Nemours Biomedical Research Statistics Course

Relationship Between Variables

Li Xie

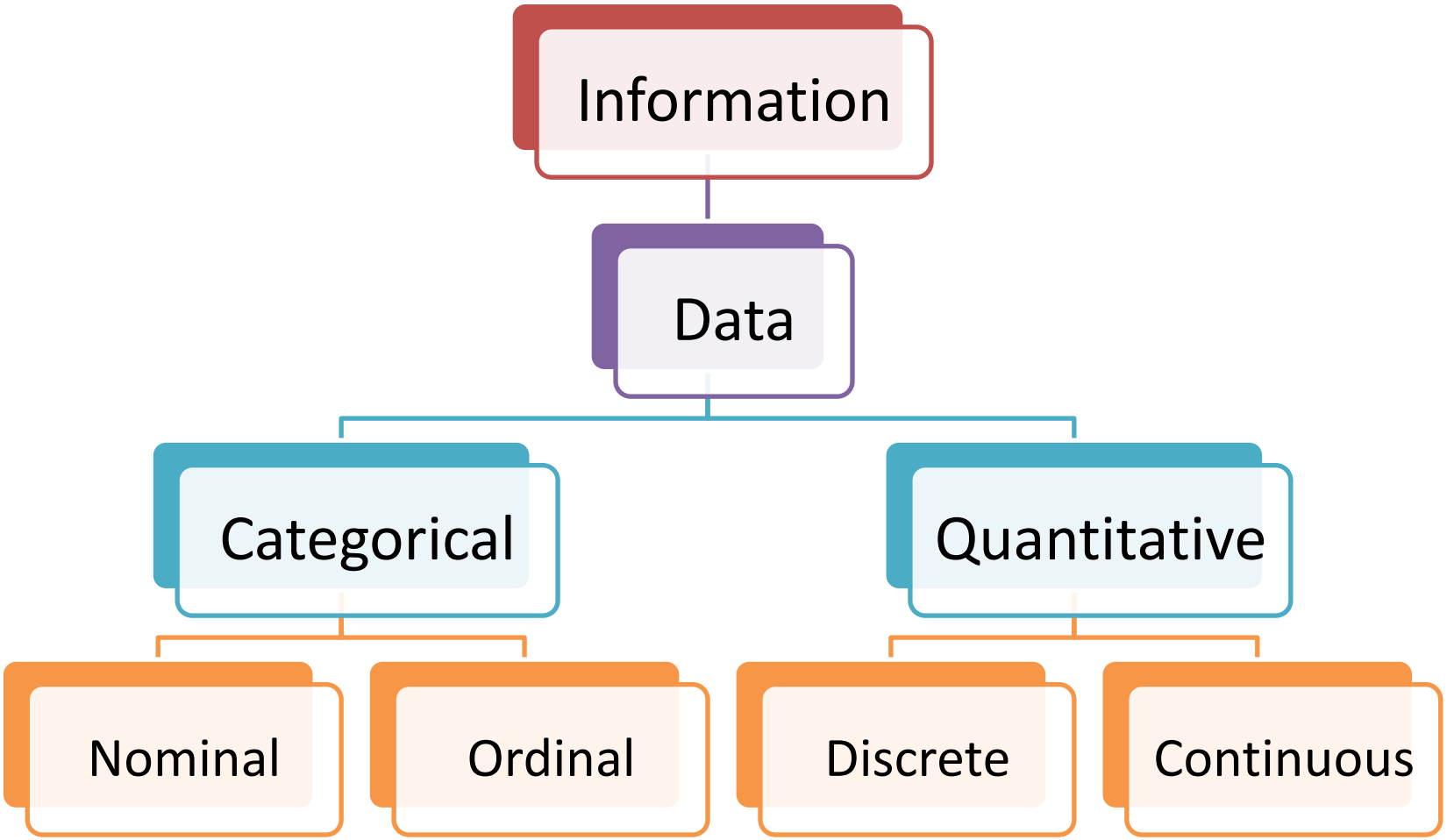
Nemours Biostatistics Core

September 16, 2014

Outline

- Recap: Variable Type
- Recap: Descriptive Statistics
- Correlation Coefficients
  - Pearson's Correlation Coefficient
  - *Spearman's Correlation Coefficient
  - *Kendall's τ
- Correlation vs Slope – Some Disambiguation

Variable Types



Descriptive Statistics

Descriptive statistics are numbers that are used to summarize and describe data.

