INTRODUCTION TO STATISTICS

Session 4

May 1st 2017

Madoka Takeuchi

Review: Measures of Center

	Always exist	Use all the data	Affected by extreme data
Mean	Yes	Yes	Yes
Median	Yes	Yes	No
Mode	No	No	No

- The Mean is used in computing other statistics, such as the variance
  -It is often not appropriate for skewed distributions

- The Median is the center number and is good for skewed distributions because it is resistant to change.

Initials	year	tv	sleep
AB	1	3	6
BC	4	6	7
CD	1	2	5
DE	2	4	6
EF	2	5	8
FG	2	3	5
GH	2	4	6
HI	2	2	6.5
IJ	2	3	7
JK	1	1	3.5
KL	1	3	12
LM	2	3	6

In class exercise- SPSS

- Data was collected on 12 students regarding average daily hours spent watching tv and hours of sleep
- Input the data into spss
  - How was the initial variable inputted?
- Describe the data-give the sample size, measures of center, measures of spread for each variable
- By looking at the relationship between the mean the median, what can be said about the distribution for the hours spent watching tv and hours of sleep?
  - Distribution can be visualized using graphs-boxplot/histograms…

SPSS: Histogram

- ANALYZE> FREQUENCIES>CHART>HISTOGRAMS
- *Option to show normal curve on the histogram*
- Bar chart is very similar to histogram
  - Histogram is always in frequencies but bar chart can be in percentages
  - Build a bar chart to see the difference!

SPSS: Boxplot

- GRAPHS>LEGACY DIALOGS>BOXPLOT
  - Choose simple
  - In the *data in chart are* box choose:
    - *summary of separate variables*- when want to get summary of the variable- i.e boxplot of weight
      - In the previous exercise- boxplot of hours of TV/sleep
    - *Summary for groups of cases*- when want to get a boxplot for a variable separate by category- i.e. boxplot of weight by gender
      - Need to define category in the *category axis* box
      - In the previous exercise- boxplot of hours of TV/sleep by year

Central Limit Theorem (CLT)

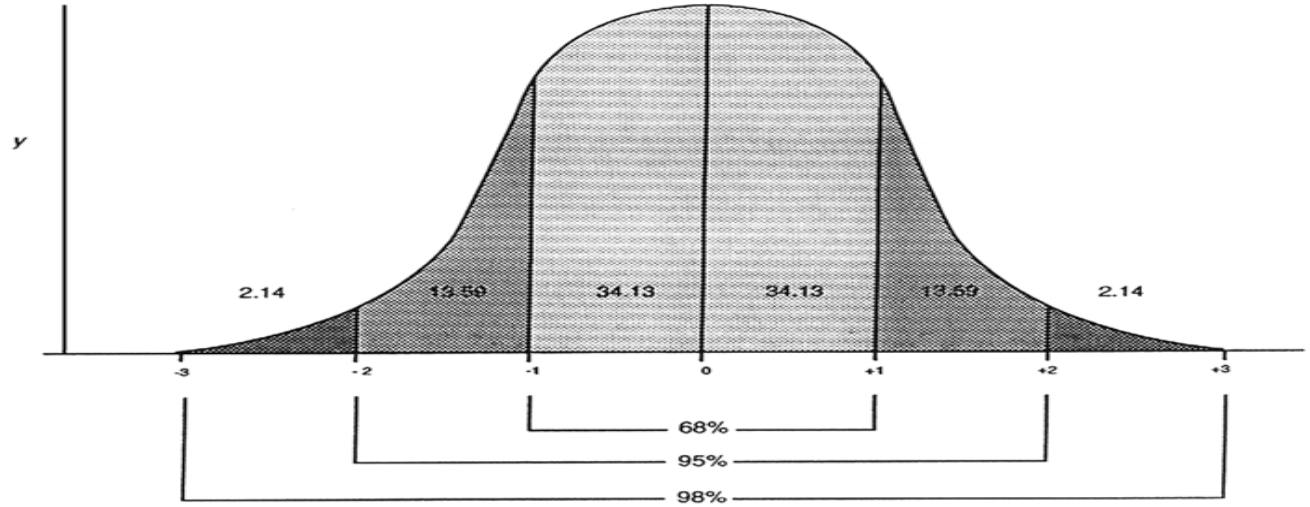
- The central limit theorem explains why many distributions tend to be close to the normal distribution.
  - the central limit theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed variables,(data points) will be approximately normal distributed.

Central Limit Theorem (CLT)

- CLT- if sufficiently large random samples are taken with replacement from a population with mean μand standard deviation σ, the distribution of means of the samples will be approximately normally distributed.
  - This will hold even if the source population is skewed- however n>30
  - If the source population is normally distributed- n can be less than 30

Empirical rule

- The empirical rule is only valid for bell-shaped (normal) distributions.



- Approximately 68% of the data values fall within (±) one standard deviation of the mean.
- Approximately 95% of the data values fall within (±) two standard deviations of the mean.
- Approximately 99.7% of the data values fall within (±) three standard deviations of the mean.

Association of Data

- When exploring data, we are often interested in the relation between the variables.

  i.e. Is there a relation between age and weight?

  Did a company`s marketing strategy increase sales?

  Is owning a pet associated to longer life?

