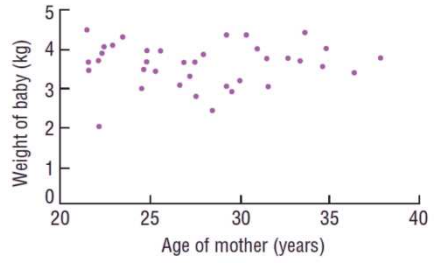
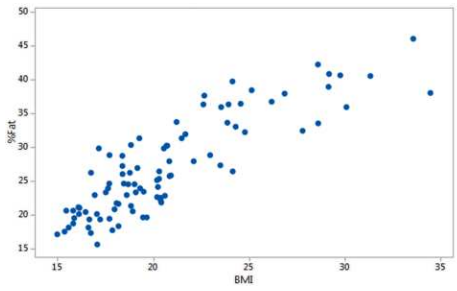
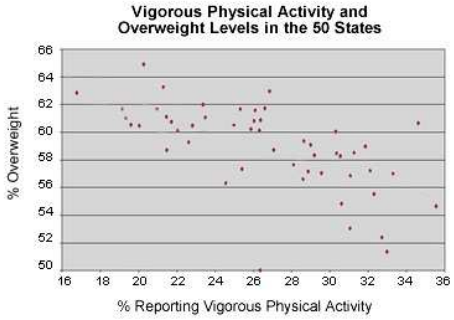
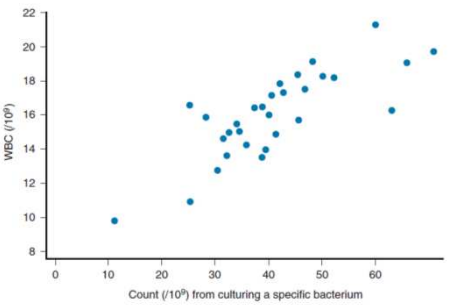
Examining associations (Correlation)

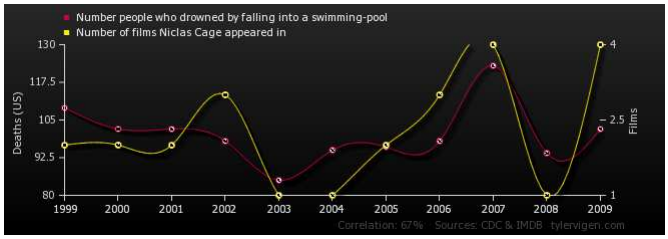
Examples of associations

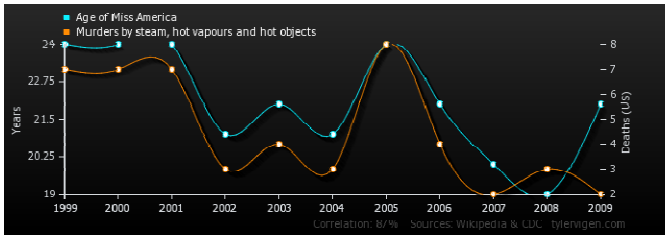










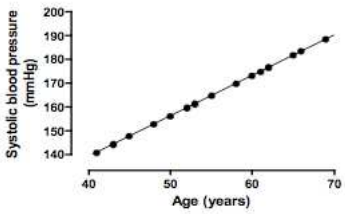


http://tylervigen.com/old-version.html

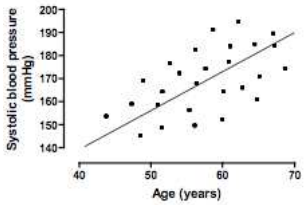
Correlation coefficient

- Measuring the strength of association between 2 continuous variables (i.e. 'taking measurements on people or things')
- Not interested in predicting one variable from the other
  - To assess whether two variables are linearly associated - use Pearson correlation coefficient (r)
  - To assess whether two variables are monotonically associated - use Spearman rank correlation coefficient (ρ, pronounced as rho)
- r and ρ take values between -1 and 1
- r or ρ=-1 or +1 ➡ Perfect association (negative/positive)
- r or ρ=0 ➡ No association

