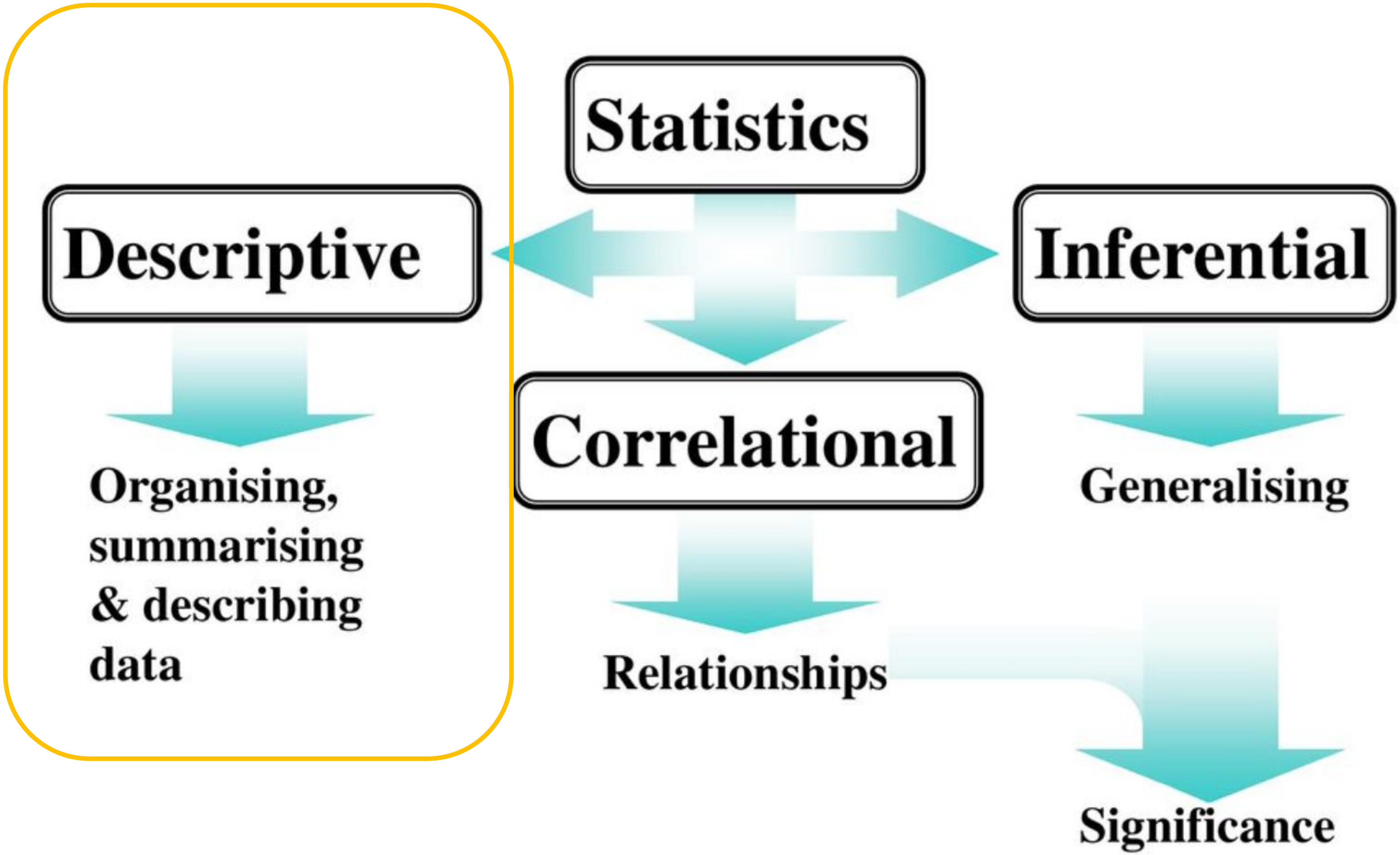
BRSM
Descriptive Statistics, Correlation

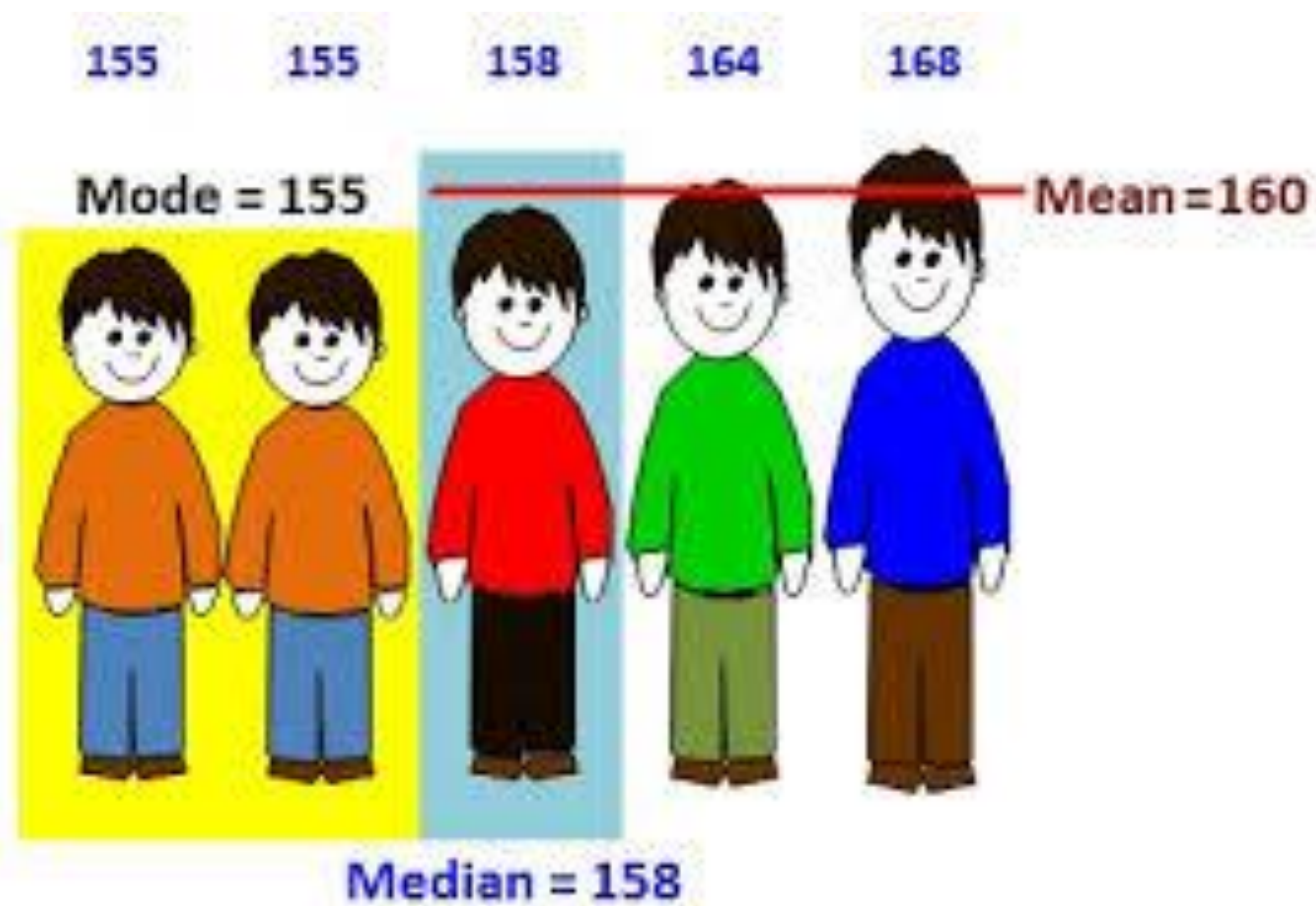
Vinoo Alluri & Bapi Raju


Statistics
Descriptive
Inferential
Correlational
Organising, summarising & describing data
Generalising
Relationships
Significance