- In this lecture, will focus on the relation between two variables

Response and Explanatory variables

- Response Variable/dependent variable
  - measures the outcome
  - denoted as y
- Explanatory Variable/ independent variable
  - explains the change in the response variable
  - may be the cause of change in the response variable
  - denoted as x

$$y=ax+b$$

Examples

1. A study looks at the affects of alcohol consumption on heart disease
   - Response variable- death due to heart disease
   - Explanatory variable- number of alcohol drinks consumed

2. A study observing the association of fast food and obesity
   - Response variable?
   - Explanatory variable?

Causation vs. Association

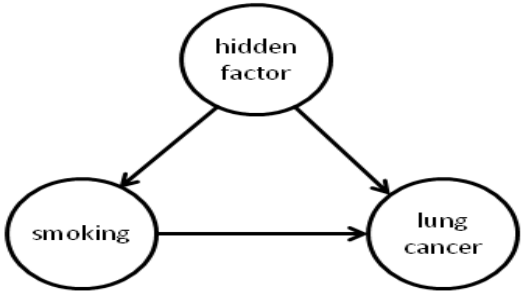
- A study shows that there is a relationship between the hours worked and income earned- increased working hours is associated with an increase in income.
- BUT, does this mean that fewer hours spent at work results in lower pay?

Causation vs. Association/Correlation

- Causation
  - Definition- relationship between two events where one event is affected by the other- *one event is the result of the occurrence of the other event-* **cause and effect**
    - one event, or variable, increases or decreases as a result of other events
      - one event is 100 percent certain to cause the other event
- Association/correlation
  - Definition- "the relationship between things that happen or change together."
    - Although there may be a relationship between two events, one event does not necessarily 100% cause the other event
    - Correlation- the relationship between the events/variables is linear

Causation or Correlation?

- Smoking and lung cancer?
  - We often say that smoking causes lung cancer, but not everyone who smokes develops lung cancer.
    - Statistically- There is a strong correlation between smoking and lung cancer



- Excessive overeating and weight gain
  - One can say that excessive overeating causes weight gain since close to 100% of people who overeat will gain weight
    - Can this be debated?
    - Statistically is this correlation?

Example

- A study looks at the association between the number of hours spent studying for a test and test score
  - A small study collected hours spent studying and test score data from 10 students

- What is the response/dependent variable?
- What is the explanatory/independent variable?

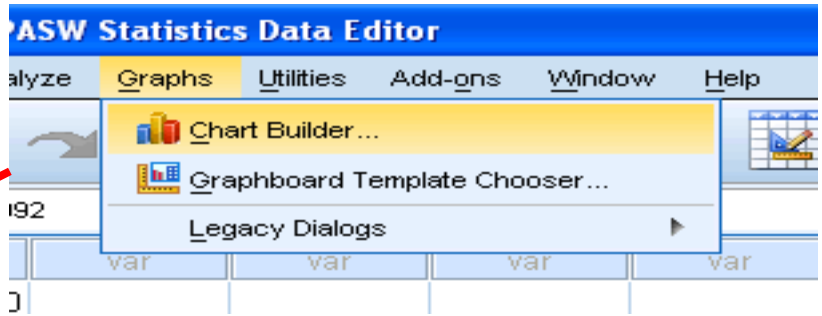
Data from the 10 students- hours spent studying and actual test score

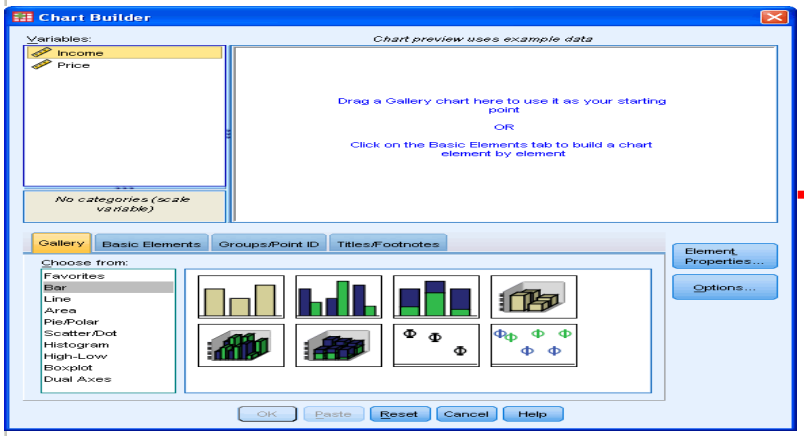
Student	Hours spent	Test Score
1	1	81
2	5	75
3	7	89
4	10	88
5	3	65
6	5	92
7	6	71
8	2	78
9	1	54
10	0	68

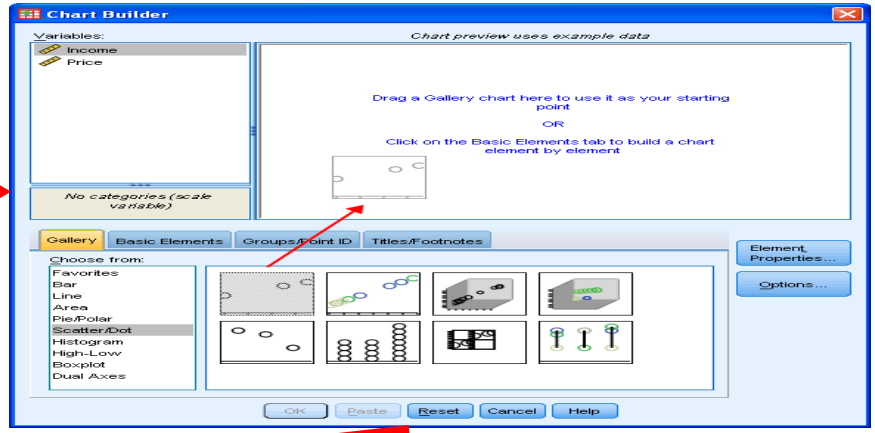
Student dataset in SPSS

- Input the previous data into SPSS
- After inputting the data, want to visualize the data points
  - In SPSS: Graphs→ Chart Builder→ Scatter/dot
  - y/dependent variable= score
  - x/independent variable= time

