Descriptives, Crosstabs, Correlation

Methodology of Conflict and Democracy Studies

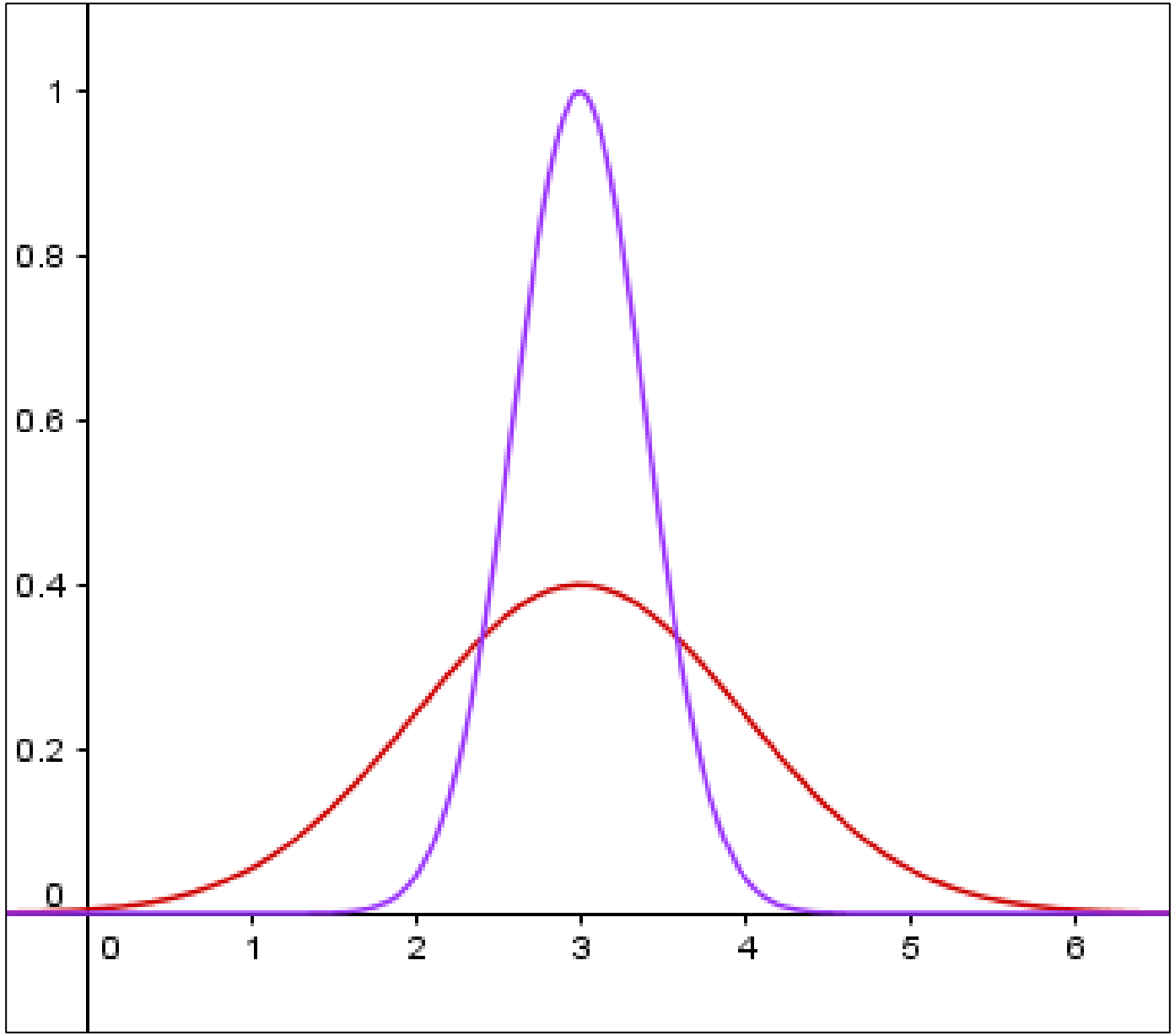
November 28

Aim of this lecture

- How to obtain basic information about your data
- Control of the assumptions
- Association of two variables:
  - Crosstabs (Contingency tables)
  - Correlation

Descriptive Statistics

- Basic measures to summarize the characteristics of your data
- Various types:
  - Central tendencies – mean, median
  - Dispersion – variance, minimum, maximum
- Not all descriptives are suitable for all types of variables
- We use them to describe and explore your data