- Categorical variable: proportion
- Quantitative variable: mean, median, variance, standard deviation

$$\text{Median} = \frac{1}{2}(n+1)\text{th value, where } n \text{ is the number of data values in the sample}$$

Sample Mean

$$\bar{x} = \frac{\sum x}{n}$$

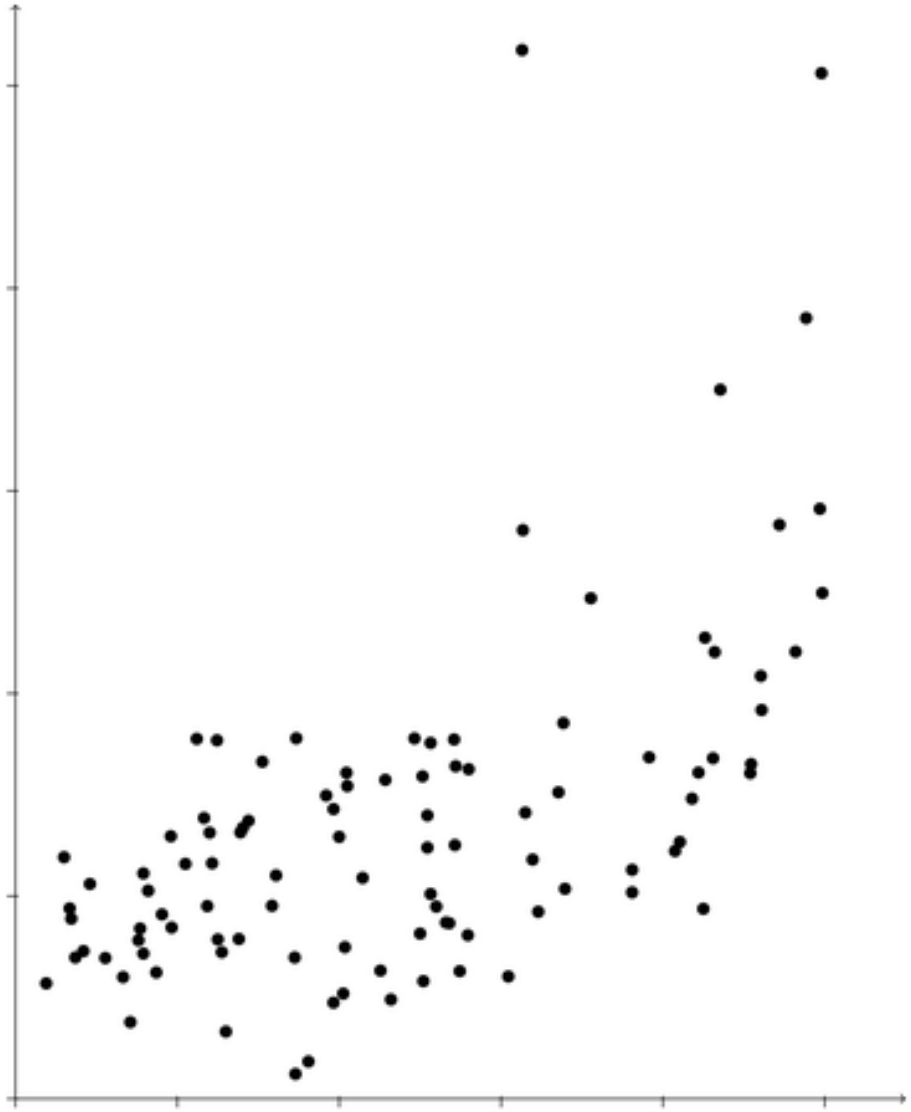
Sample Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Relationship Between Quantitative Variables



Scatterplot carries 3 types of information about the relationship between 2 quantitative variables:

1. Linearity of relationship
2. Strength of relationship
3. Direction of relationship

Alternatively (to scatterplot), such information could be conveyed numerically by simple correlation coefficients.

Correlation

- Correlation is a measure of the quantitative relationship between variables. The calculation of statistical correlation does NOT need scientific basis between X and Y.