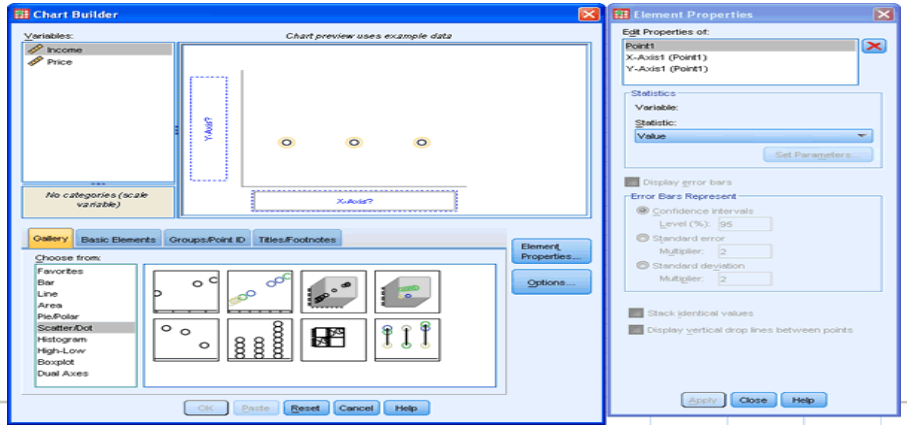
1.Graphs→ Chart Builder





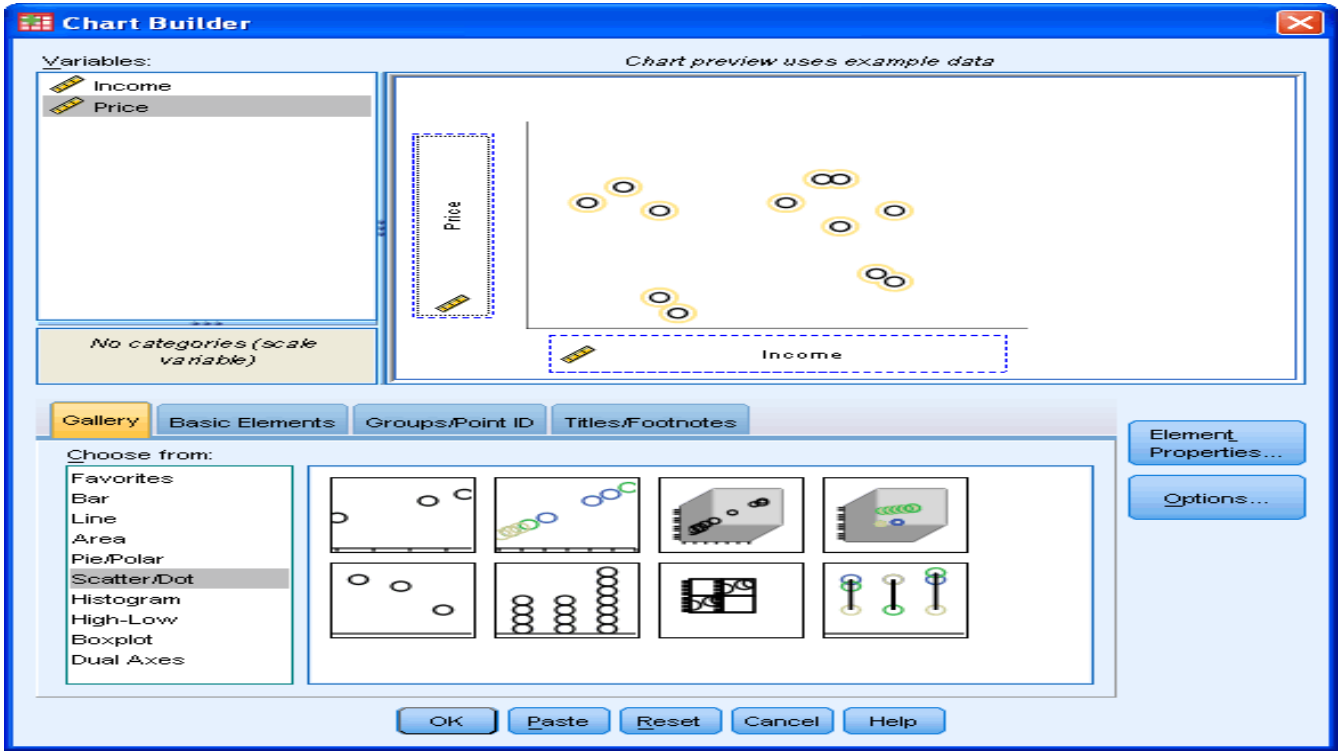


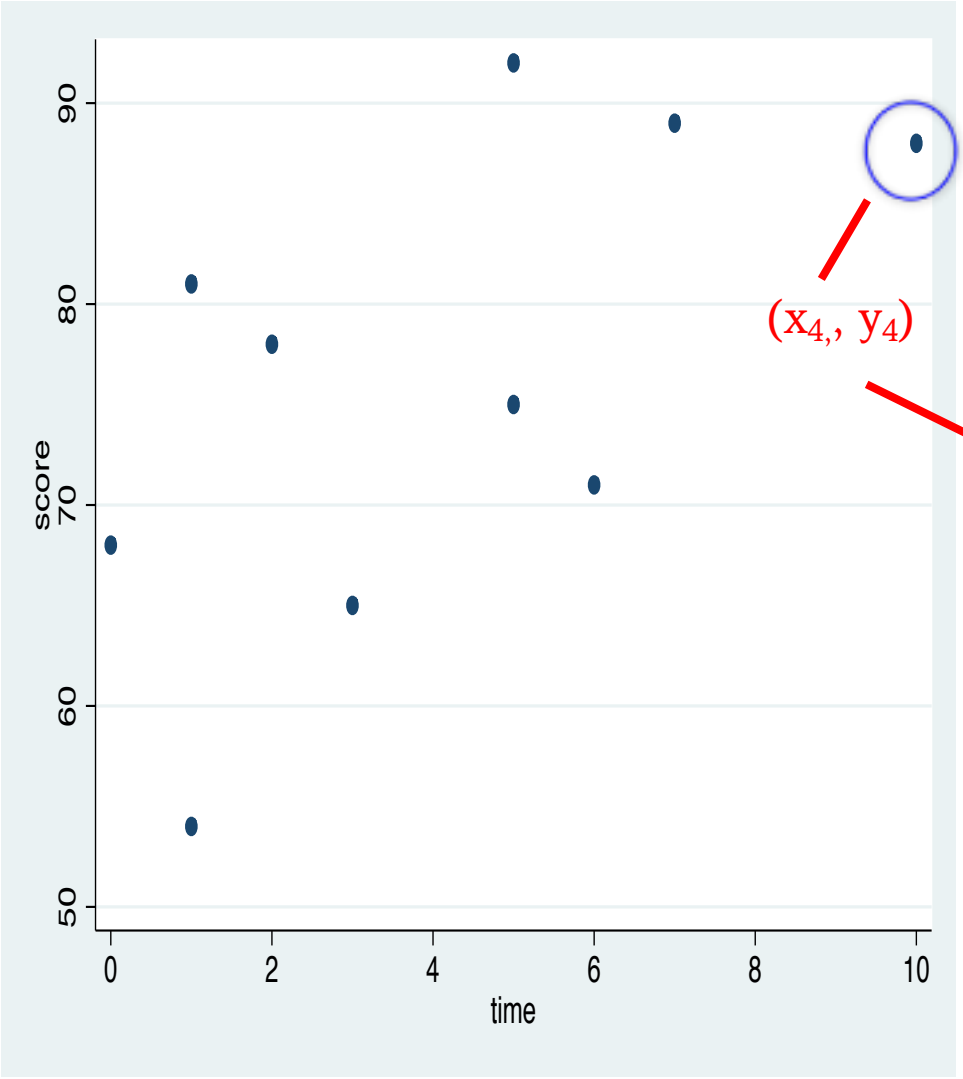
2. Select Scatter/Dot and drag the simple scatter into the Chart Preview Area



Drag the Independent/explanatory variable into the “x-axis?” box
-in our example drag “time” into the “x-axis?” box

Drag the Dependent/Response variable into the “y-axis?” box
-in our example drag “score” into the “y-axis?” box