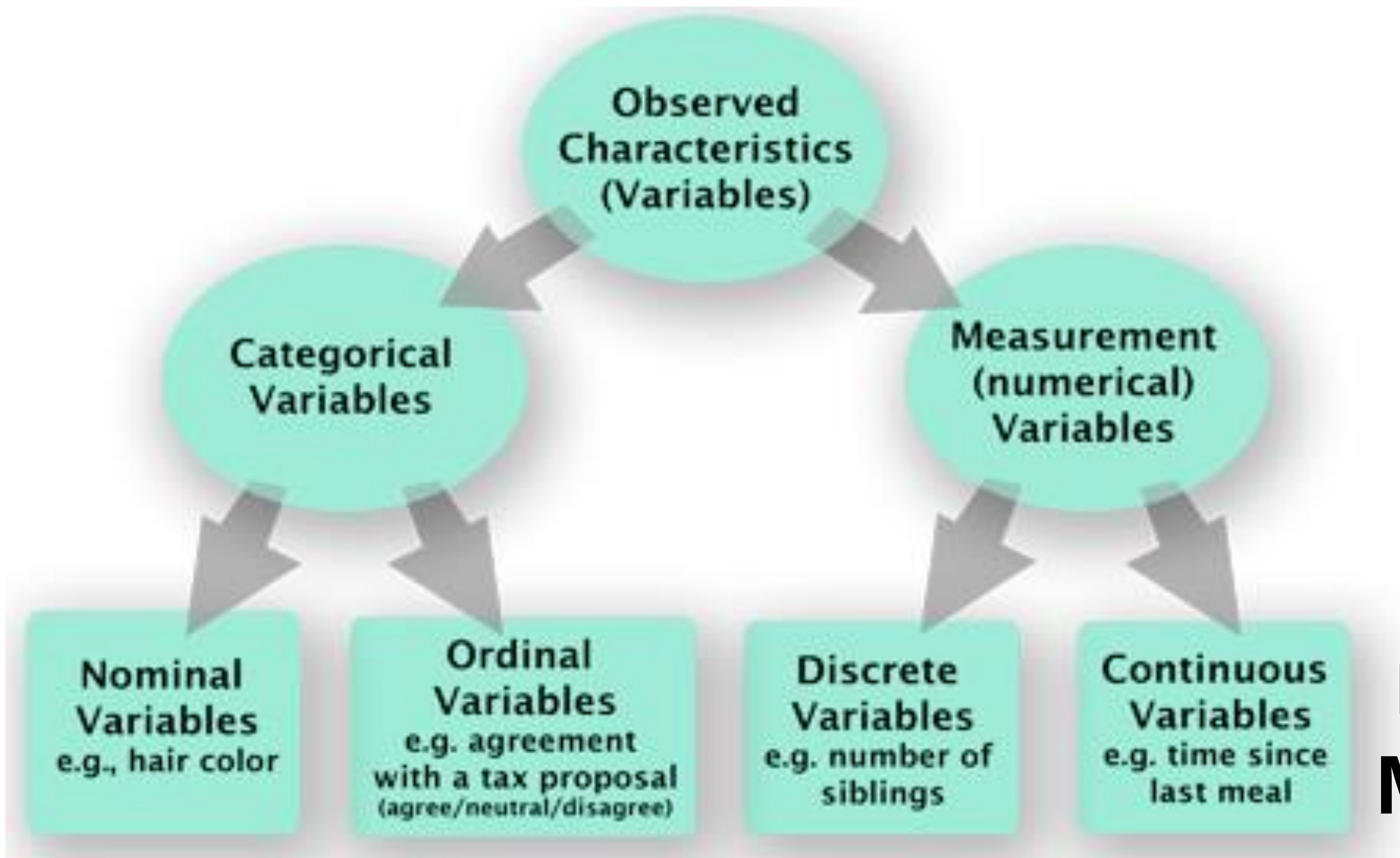
Descriptive Statistics

- Common descriptive statistics are:
  - Measure of **central tendency**
    - the most typical value of a given group of values
  - Measure of **dispersion**
    - how much all the other values in the group vary around the typical value

Measures of central tendency



Central Tendency for Variable Types



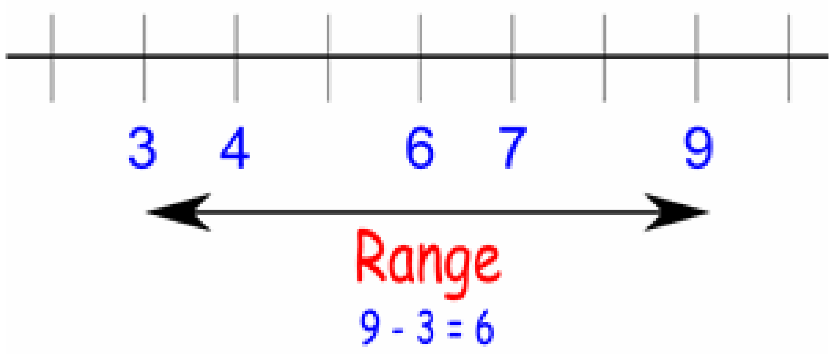
MODE
Median

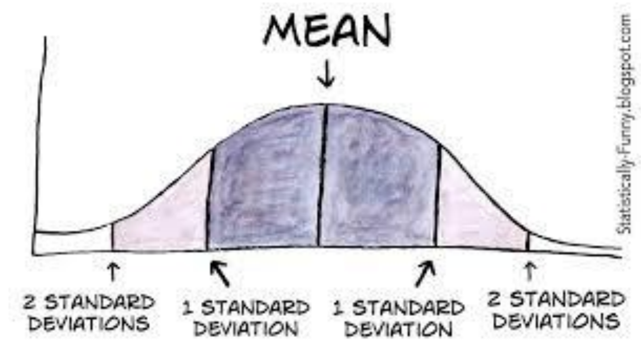
MODE **Median** **Mean** **Mean**

Measures of central tendency

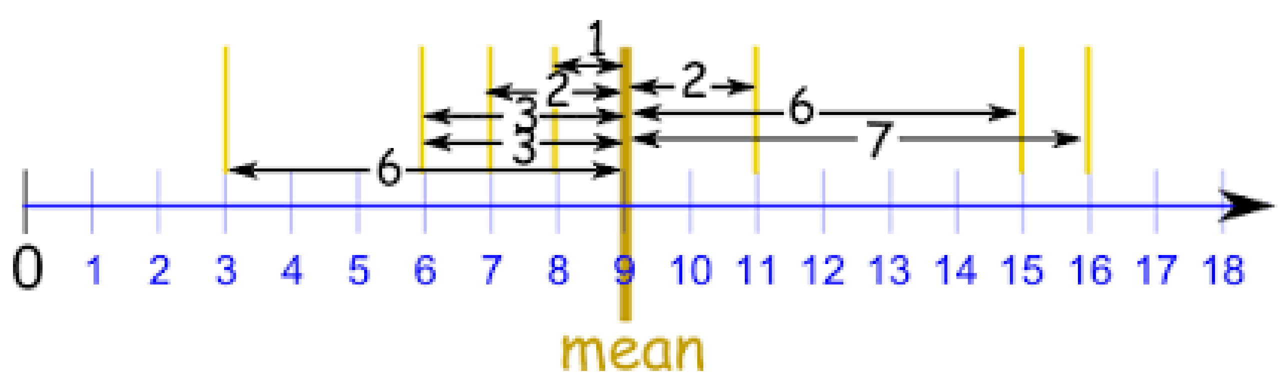
	Advantages	Disadvantages
Mean		
Median		
MODE		