1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6

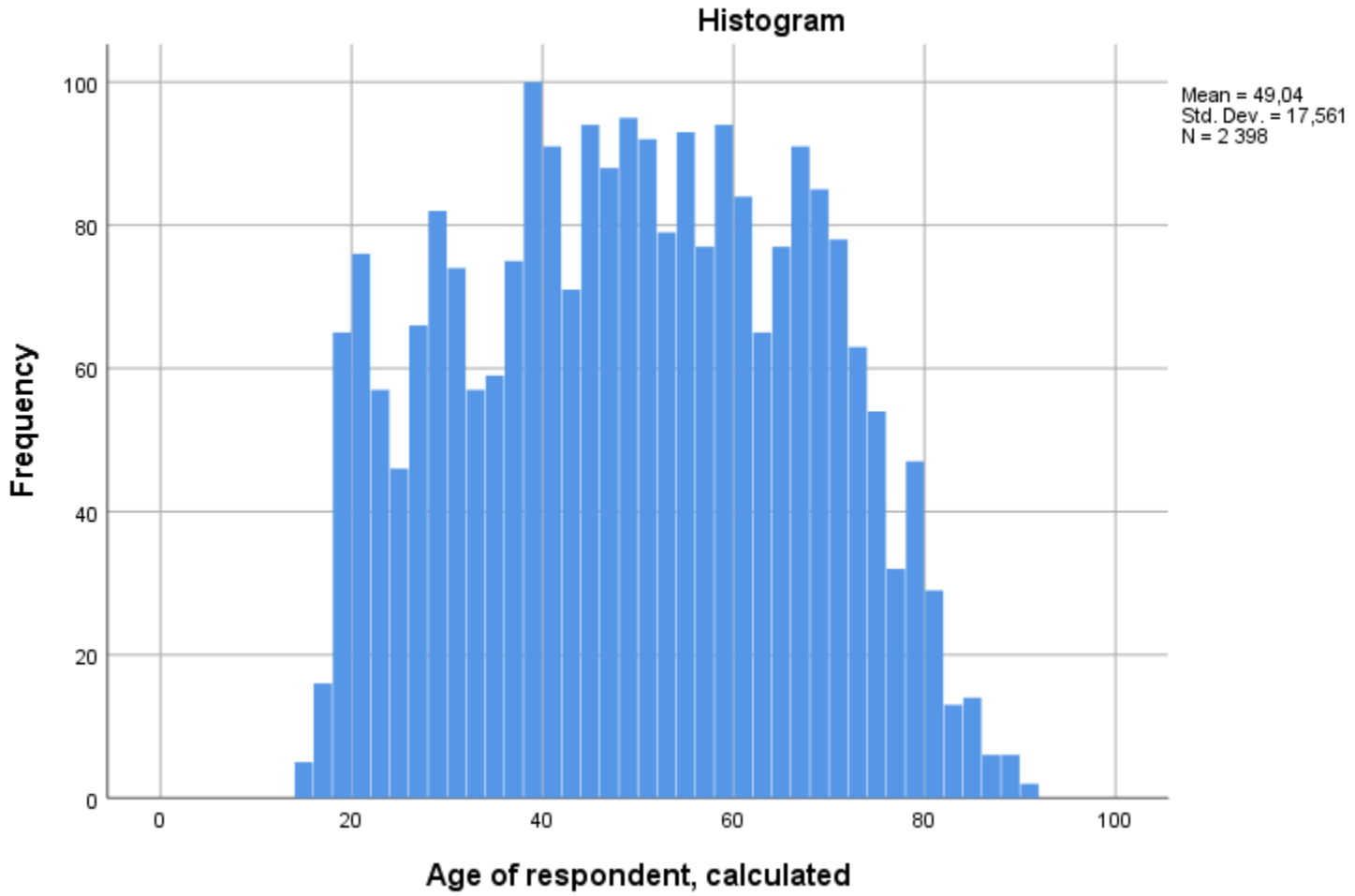
How to Obtain Descriptives in SPSS

- Analyze > Descriptive Statistics > Frequencies
- Move variables of interest to the right
- In ‘Statistics’ choose all measures you require

Statistics

Age of respondent, calculated

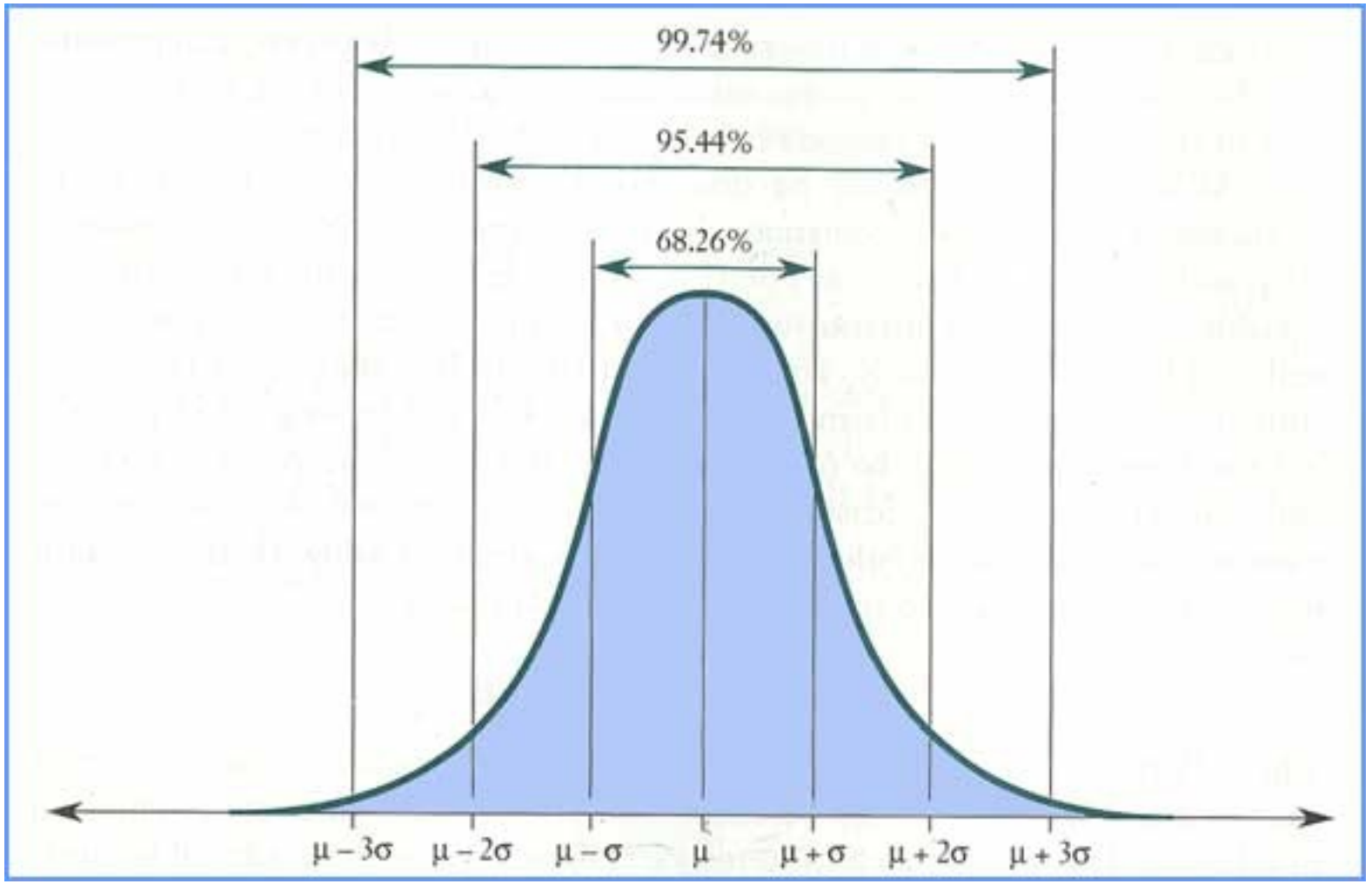
N	Valid	2398
	Missing	0
Mean		49,04
Median		49,00
Mode		50
Std. Deviation		17,561
Minimum		15
Maximum		90
Sum		117591