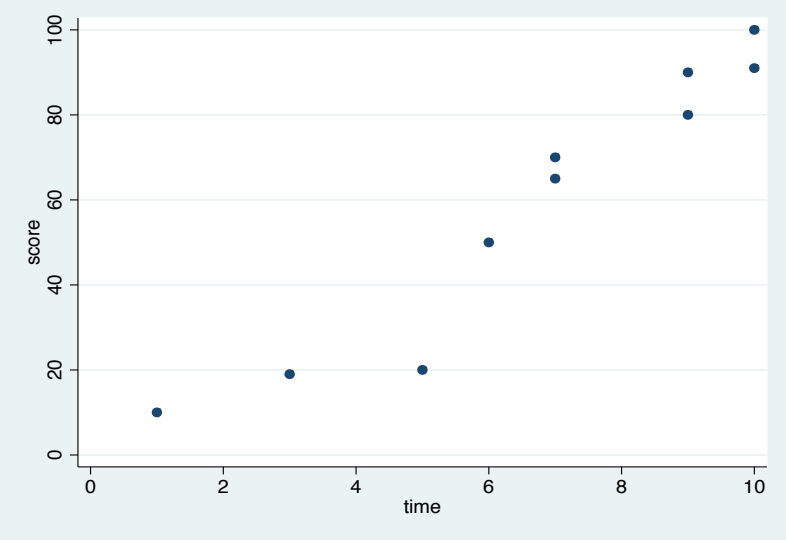


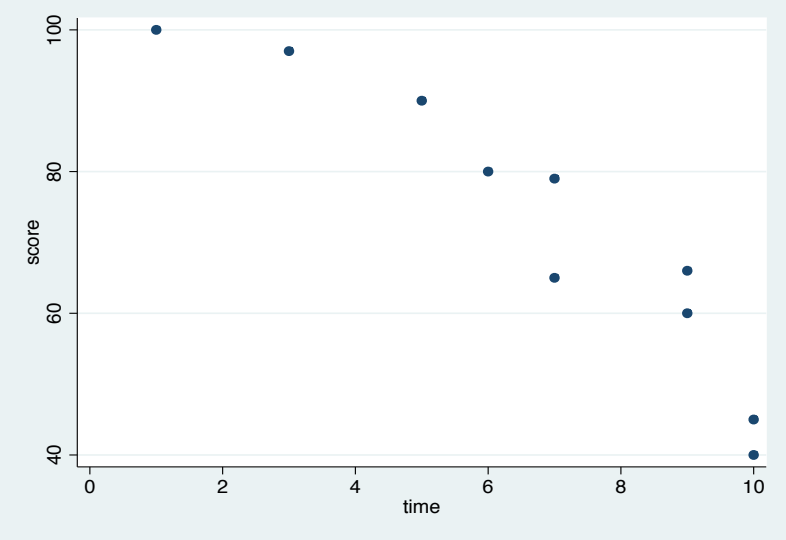


Student	Hours spent	Test Score
1	1	81
2	5	75
3	7	89
4	10	88
5	3	65
6	5	92
7	6	71
8	2	78
9	1	54
10	0	68

Positive/Negative Association

- Positive Association- increasing the independent variable increases the dependent variable
- Negative Association-increasing the independent variable decrease the dependent variable



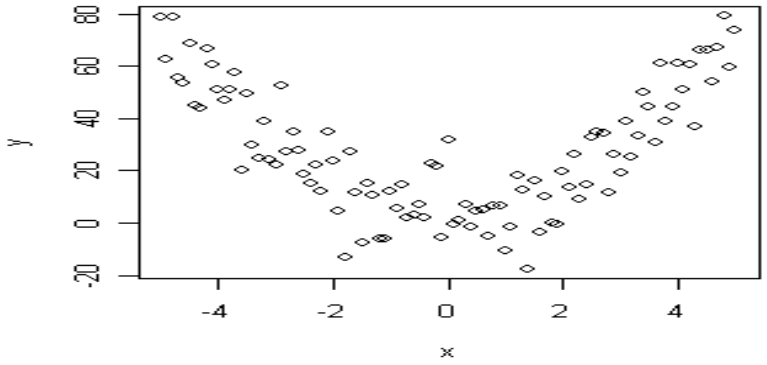


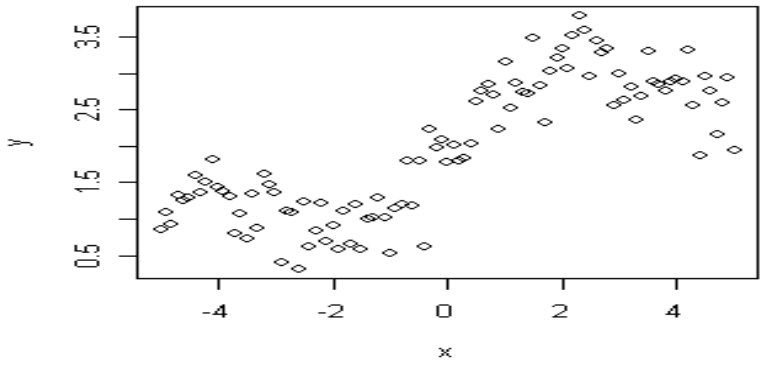
Correlation Coefficient

- Statistical correlation is measured using the correlation coefficient (r) which describes the degree of relationship between two variables
- r- takes on a value from -1 to 1
- $r<0$ -indicates a negative relationship between the variables
  - when one variable increases the other variable decreases, or when one variable decreases the other variable increases
- $r>0$ indicates a positive relationship between the variables
  - when one variable decreases the other also decreases, or when one variable increases the other also increases
- $r=0$ indicates that there is no relationship between the variables.

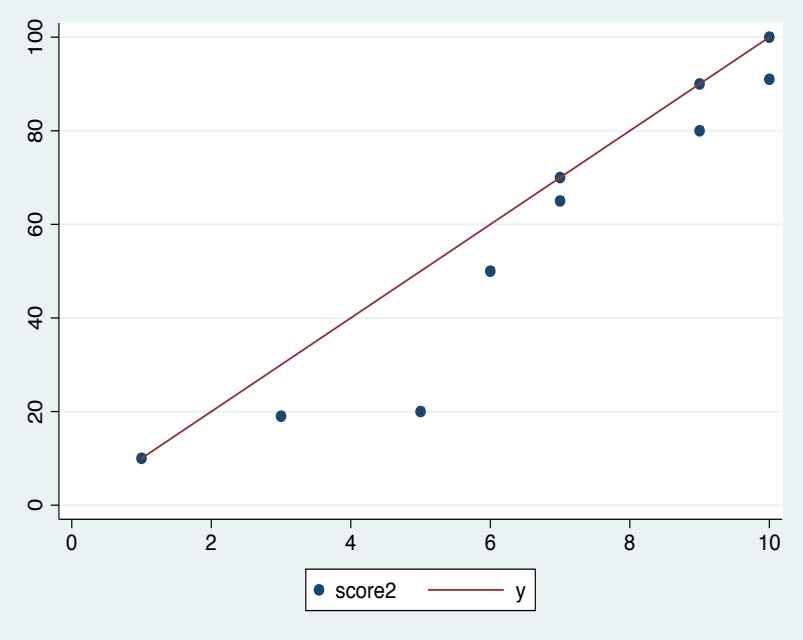
Correlation Coefficient cont`d

- *r* has no units
- *r* is not a proportion or percent-
  - *r*=0.8 is not twice as strong of a correlation as *r*=0.4
- *r* can be misleading – low *r* does not mean that there is no relation- always plot data to understand the association of the data