Measures of dispersion/spread



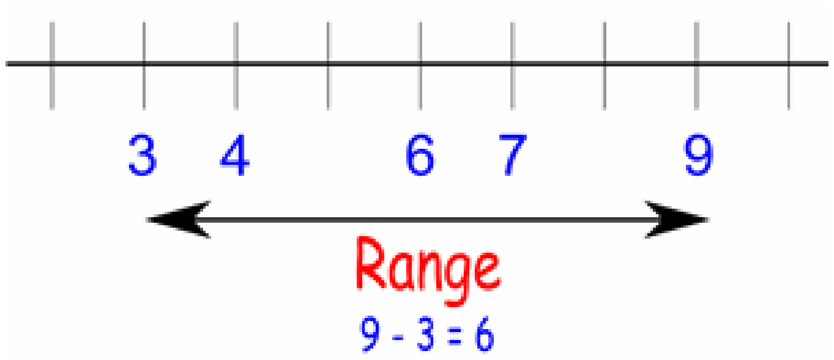
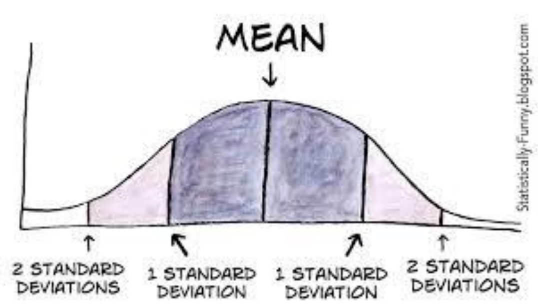


$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

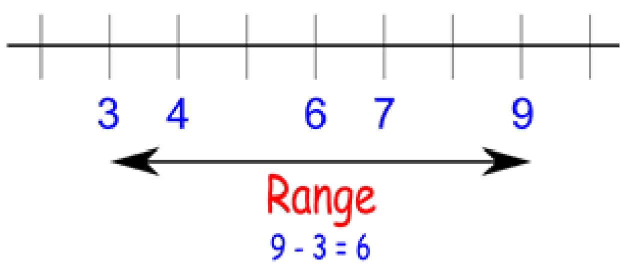


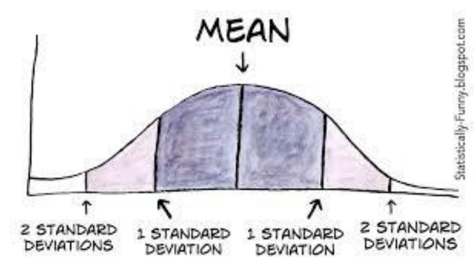
$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

Measures of dispersion/spread

	Advantages	Disadvantages
	——	distorted by extreme values no indication of grouping around the mean
 $\mathrm{SD} = \sqrt{\frac{\sum \lvert x - \bar{x} \rvert^2}{n}}$	- Fundamental to significance testing, and forms basis of Analysis of Variance (ANOVA) - Enables population parameters to be estimated from a sample of people	——

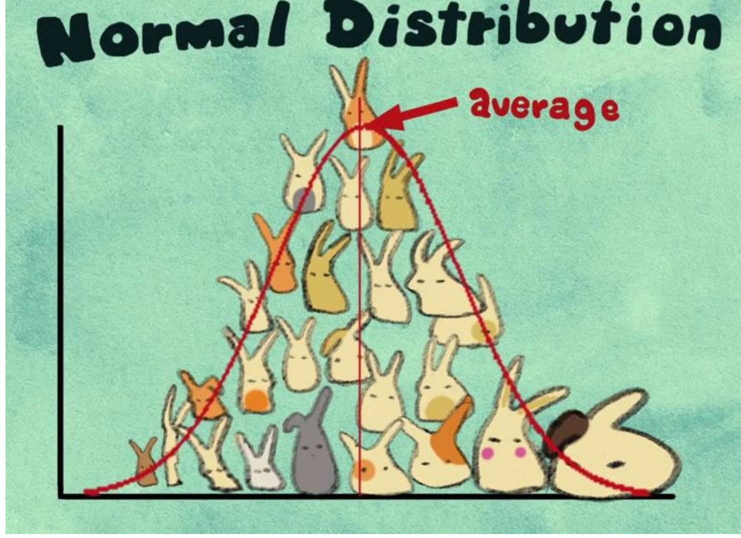
MEAN ? MODE MEDIAN





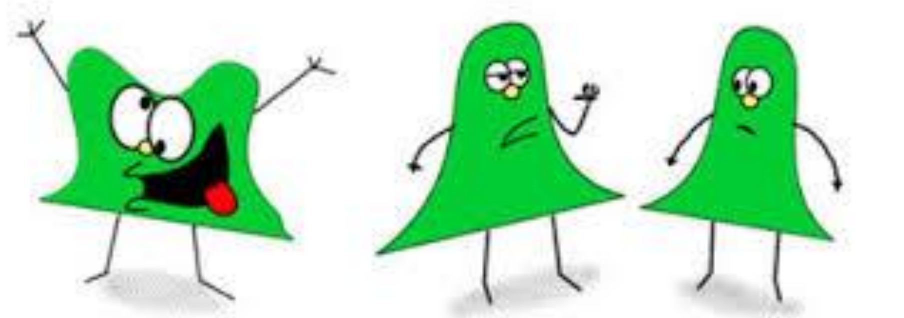
When do these measures fail to be representative ????

Normal Distribution

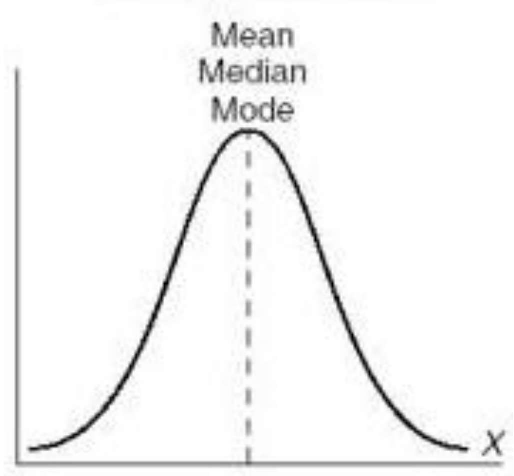


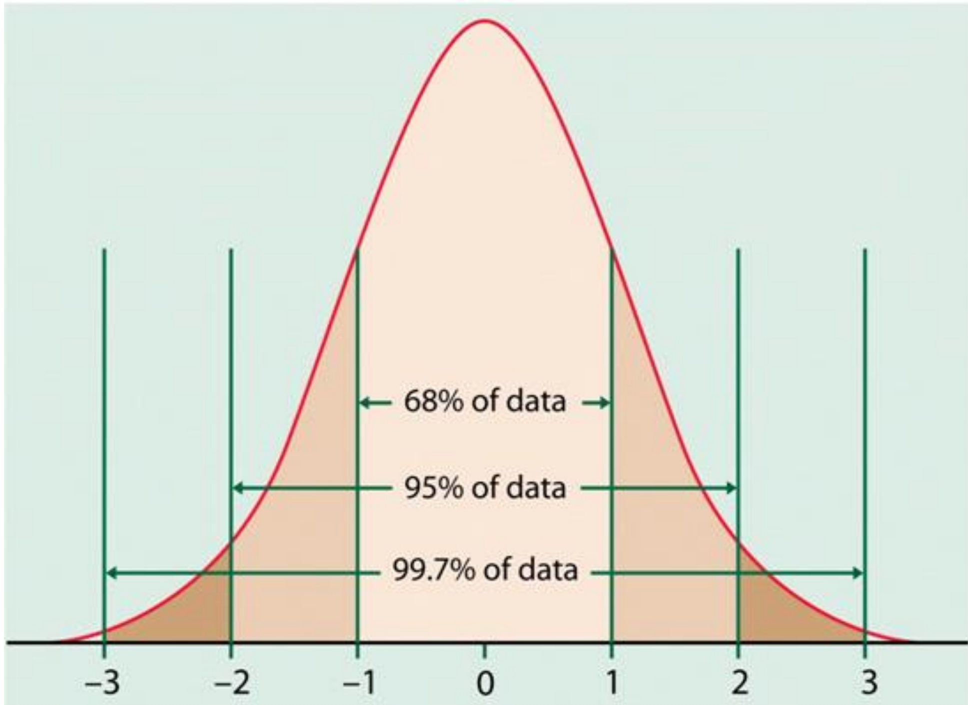
- A bell-shaped mathematical curve describing how values are distributed
- Data taken from a sample is **assumed** to be 'normally distributed', and must approximate this shape in order to use parametric tests of significance
- *Inferential statistics* (eg: t-tests, F-tests, regression analyses) require in some sense that the numeric variables are approximately normally distributed
- *Note:* it does not fit all populations