- Some simple popular correlation coefficients:
  - Pearson product-moment correlation coefficient
  - Spearman's correlation coefficient
  - Kendall's τ

Pearson's Correlation Coefficient

- A unitless measure of the LINEAR correlation between two variables X and Y, -1 ≤ Pearson's corr ≤ 1.
- Interpretation:

  1 total positive linear correlation ("direct correlation")

  0 no linear correlation

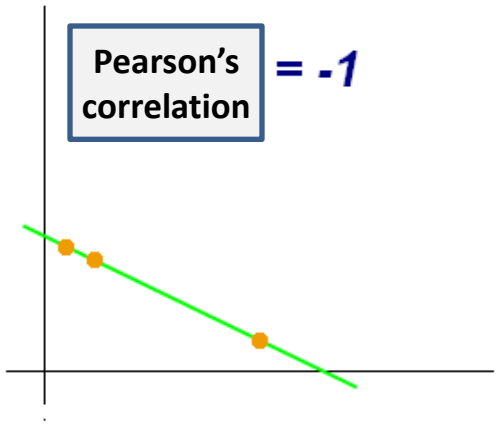
  −1 total negative linear correlation ("inverse correlation")

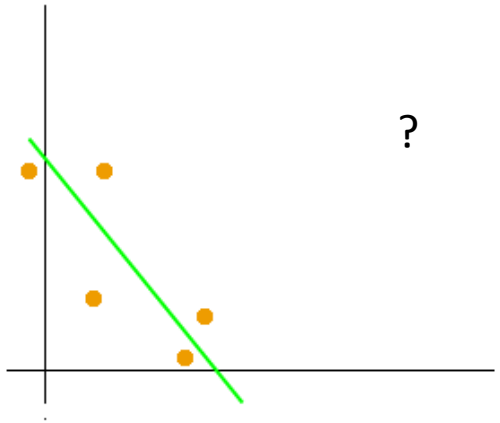
Pearson's correlation $= \dfrac{\operatorname{cov}(x, y)}{\sigma_X \sigma_Y}$

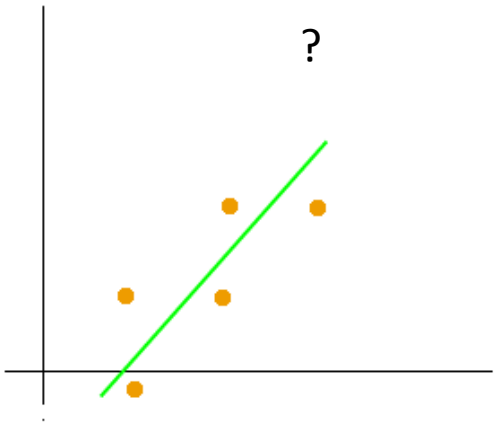
Pearson's correlation $=$ Covariance of X and Y / (Standard deviation of X × Standard deviation of Y)

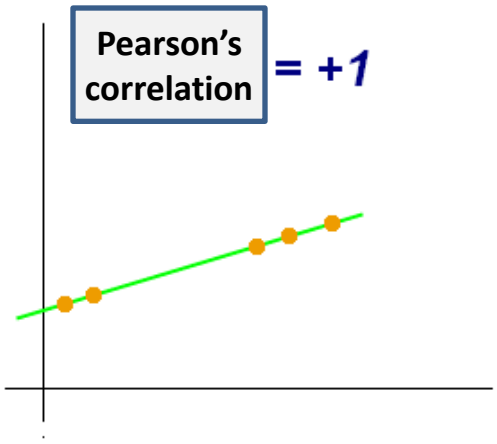
Pearson's correlation $=$ How x changes as y changes / (Variability of x × Variability of y)