Assumptions of Data

- Not all data are suitable for all statistical tests
- Parametric and Non-parametric tests
- Parametric tests as a preference v. higher requests on data

Normal Distribution



How to Check the Distribution

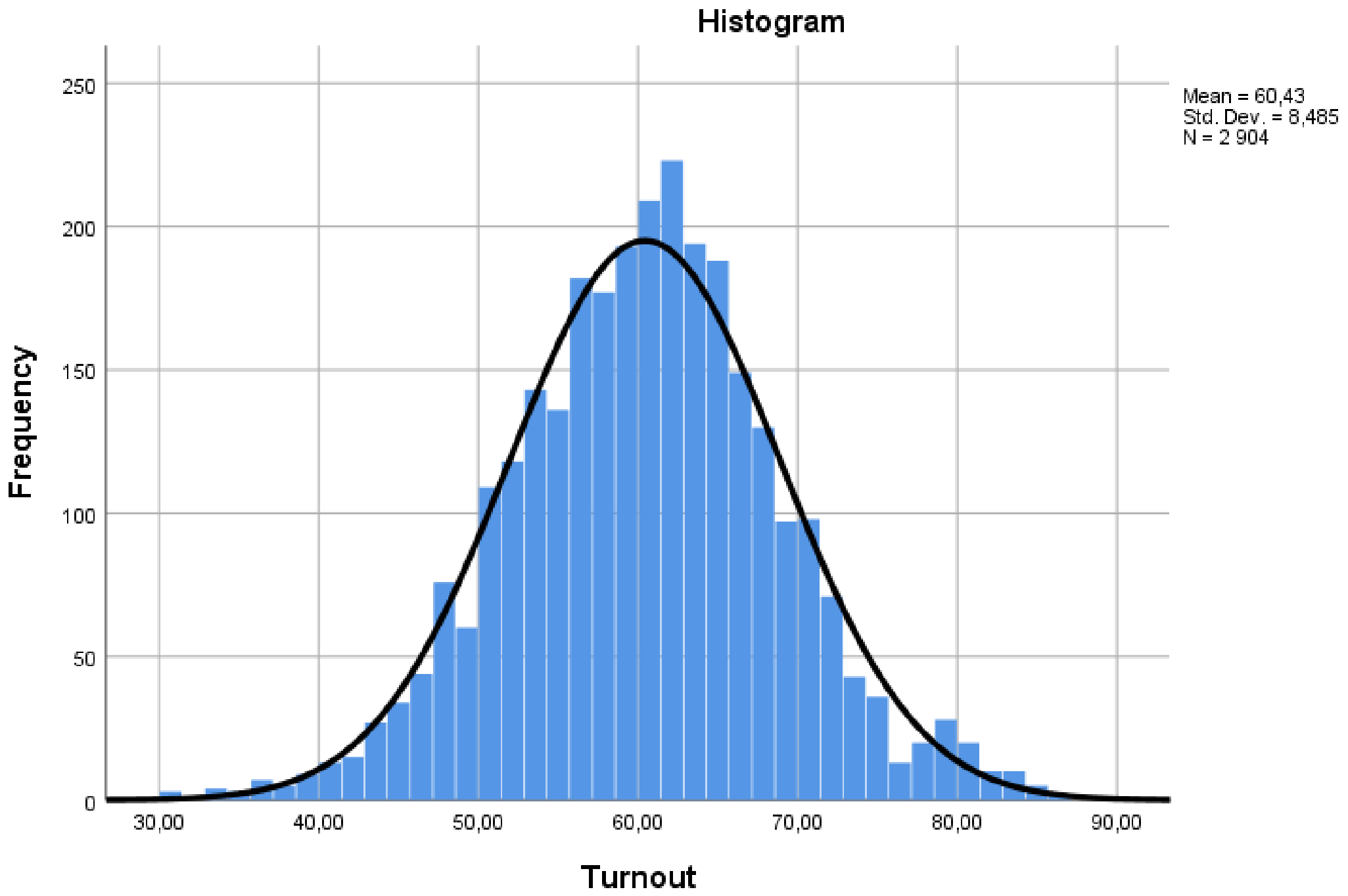
1) Visual control – Histogram

2) Statistical tests:

- Kolmogorov-Smirnov
- Shapiro-Wilk

1) Histogram

- Analyze > Descriptive Statistics > Frequencies
- In ‘Charts’ choose ‘Histogram’
- Select ‘Show normal curve on histogram’ to draw a line corresponding to normal distribution


Histogram
Mean = 60,43
Std. Dev. = 8,485
N = 2 904
Frequency
250
200
150
100
50
0
30,00
40,00
50,00
60,00
70,00
80,00
90,00
Turnout

2) Statistical Tests

- Kolmogorov-Smirnov (Shapiro-Wilk)
  - Both test the null hypothesis that your data are normally distributed

- Results:
  - Significant ($p <= 0.05$) – we reject the null hypothesis
  - Not significant ($p > 0.05$) – we keep the null hypothesis

- With large samples the tests tend to lead to significant results without meaningful reason → use histogram instead

How to *read* the significance in SPSS outputs

SPSS output	Significance
,900	10 %
,750	25 %
,500	50 %
,200	80 %
,100	90 %
,050	**95 %**
,010	**99 %**
,001	**99.9 %**
,000	**> 99.9 %**

= (1 – SPSS output) * 100

Example: (1 – 0.234) * 100 = 0.766 * 100 = 76.6 %

2) Statistical Tests

- Analyze > Descriptive Statistics > Explore
- Place variable of your interest into ‘Dependent List’
- In ‘Plots’ select ‘Normality plots with tests’

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Turnout	,018	2904	,039	,998	2904	,000

a. Lilliefors Significance Correction

Association of Two Variables

- Depends on types of variables

- Crosstabs:
  - Suitable for two categorical variables
  - Low amount of categories in your variables (but at least two per variable)

- Correlation:
  - Two scale variables, scale and ordinal, two ordinal variables
  - Specific case – scale and binary variable

Crosstabs

- Contingency tables
- Describe interaction of two categorical variables
- Age groups of people v. turnout in election (yes/no)
- Allow generalization to population