Normal Distribution

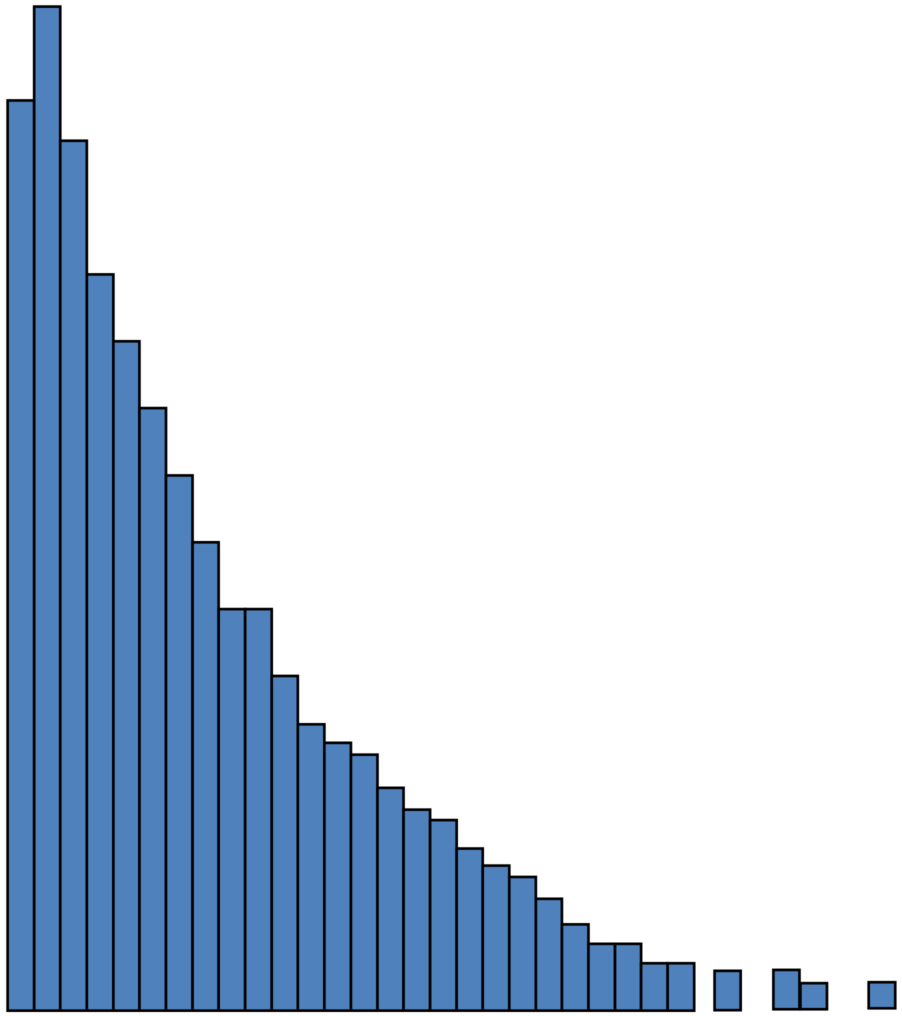

"KEEP YOUR EYE ON THAT GUY, TOM. HES NOT, YOU KNOW...NORMAL!"

- symmetrical about the horizontal axis midpoint
- mean, median, and mode all fall on the midpoint
- No matter what μ and σ are, the area between
  - μ-σ and μ+σ is about 68%;
  - μ-2σ and μ+2σ is about 95%;
  - μ-3σ and μ+3σ is about 99.7%
- Almost all values fall within 3 standard deviations



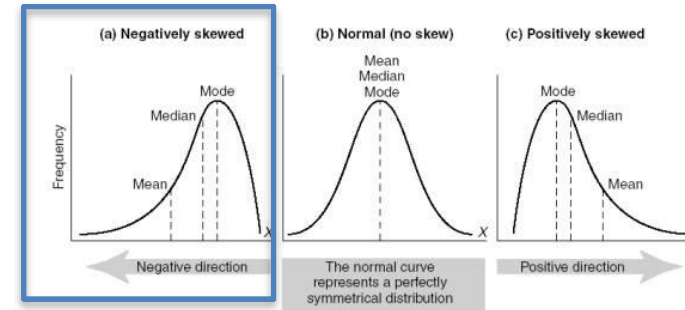


Skewed Distribution

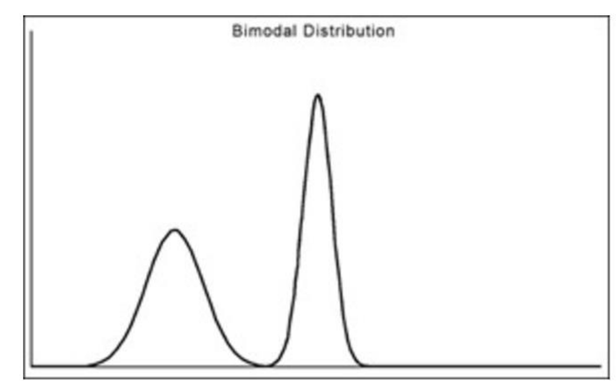


- Resembles an exponential distribution
- Lots of extreme values far from mean or mode
- Not straightforward to do useful statistical tests with this type of distribution

Skewed Distribution

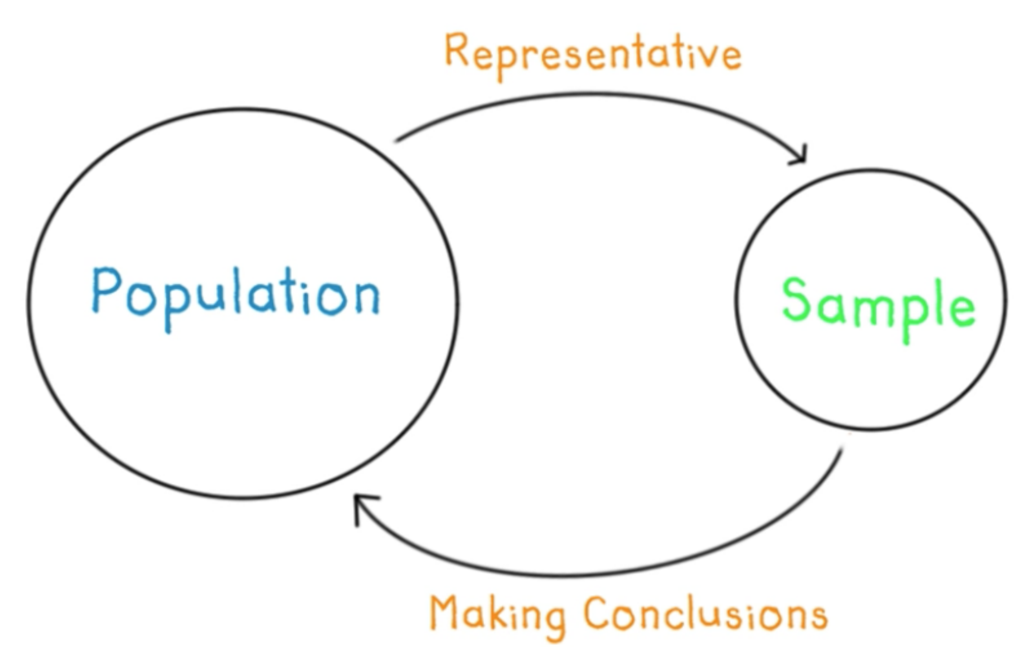


- **Negative skew**
  - Result from relatively easy tasks, due to a ceiling effect
- **Positive skew**
  - Results from tasks which are hard to improve upon, due to a floor effect (such as RT —reaction time)
- **Bimodal**
  - Two distinct peaks
  - probable indicator of groups
  - ex: completion time of marathon runners