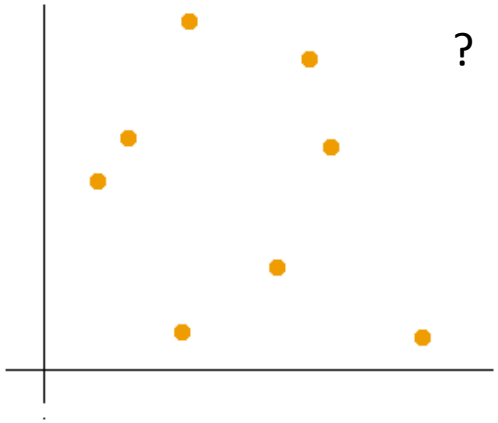
Visualization



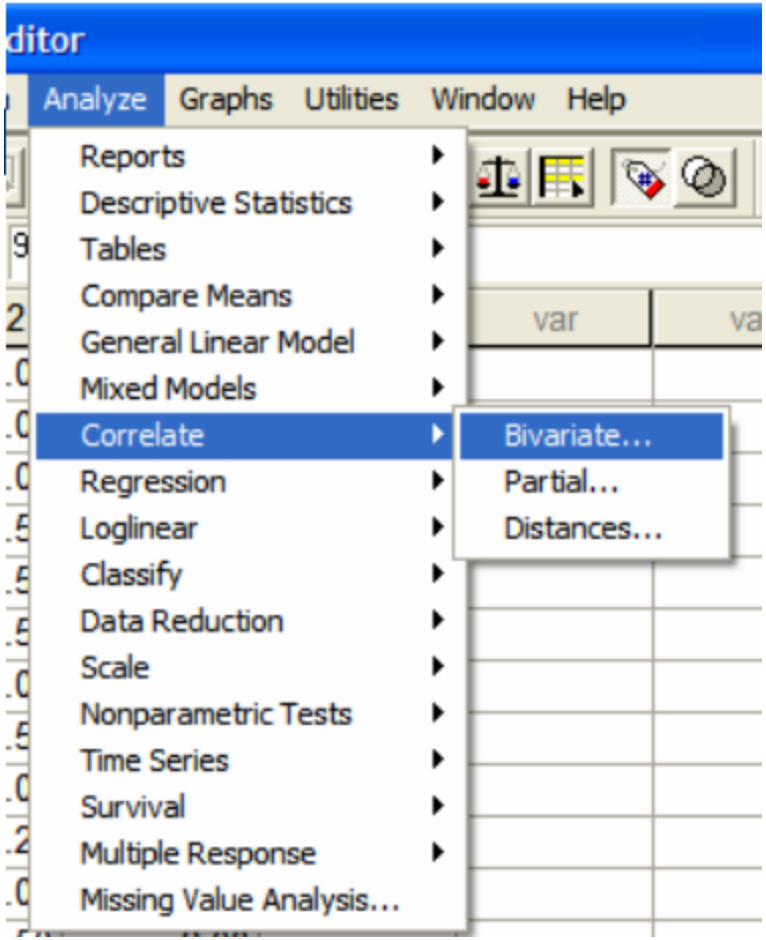


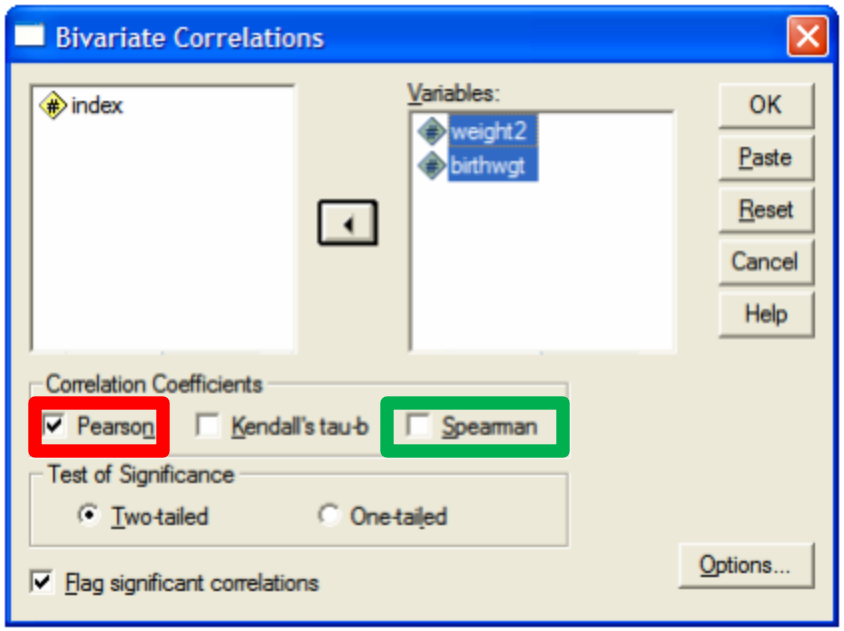




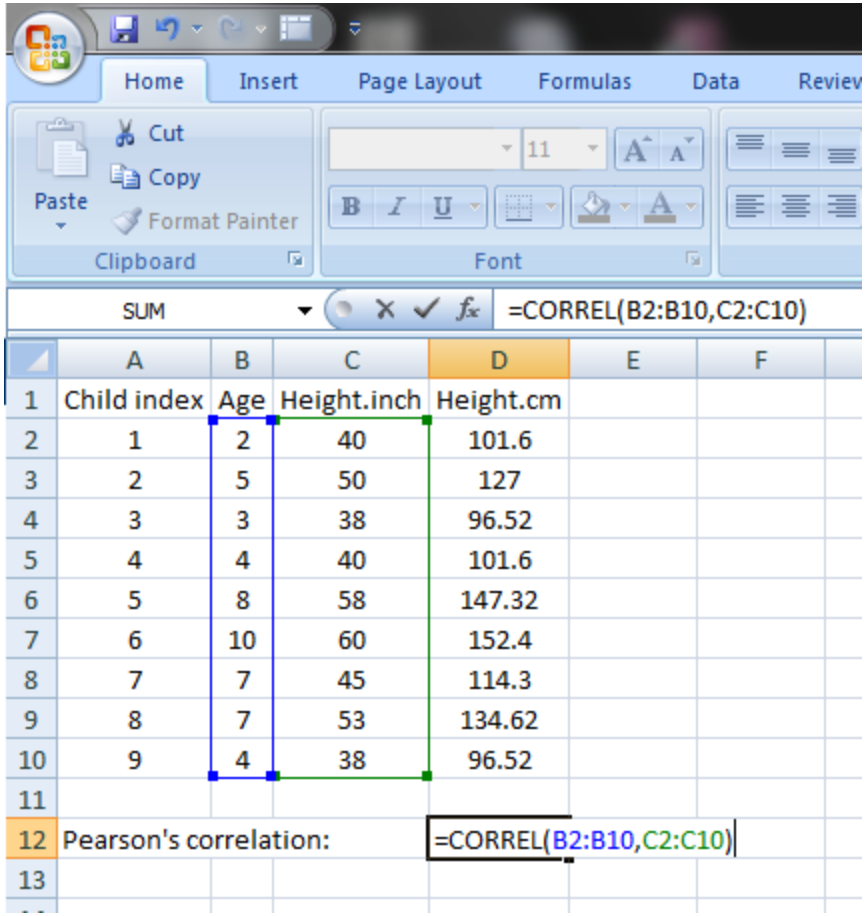


Pearson's & Spearman's Correlation Coefficients in SPSS





Pearson's Correlation in Excel

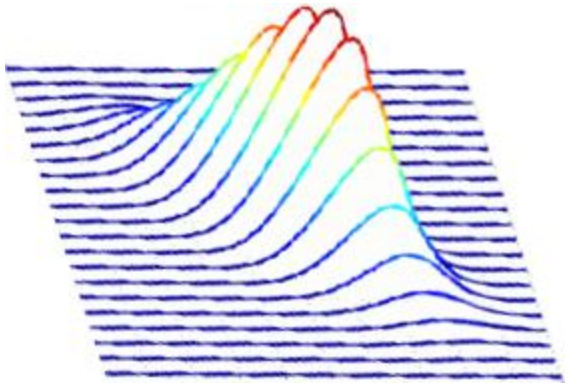