Crosstabs

- Analyze > Descriptive statistics > Crosstabs
- Select variables for Columns and Rows
- Features:
  - Cells – counts, percentages, residuals
  - Statistics – Chi-square, Cramer’s V
- Try not to fill your crosstab with too many features

Counts: Observed

Age * Voted in election Crosstabulation

Count

		Voted in election		
		No	Yes	Total
Age	18 - 35	271	248	519
	36 - 59	390	655	1045
	60 - 90	186	556	742
Total		847	1459	2306

Counts: Observed
Percentages: Row

Age * Voted in election Crosstabulation

			Voted in election		
			No	Yes	Total
Age	18 - 35	Count	271	248	519
		% within Age	52,2%	47,8%	100,0%
	36 - 59	Count	390	655	1045
		% within Age	37,3%	62,7%	100,0%
	60 - 90	Count	186	556	742
		% within Age	25,1%	74,9%	100,0%
Total		Count	847	1459	2306
		% within Age	36,7%	63,3%	100,0%

Counts: Observed
Percentages: Column

Age * Voted in election Crosstabulation

			Voted in election		
			No	Yes	Total
Age	18 - 35	Count	271	248	519
		% within Voted in election	32,0%	17,0%	22,5%
	36 - 59	Count	390	655	1045
		% within Voted in election	46,0%	44,9%	45,3%
	60 - 90	Count	186	556	742
		% within Voted in election	22,0%	38,1%	32,2%
Total		Count	847	1459	2306
		% within Voted in election	100,0%	100,0%	100,0%

Counts: Observed
Percentages: Row

- Younger people do not vote to the same extent than older people
- But can we apply this to the whole population?

Age * Voted in election Crosstabulation

			Voted in election		
			No	Yes	Total
Age	18 - 35	Count	271	248	519
		% within Age	52,2%	47,8%	100,0%
	36 - 59	Count	390	655	1045
		% within Age	37,3%	62,7%	100,0%
	60 - 90	Count	186	556	742
		% within Age	25,1%	74,9%	100,0%
Total		Count	847	1459	2306
		% within Age	36,7%	63,3%	100,0%

Chi-square, Cramer's V

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	97,142[a]	2	,000
Likelihood Ratio	97,604	2	,000
Linear-by-Linear Association	96,677	1	,000
N of Valid Cases	2306		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 190,63.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,205	,000
	Cramer's V	,205	,000
N of Valid Cases		2306	

- There is a relationship between age and turnout, and it applies to the population
- But is it okay to end the analysis at this point? Can we find out **more**?

Counts: Observed + Expected

Age * Voted in election Crosstabulation

			Voted in election		
			No	Yes	Total
Age	18 - 35	Count	271	248	519
		Expected Count	190,6	328,4	519,0
	36 - 59	Count	390	655	1045
		Expected Count	383,8	661,2	1045,0
	60 - 90	Count	186	556	742
		Expected Count	272,5	469,5	742,0
Total		Count	847	1459	2306
		Expected Count	847,0	1459,0	2306,0

Counts: Observed + Expected
Residuals: Unstandardized

Age * Voted in election Crosstabulation

			Voted in election		
			No	Yes	Total
Age	18 - 35	Count	271	248	519
		Expected Count	190,6	328,4	519,0
		Residual	80,4	-80,4	
	36 - 59	Count	390	655	1045
		Expected Count	383,8	661,2	1045,0
		Residual	6,2	-6,2	
	60 - 90	Count	186	556	742
		Expected Count	272,5	469,5	742,0
		Residual	-86,5	86,5	
Total		Count	847	1459	2306
		Expected Count	847,0	1459,0	2306,0

Counts: Observed + Expected
Residuals: Adjusted standardized

Age * Voted in election Crosstabulation

			Voted in election		
			No	Yes	Total
Age	18 - 35	Count	271	248	519
		Expected Count	190,6	328,4	519,0
		Adjusted Residual	8,3	-8,3	
	36 - 59	Count	390	655	1045
		Expected Count	383,8	661,2	1045,0
		Adjusted Residual	,5	-,5	
	60 - 90	Count	186	556	742
		Expected Count	272,5	469,5	742,0
		Adjusted Residual	-8,0	8,0	
Total		Count	847	1459	2306
		Expected Count	847,0	1459,0	2306,0

Counts: Observed + Expected
Residuals: Adjusted standardized
Chi-square, Cramer's V

Age * Voted in election Crosstabulation

			Voted in election		
			No	Yes	Total
Age	18 - 35	Count	271	248	519
		Expected Count	190,6	328,4	519,0
		Adjusted Residual	8,3	-8,3	
	36 - 59	Count	390	655	1045
		Expected Count	383,8	661,2	1045,0
		Adjusted Residual	,5	-,5	
	60 - 90	Count	186	556	742
		Expected Count	272,5	469,5	742,0
		Adjusted Residual	-8,0	8,0	
Total		Count	847	1459	2306
		Expected Count	847,0	1459,0	2306,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	97,142[a]	2	,000
Likelihood Ratio	97,604	2	,000
Linear-by-Linear Association	96,677	1	,000
N of Valid Cases	2306		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 190,63.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,205	,000
	Cramer's V	,205	,000
N of Valid Cases		2306	

Why Not Make It Too Complicated?

Age * Voted in election Crosstabulation

			Voted in election		
			No	Yes	Total
Age	18 - 35	Count	271	248	519
		Expected Count	190,6	328,4	519,0
		% within Age	52,2%	47,8%	100,0%
		% within Voted in election	32,0%	17,0%	22,5%
		% of Total	11,8%	10,8%	22,5%
		Residual	80,4	-80,4	
		Adjusted Residual	8,3	-8,3	
	36 - 59	Count	390	655	1045
		Expected Count	383,8	661,2	1045,0
		% within Age	37,3%	62,7%	100,0%
		% within Voted in election	46,0%	44,9%	45,3%
		% of Total	16,9%	28,4%	45,3%
		Residual	6,2	-6,2	
		Adjusted Residual	,5	-,5	
	60 - 90	Count	186	556	742
		Expected Count	272,5	469,5	742,0
		% within Age	25,1%	74,9%	100,0%
		% within Voted in election	22,0%	38,1%	32,2%
		% of Total	8,1%	24,1%	32,2%
		Residual	-86,5	86,5	
		Adjusted Residual	-8,0	8,0	
Total		Count	847	1459	2306
		Expected Count	847,0	1459,0	2306,0
		% within Age	36,7%	63,3%	100,0%
		% within Voted in election	100,0%	100,0%	100,0%
		% of Total	36,7%	63,3%	100,0%