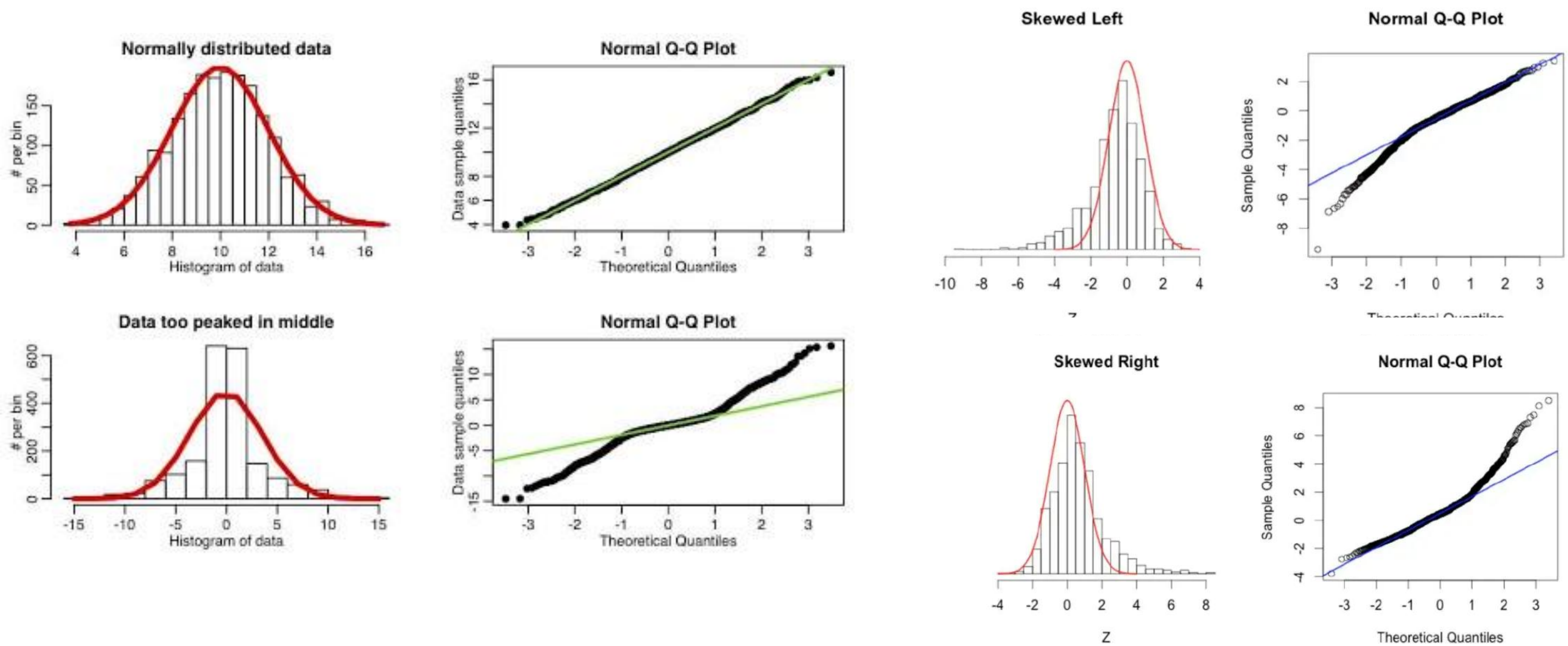


Normality in Real-World Data

- real-world data is usually skewed
- parametric tests assume that we are sampling from a normally distributed population



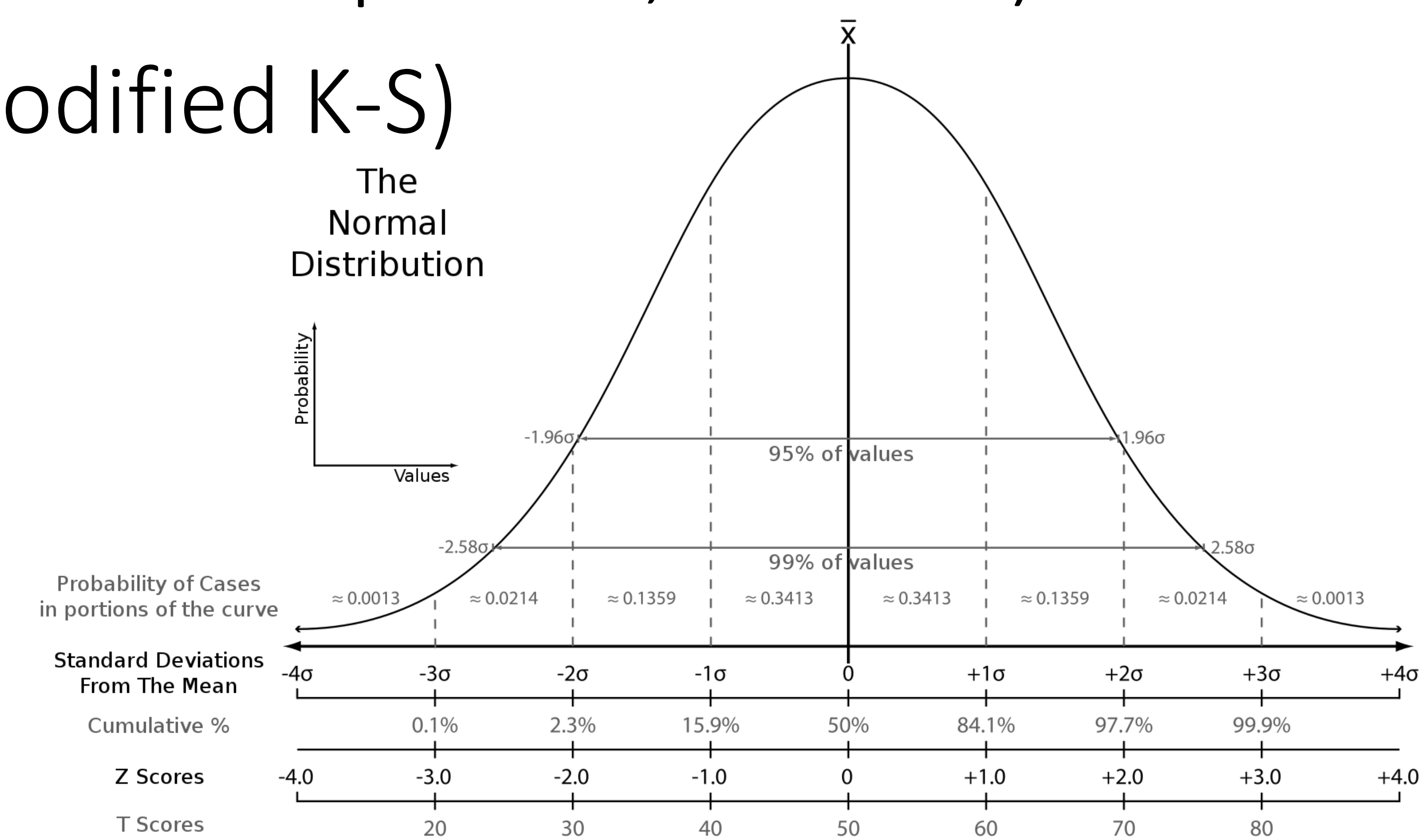
Testing Normality



- Q-Q plot: graphical technique (can also use it to test any theoretical distribution)
- theoretical quantiles plotted on x-axis and sample quantiles plotted on y-axis