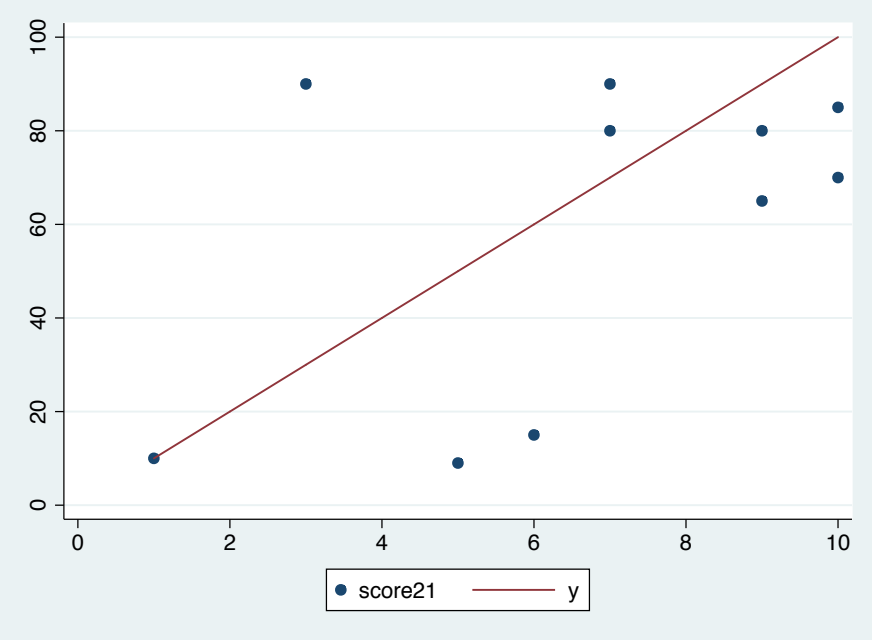




Positive correlation (positively linear)

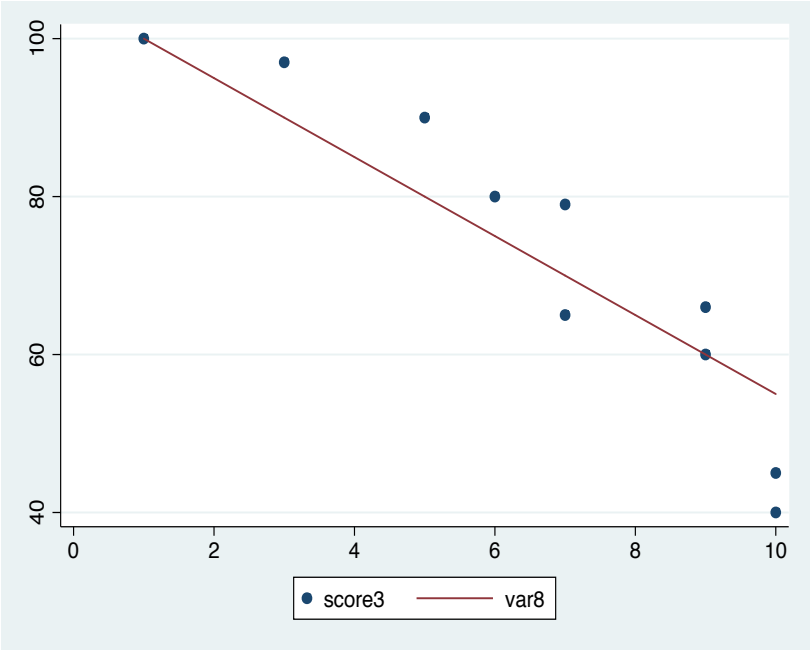


R=0.9617

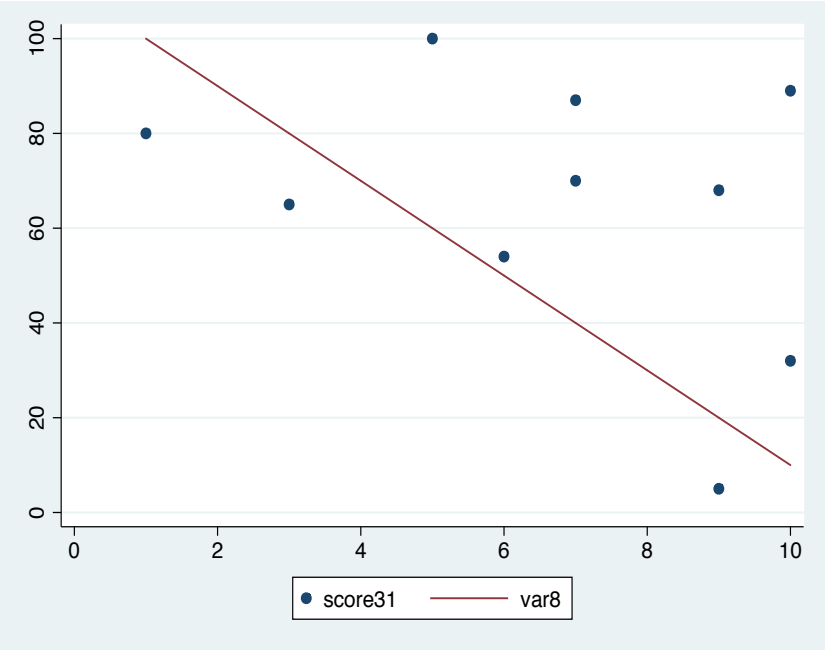


R=0.5179

Negative Correlation (Negatively Linear)