Correlation

- Association between two variables (for other cases than crosstabs)
- Examples: two scale variables, scale and ordinal, two ordinal variables
- Three coefficients:
  - Pearson
  - Spearman
  - Kendall

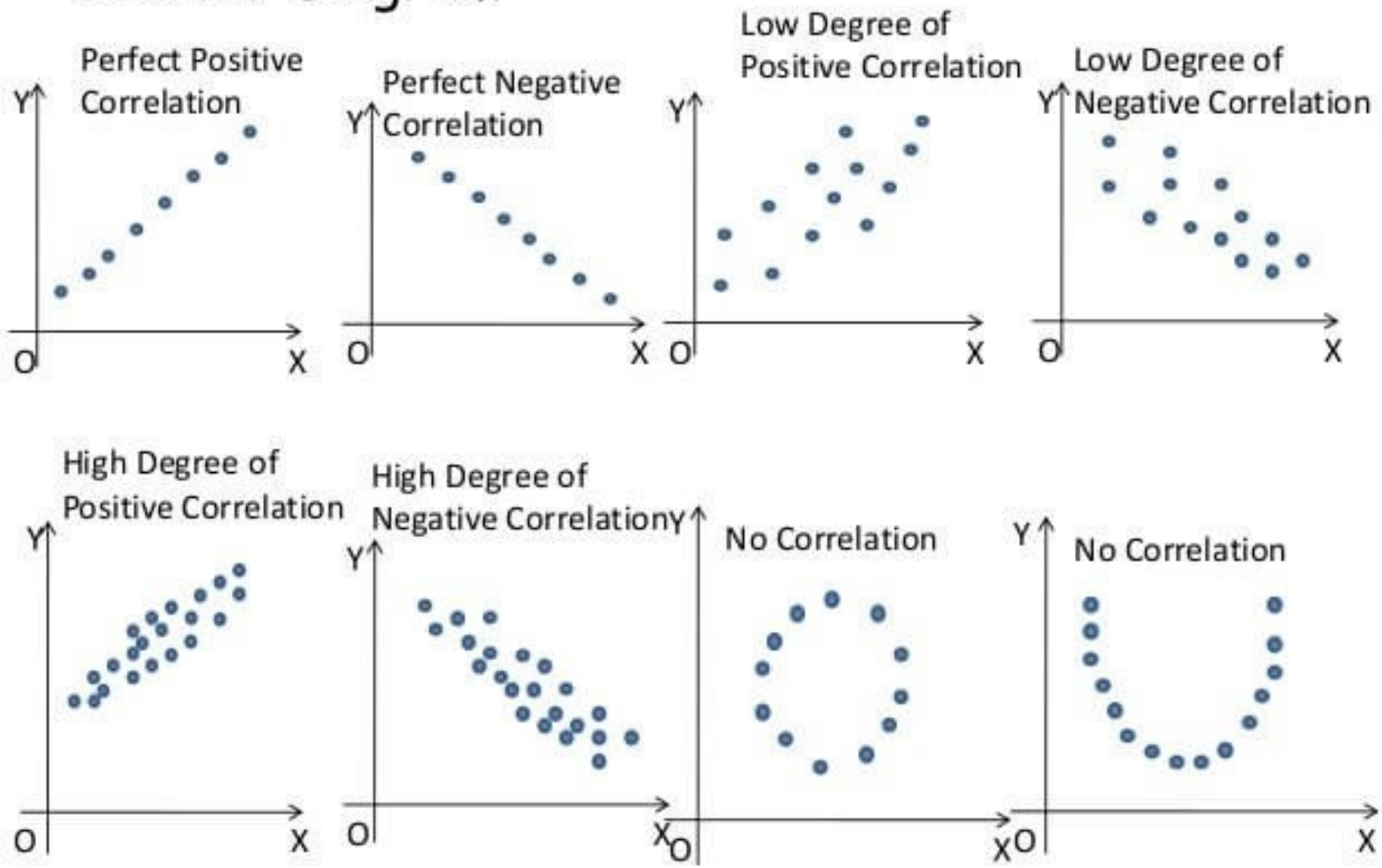
Correlation

- Results vary on a scale between -1 and 1

- Interpretation:
  - Zero means no association between the variables
  - Rising distance from zero shows rising association (regardless the direction – negative or positive)
  - -1: perfect negative association
  - 1: perfect positive association

- Beware of false absence of association
- Always good to visualize data before calculating correlations