Testing Normality

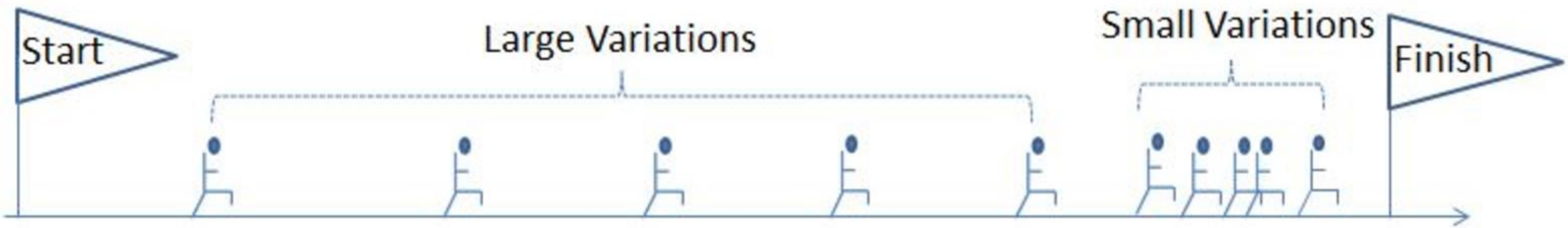
- Tests to assess normality (null hypothesis: data are sampled from a population that follows a normal distribution)
    - Kolmogorov-Smirnov (≥ 50)
    - Shapiro-Wilk (for smaller sample size, i.e. < 50)
    - Anderson-Darling (modified K-S)
    - Lilliefors test
    - Cramer-von Mises
    - etc..

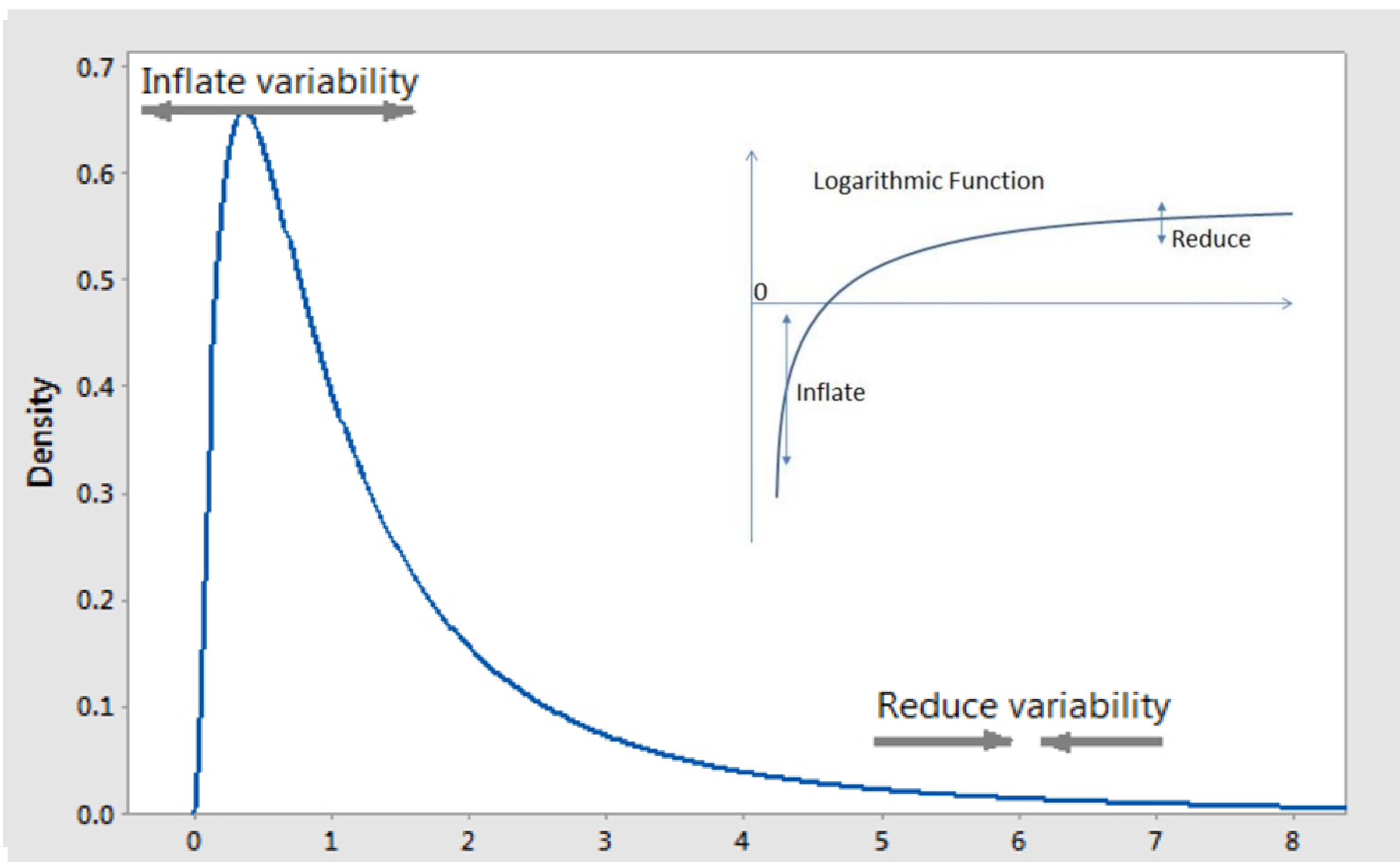


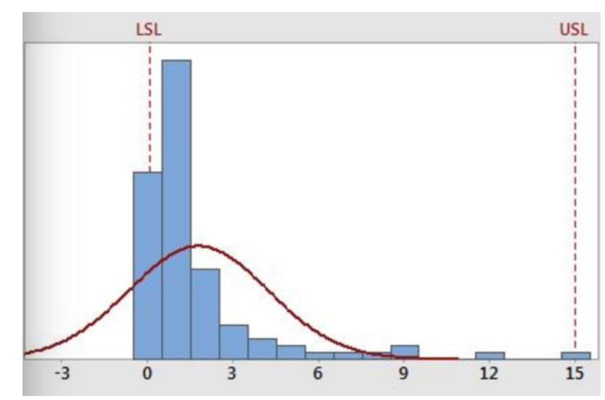
Testing Normality

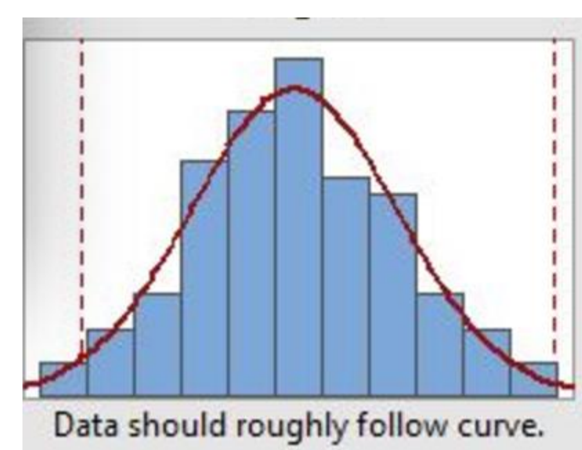
- For non-normal data
  - transform to normal distribution (eg: sqrt, log)
    - if it works - use parametric tests
    - if still not normal - use non-parametric tests
  - if you have groups of data, you **MUST** test each group for normality.