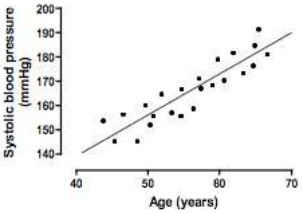
To evaluate the strength of an association



Very strong association



Lots of scatter



Less scatter, stronger association

Correlation=1

Perfect (positive) association

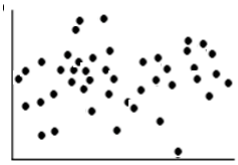
r>0 ➡ one variable increases, so does the other

Correlation=-1

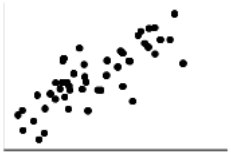
Perfect (negative) association

r<0 ➡ one variable increases, the other decreases

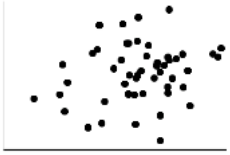
Correlation=0



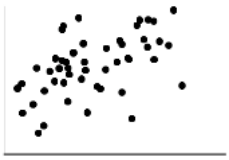
Correlation=0.8



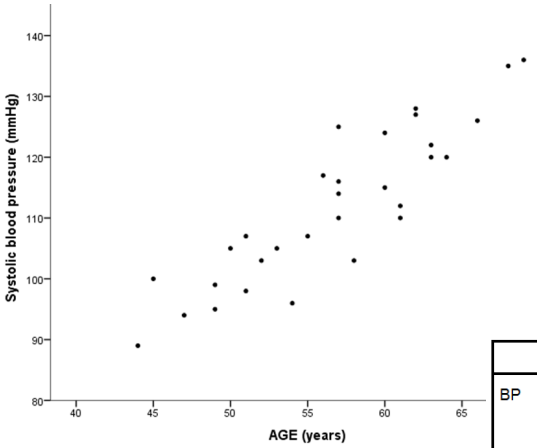
Correlation=0.2



Correlation=0.5



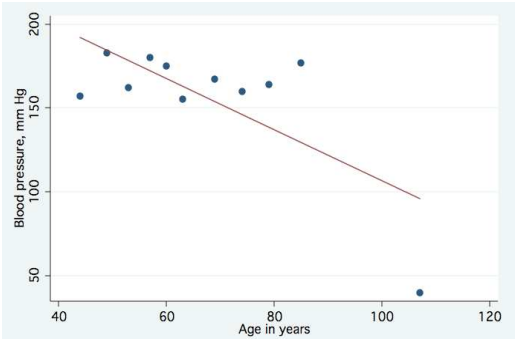
SPSS output



Correlations

		BP	AGE
BP	Pearson Correlation	1	.891**
	Sig. (2-tailed)		.000
	N	30	30
AGE	Pearson Correlation	.891**	1
	Sig. (2-tailed)	.000	
	N	30	30

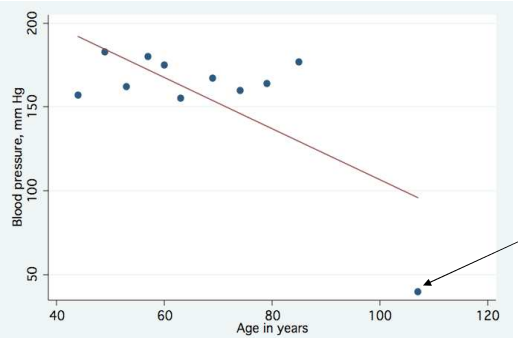
**. Correlation is significant at the 0.01 level (2-tailed).



Pearson correlation coefficient (r) = -0.70 (p-value=0.017)

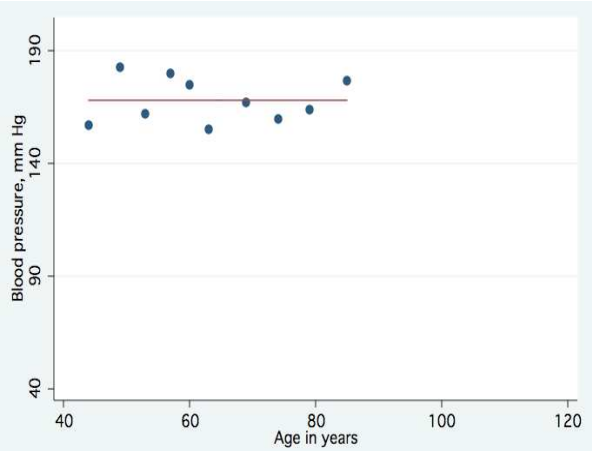
What is your conclusion now?

