar Graphs**
- **Additional Examples**

A Problem using the old way



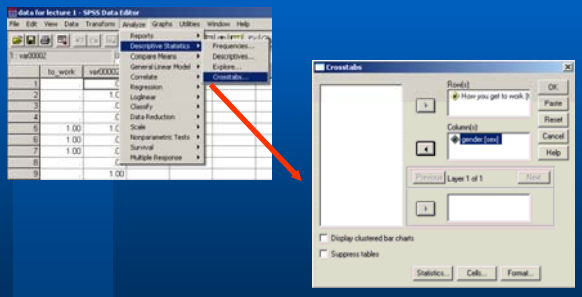
Crosstabs

GENDER * ALGEBRA8 Crosstabulation

Count

		ALGEBRA8		
		No	Yes	Total
GENDER	Male	104	108	212
	Female	124	126	250
Total		228	234	462

Crosstabs on SPSS



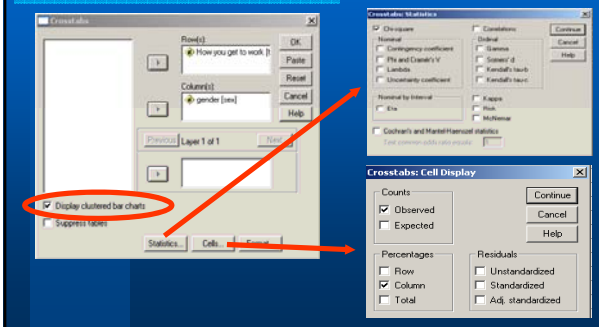
Crosstabs on SPSS

How do you get to work? * gender Crosstabulation

Count

		gender		
		male	female	Total
How do you get to work?	by car	349	323	672
	by bus	86	239	325
	walk	21	48	69
	bike	35	85	120
Total		491	695	1186

Crosstabs with Percentages and Counts



Crosstabs with Percentages and Counts

How do you get to work? * gender Crosstabulation

			gender		
			male	female	Total
How do you get to work?	by car	Count	349	323	672
		% within gender	71.1%	46.5%	56.7%
	by bus	Count	86	239	325
		% within gender	17.5%	34.4%	27.4%
	walk	Count	21	48	69
		% within gender	4.3%	6.9%	5.8%
	bike	Count	35	85	120
		% within gender	7.1%	12.2%	10.1%
Total		Count	491	695	1186
		% within gender	100.0%	100.0%	100.0%

Clustered Bar Graph