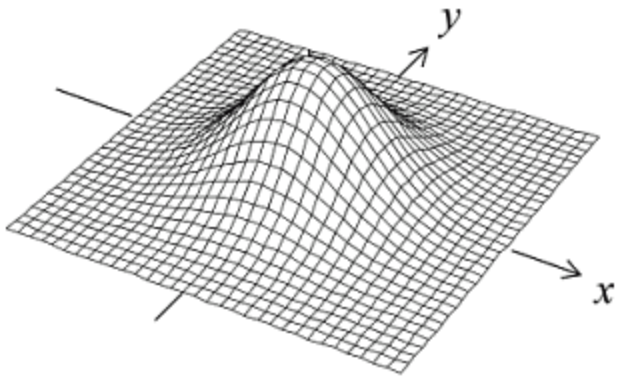
	A	B	C	D	E	F
1	Child index	Age	Height.inch	Height.cm		
2	1	2	40	101.6		
3	2	5	50	127		
4	3	3	38	96.52		
5	4	4	40	101.6		
6	5	8	58	147.32		
7	6	10	60	152.4		
8	7	7	45	114.3		
9	8	7	53	134.62		
10	9	4	38	96.52		
11						
12	Pearson's correlation:			=CORREL(B2:B10,C2:C10)		
13						