Scatter Diagram



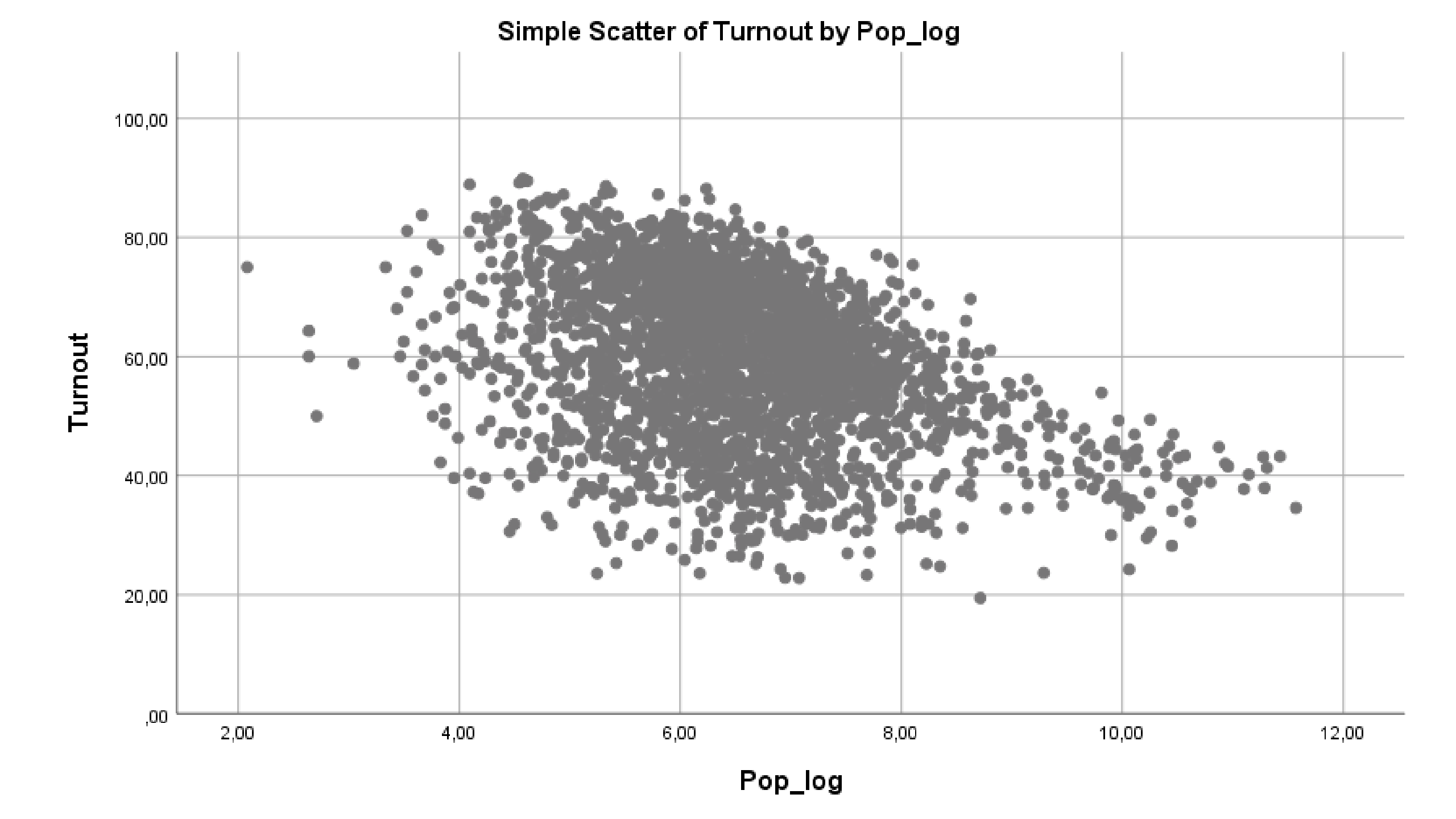
Pearson's Correlation Coefficient

- Parametric operation
- Requirements:
  - Scale data (exemption – scale and binary)
  - If we aim to apply the findings to the population, we need normally distributed data (or a large sample)
- Sensitive to outliers

Pearson’s Correlation Coefficient

- Visualize the data
  - Graphs > Chart Builder
  - Select Scatter/Dot a variables of your interest

- Correlation
  - Analyze > Correlate > Bivariate
  - Select variables and the proper coefficient (PCC is set by default)
  - For significance select ‘Flag significant correlations’


Simple Scatter of Turnout by Pop_log
Turnout
100,00
80,00
60,00
40,00
20,00
,00
2,00
4,00
6,00
8,00
10,00
12,00
Pop_log