Outliers and correlations



- Pearson correlation coefficient (r) = -0.70 (p-value=0.017)
- Spearmans rank correlation coefficient (ρ) = -0.25 (p-value =0.450)

Pearson coefficient is very sensitive to outliers. Spearman rank coefficient is more appropriate in this example



- Pearson correlation coefficient (r) = 0.005 (p-value= 0.989)
- Spearmans rank correlation coefficient (ρ) = -0.006 (p-value =0.987)

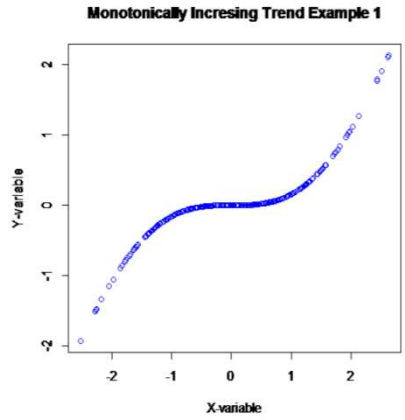
How do I deal with outliers?

- Exclude them?
- Include them?

Before deciding:

- Data checks
- Analysis with and without outlier

- If data point is correct, check if there is anything unusual *!*
- There might be something scientifically important which may
- explain why it is an outlier



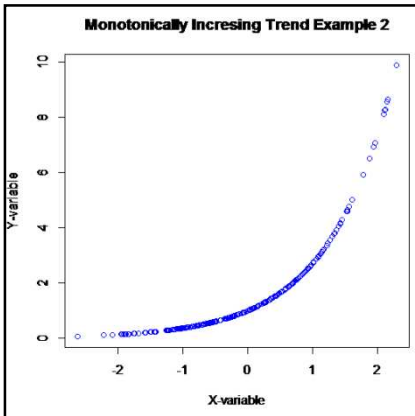
Pearsons = 0.85

Spearmans = 1

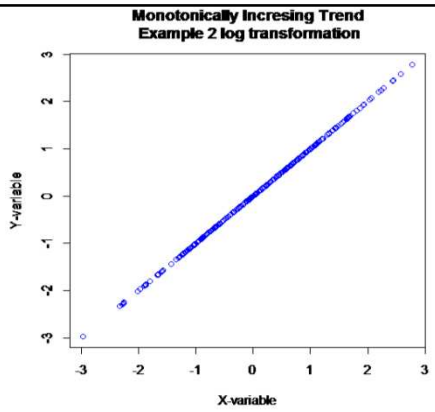
When non-linear, Spearman is better

Assumptions for Pearson's

- Each variable does not have to be Normally distributed
- But if you don't see an approximate straight line, taking a transformation(e.g. log) might fix this (i.e. make it nearly straight)



Pearsons = 0.86
Spearmans = 1



Pearsons = 1
Spearmans = 1

Correlations and p-values

- Statistical packages will provide a p-value for the correlation coefficient
- Need to interpret it carefully, since a small p-value (eg 0.01) could be obtained for a weak correlation (eg 0.09) that came from a large number of observations (eg 5000)
- What matters is the size of the correlation
- P-values might be useful when the study size is not large (say <100)
- A 95% CI is more useful for any correlation, since it will be wide for small studies, and narrow for large studies

Confidence intervals

- 30 measurements with blood pressure and age:
  - Pearsons: 0.89, 95% CI: 0.78 to 0.95
  - Spearmans: 0.87, 95% CI: 0.75 to 0.94
- 10 measurements (without outlier)
  - Pearsons: 0.005, 95% CI: -0.63 to +0.63
  - Spearmans: -0.006, 95% CI: -0.63 to +0.63
- 11 measurements (with outlier)
  - Pearsons: -0.7, 95% CI: -0.92 to -0.17
  - Spearmans: -0.25, 95% CI: -0.74 to +0.412