Then hit "Enter"

Assumption of Pearson's Correlation

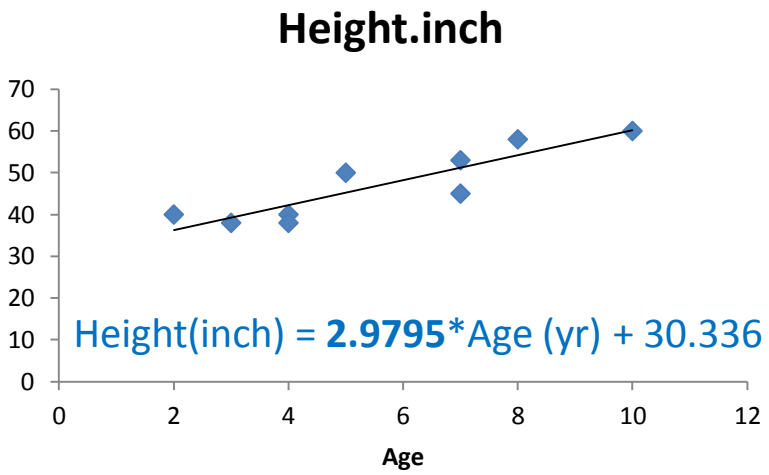
- X and Y are bivariate normal
- A reasonably linear relationship exists

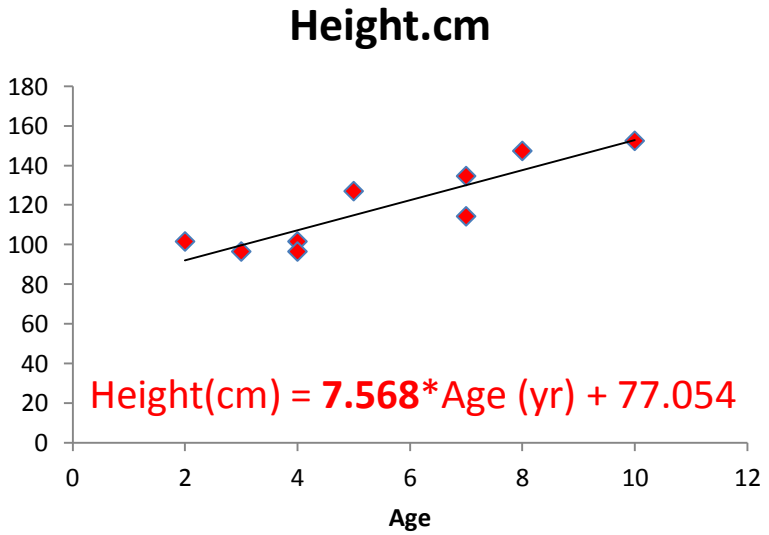




(Pearson's) Correlation vs Slope

Child ID	Age (yr)	Height (in)	Height (cm)
1	2	40	101.6
2	5	50	127
3	3	38	96.52
4	4	40	101.6
5	8	58	147.32
6	10	60	152.4
7	7	45	114.3
8	7	53	134.62
9	4	38	96.52





Pearson's correlation coefficient (0.90) is IDENTICAL in both cases

Interpretation

- What does the 0.9 mean?
- Does height cause age to increase?
- Can the increase in height with each year of growth be inferred from the Pearson's correlation coefficient?
- For an average 12-year old, calculate Height(inch) and Height(cm). What did you just do?

Height(cm) = **7.568***Age(yr) + 77.054

Height(inch) = **2.9795***Age(yr) + 30.336

Contrast Results from 3 Correlation Coefficients

Pearson corr	0.89652
Spearman's corr	0.88136
Kendall's τ	0.76471

Take-Home Point:

Employing different methods, statistics DO summarize data; Regardless of methods, statistical methods DO NOT create data.

Child ID	Age (yr)	Height (in)	Height (cm)
1	2	40	101.6
2	5	50	127
3	3	38	96.52
4	4	40	101.6
5	8	58	147.32
6	10	60	152.4
7	7	45	114.3
8	7	53	134.62
9	4	38	96.52

On the Choice of Correlation Coefficients

- Pearson: reasonably linear relationship.
- Spearman's: reasonably linear relationship with outliers
- Kendall's: 2 ordinal (rank) variables.

Epilogue

- Look before compute
- Describe before infer