EXAMPLE


Start
Large Variations
Small Variations
Finish


Inflate variability
Logarithmic Function
Reduce
0
Inflate
Reduce variability
Density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8


LSL
USL
-3
0
3
6
9
12
15


Data should roughly follow curve.

Normality Transforms

Moderately positive skewness	sqrt(X)
Substantially positive skewness	$\log_{10} X$
Substantially positive skewness (with zero values)	$\log_{10} (X + C)$
Moderately negative skewness	sqrt(K-X)
Substantially negative skewness	$\log_{10} (K-X)$

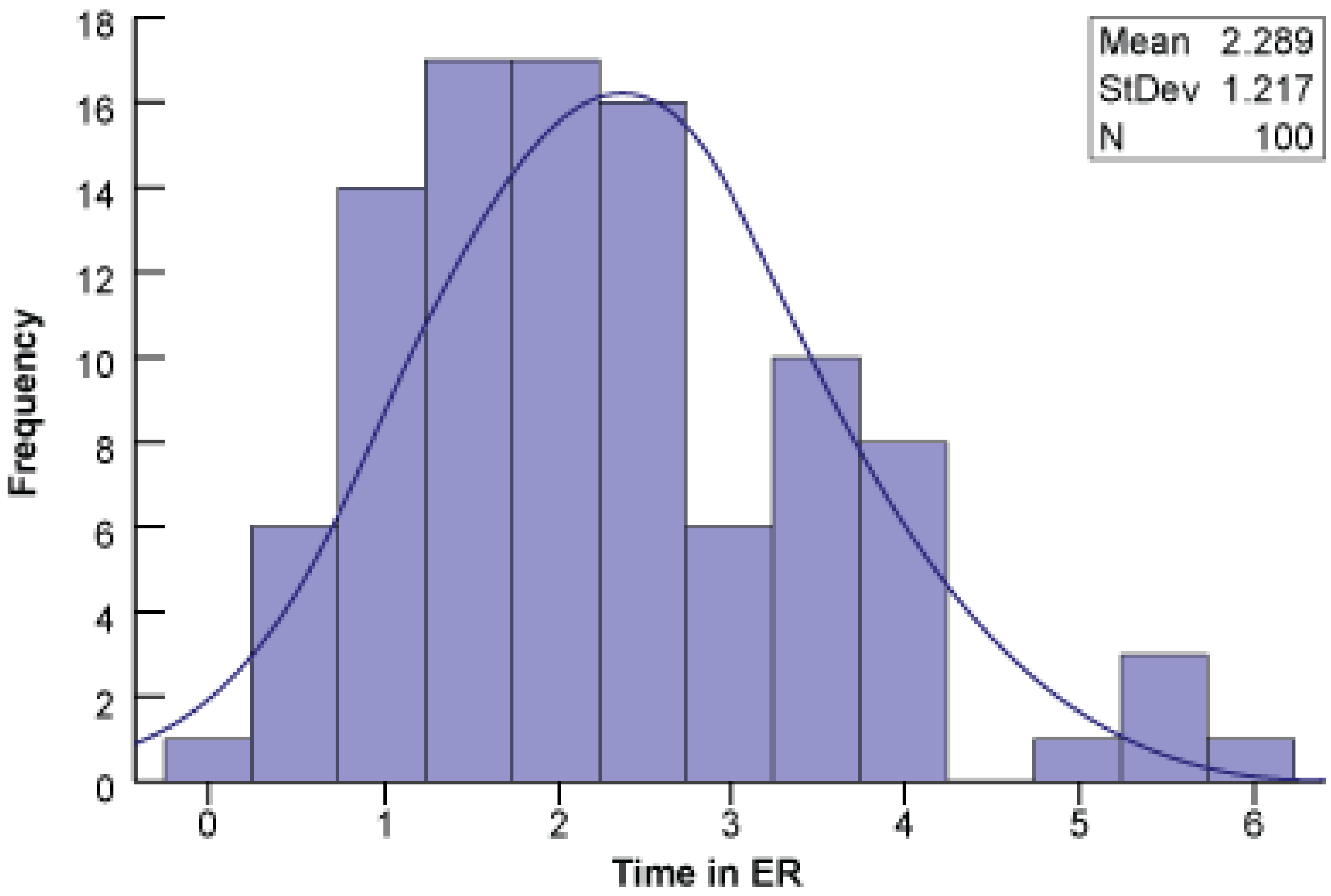
C = a constant added to each score so that the minimum score is 1

K = a constant from which each score is subtracted so that the minimum score is 1

Box-Cox transformation

- Box & Cox (1964) developed a procedure to identify an appropriate exponent (Lambda = l) to use to **transform non-normal data into a "normal shape."**
- power transformation
- increases the applicability and usefulness of statistical techniques based on the normality assumption
- is **not** a guarantee for normality
- only works if all the data is positive and greater than 0 (adding a constant (c) to all data)

hospital’s target time for processing, diagnosing and treating patients entering the ER