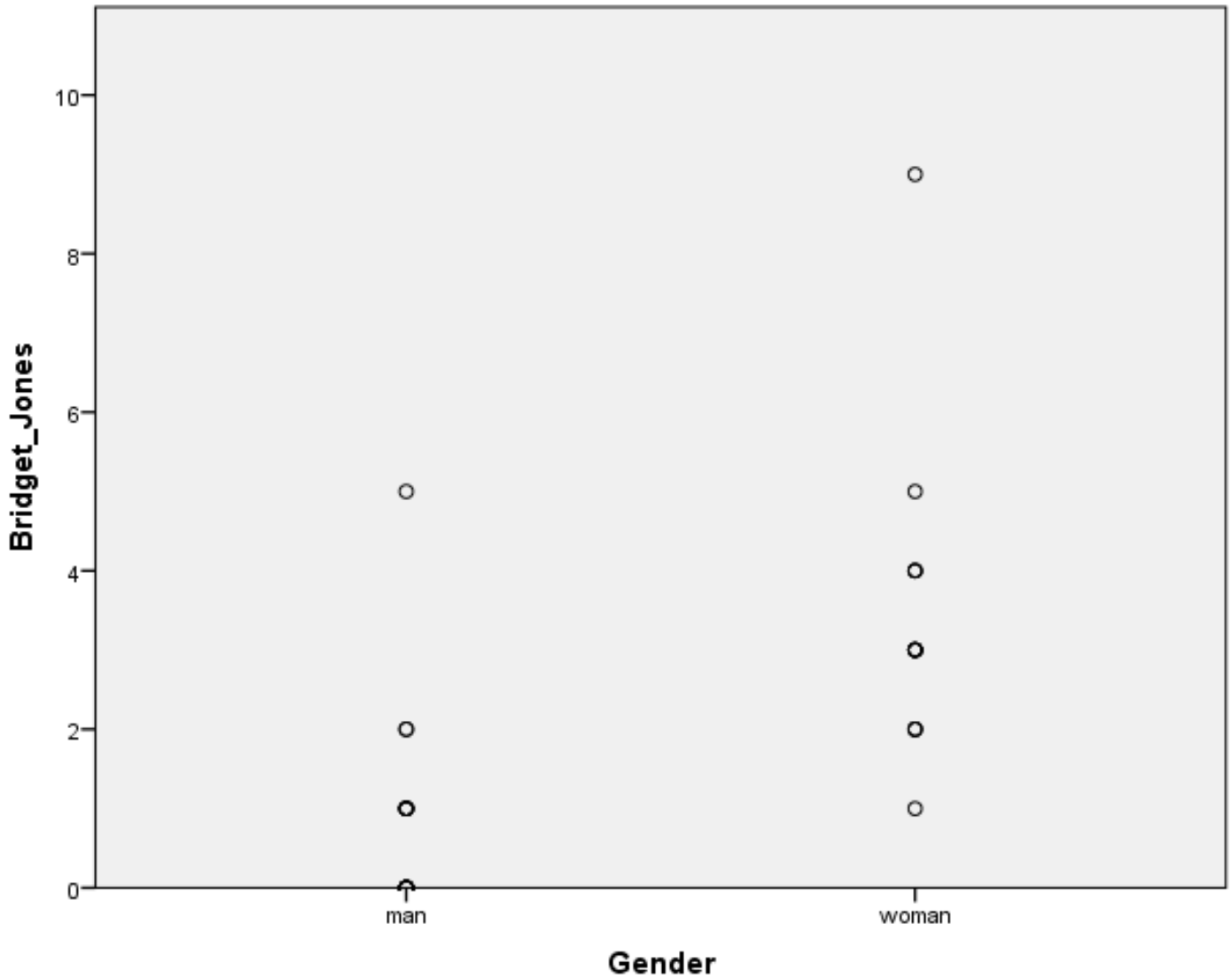
Correlations

		Pop_log	Turnout
Pop_log	Pearson Correlation	1	-,366**
	Sig. (2-tailed)		,000
	N	2926	2919
Turnout	Pearson Correlation	-,366**	1
	Sig. (2-tailed)	,000	
	N	2919	2919

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

Pearson's Correlation Coefficient

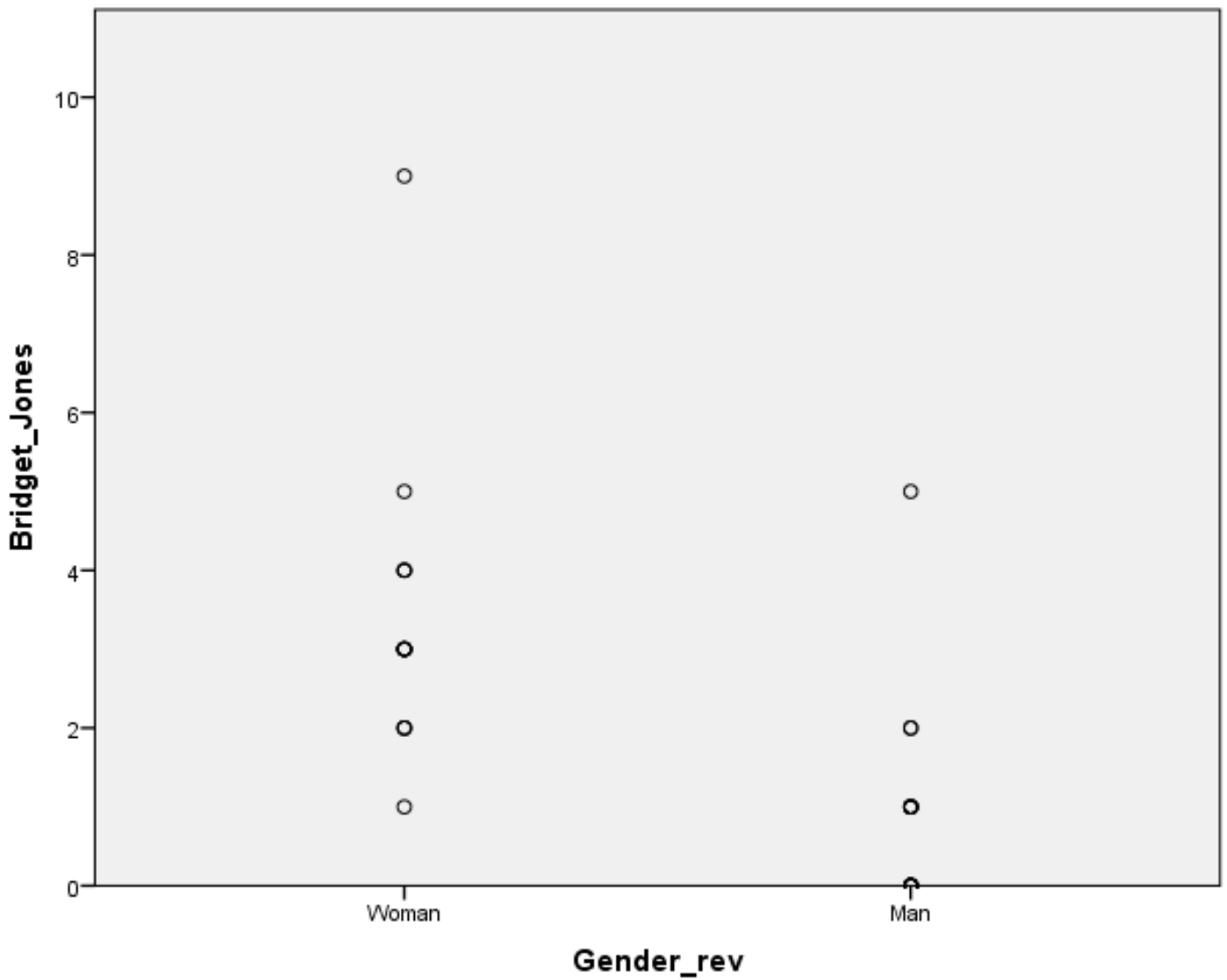
- Scale variable and binary variable
- Works the same as for two scale variables
- Beware of coding of the binary variable (be sure what values the codes represent)



Correlations

		Bridget_Jones	Gender
Bridget_Jones	Pearson Correlation	1	,677**
	Sig. (2-tailed)		,000
	N	37	37
Gender	Pearson Correlation	,677**	1
	Sig. (2-tailed)	,000	
	N	37	37

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



Correlations

		Bridget_Jones	Gender_rev
Bridget_Jones	Pearson Correlation	1	-,677**
	Sig. (2-tailed)		,000
	N	37	37
Gender_rev	Pearson Correlation	-,677**	1
	Sig. (2-tailed)	,000	
	N	37	37

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

Non-Parametric Correlation

- Spearman’s Rho and Kendall’s Tau
  - Correlation for other cases than two scale variables (or scale and binary)
  - Same interpretation as in Pearson’s CC
  - Preference of Kendall’s Tau if variables contain less categories and for smaller samples