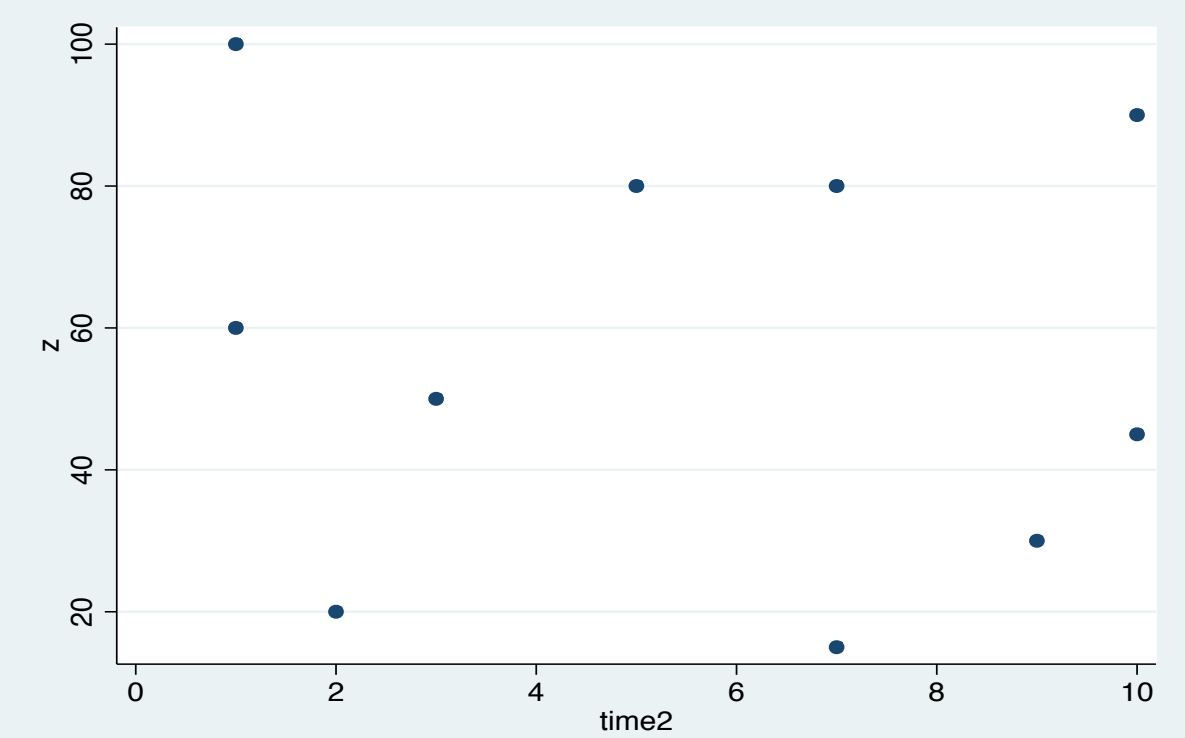


R=-0.9377



R=-0.3732

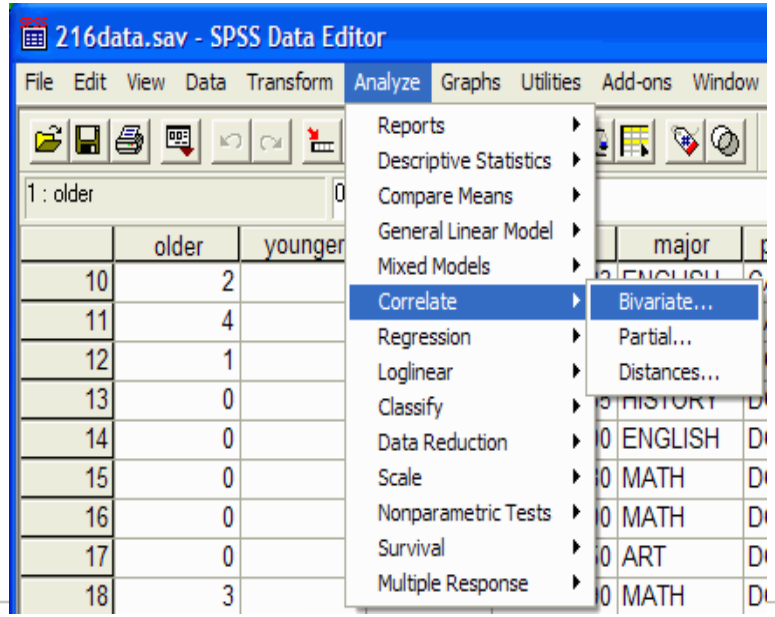
No Correlation

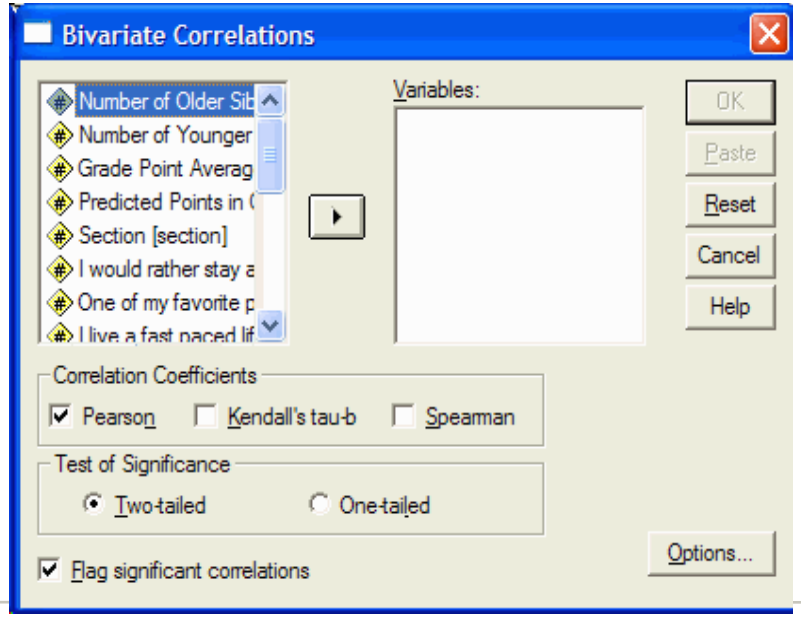


r= -0.1032

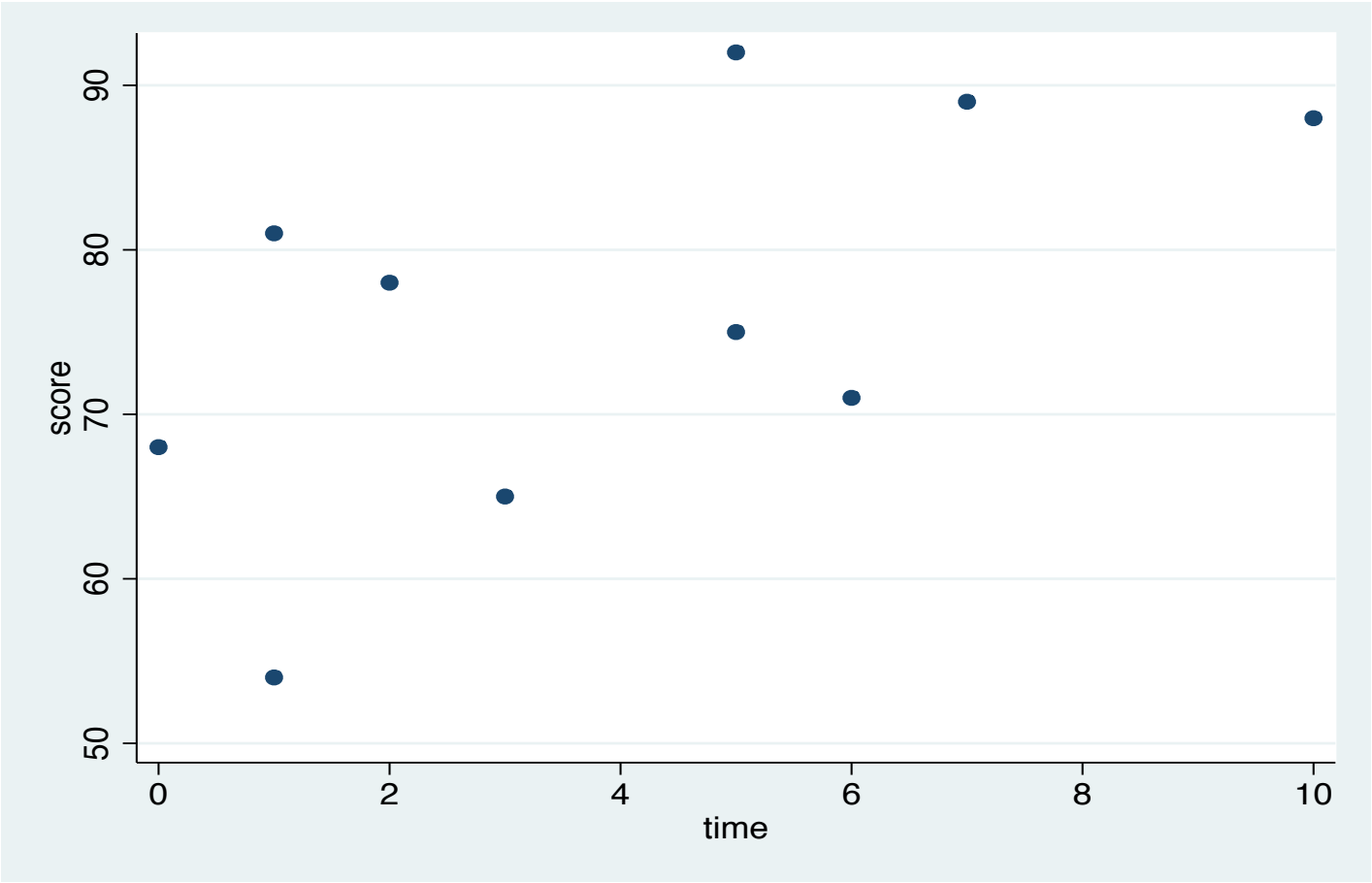
Correlation Coefficient in SPSS

- Analyze→ Correlate→ Bivariate
  - In the variables box select the two variables that are interest (time, score)
  - In the Correlation Coefficient box, Select Pearson





How is how the data correlated from the example?



r=0.6