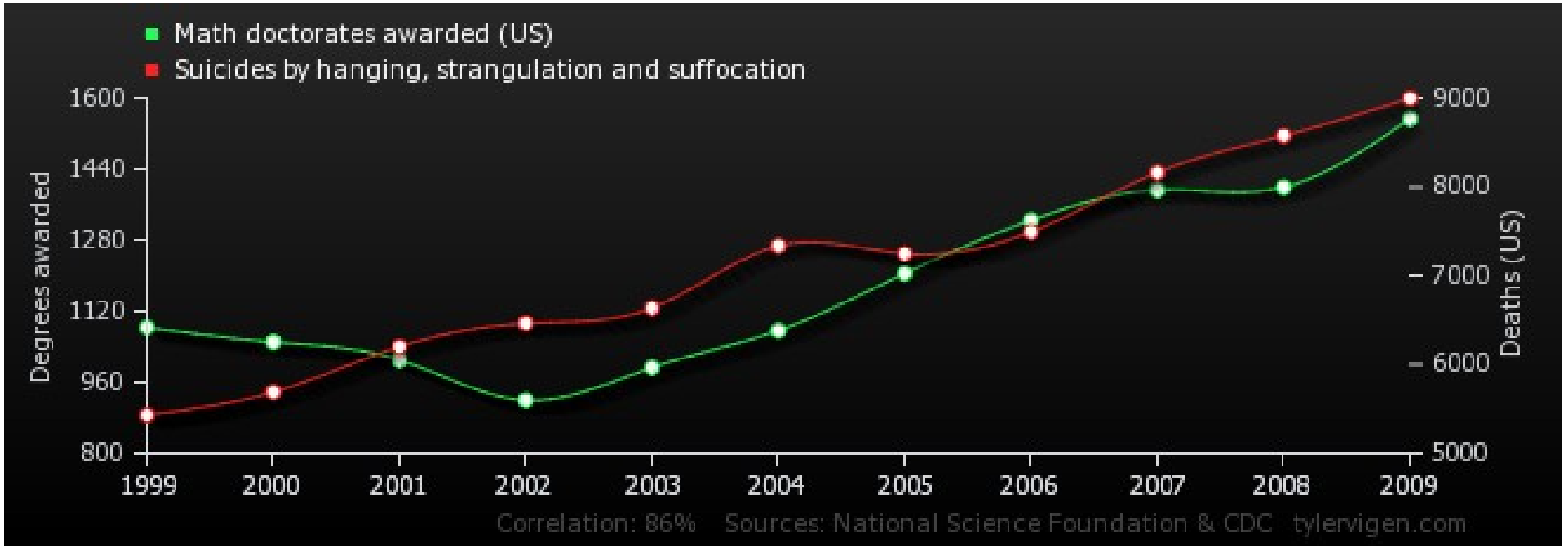
- Analyze > Correlate > Bivariate
  - Select variables and Spearman/Kendall
  - For significance select ‘Flag significant correlations’

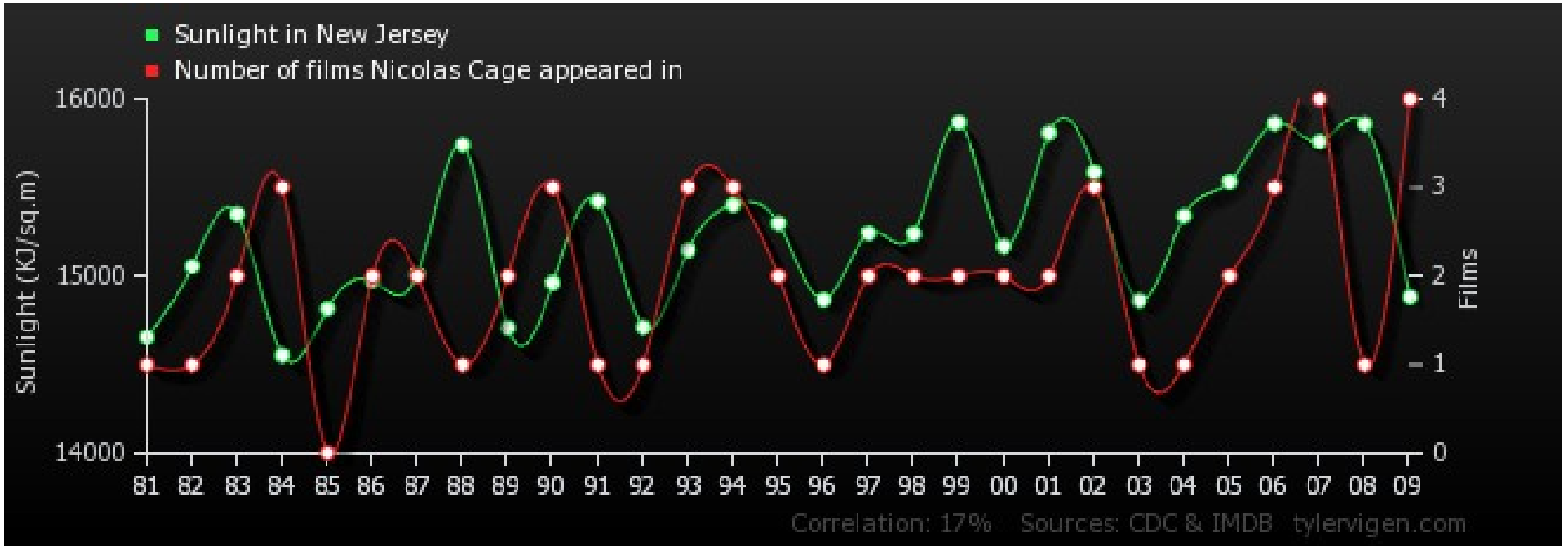
Interpretation

- Correlation does not imply causality
  - No control of other variables
  - No independent and dependent variable

- You cannot tell that one variable affects the other even in cases when such relationship seems to be meaningful and logical

- Keep the interpretation of effects of IVs on DV for the regression analysis


Math doctorates awarded (US)
Suicides by hanging, strangulation and suffocation
Degrees awarded
Deaths (US)
1600
1440
1280
1120
960
800
9000
8000
7000
6000
5000
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Correlation: 86% Sources: National Science Foundation & CDC tylervigen.com


Sunlight in New Jersey
Number of films Nicolas Cage appeared in
Sunlight (KJ/sq.m)
Films
16000
15000
14000
4
3
2
1
0
81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09
Correlation: 17% Sources: CDC & IMDB tylervigen.com