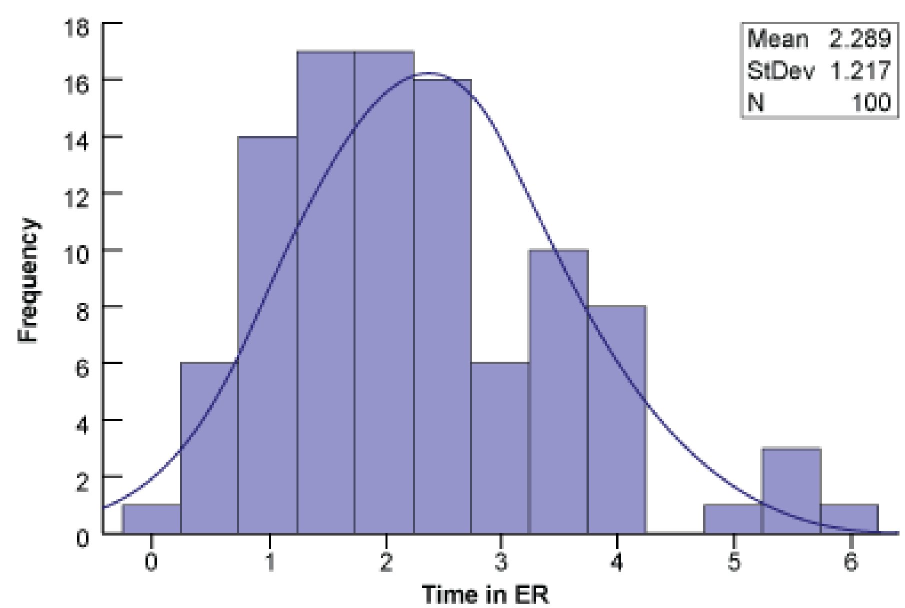


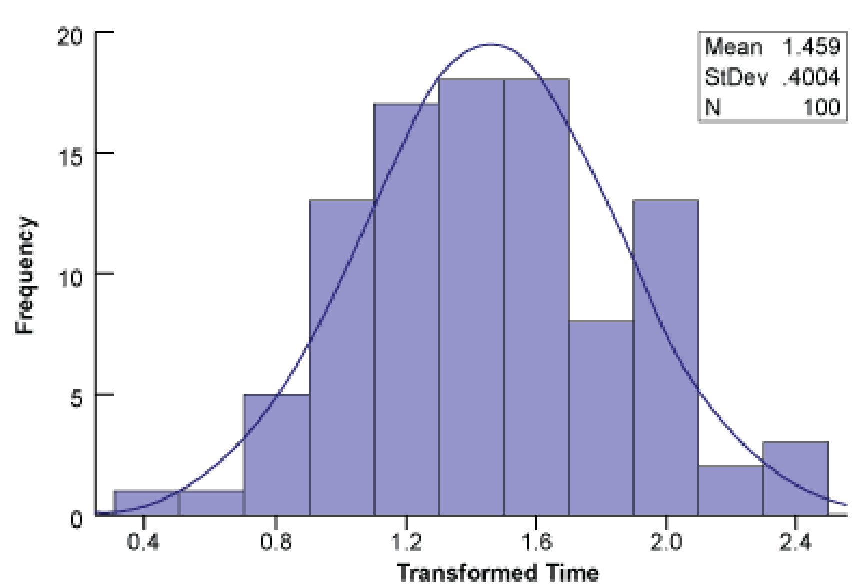
typically it is four hours or less

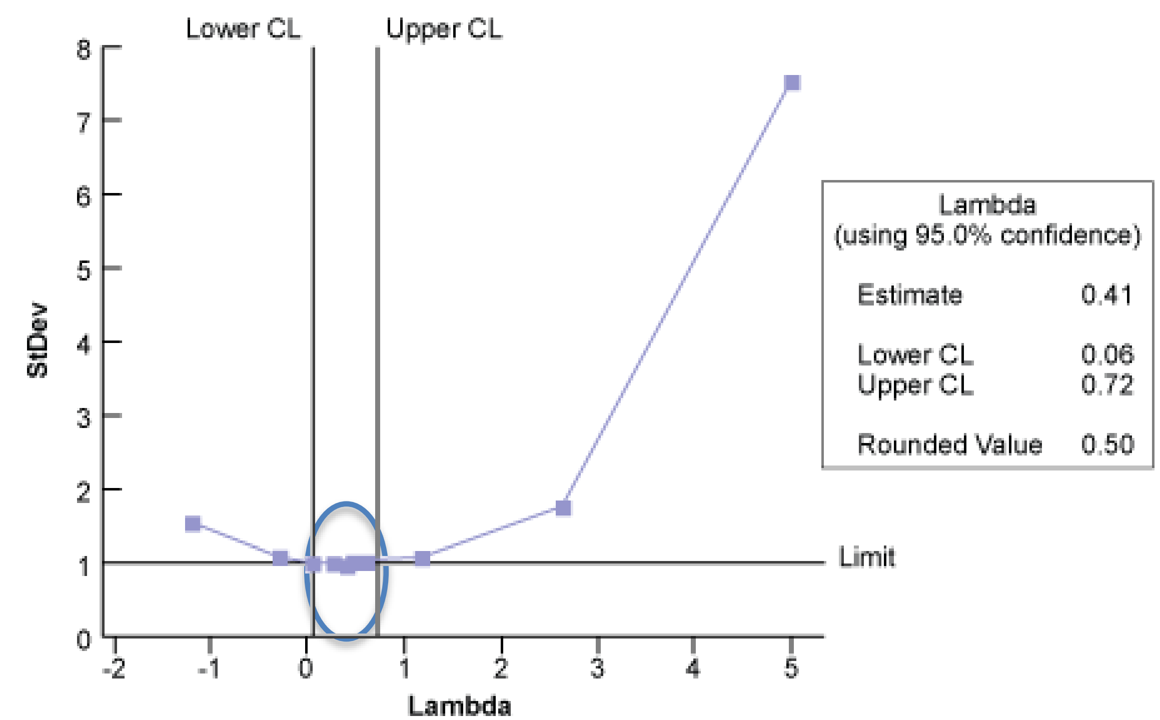
EXAMPLE

hospital's target time for processing, diagnosing and treating patients entering the ER

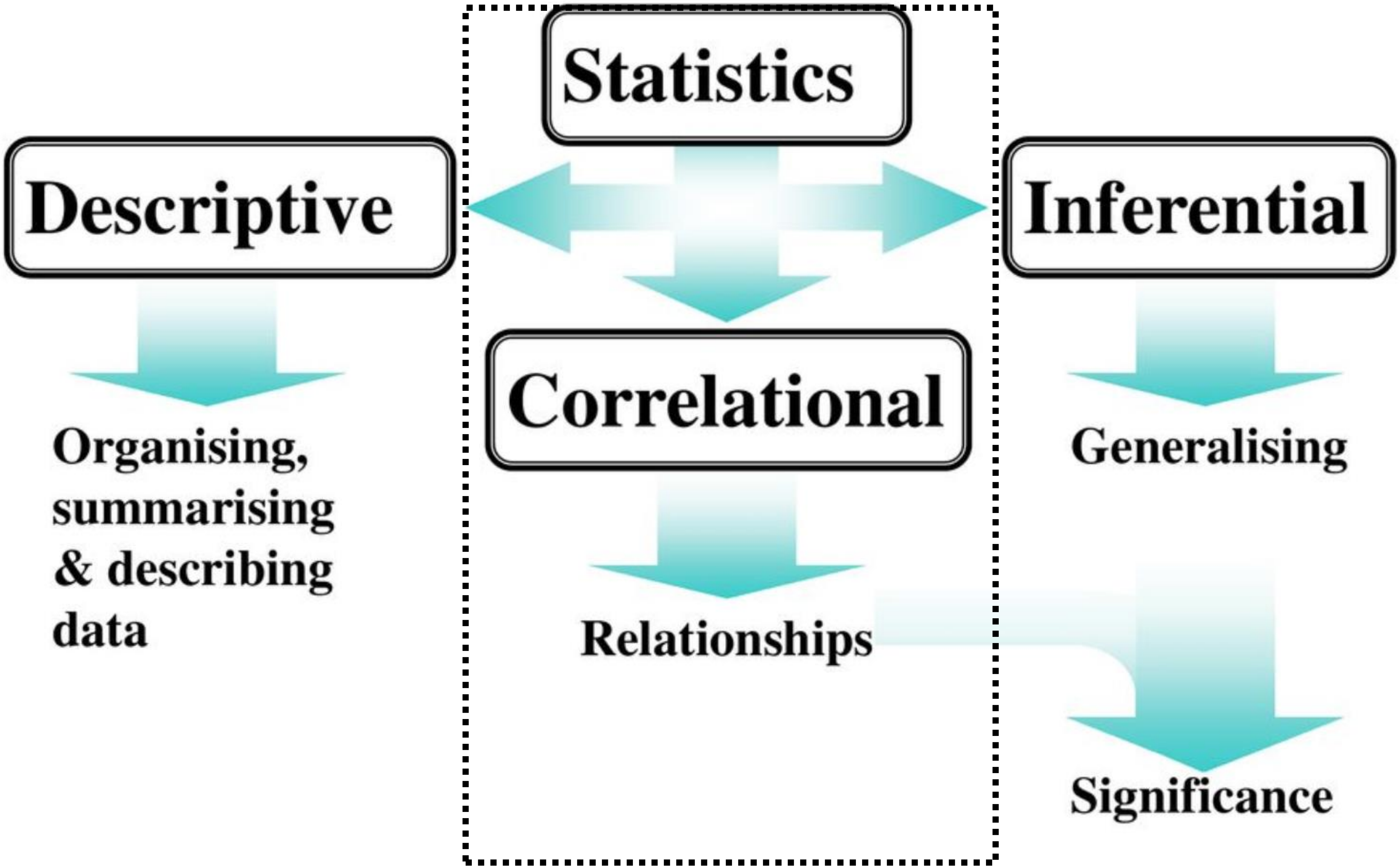
the "optimal value" is the one which results in the best approximation of a normal distribution curve



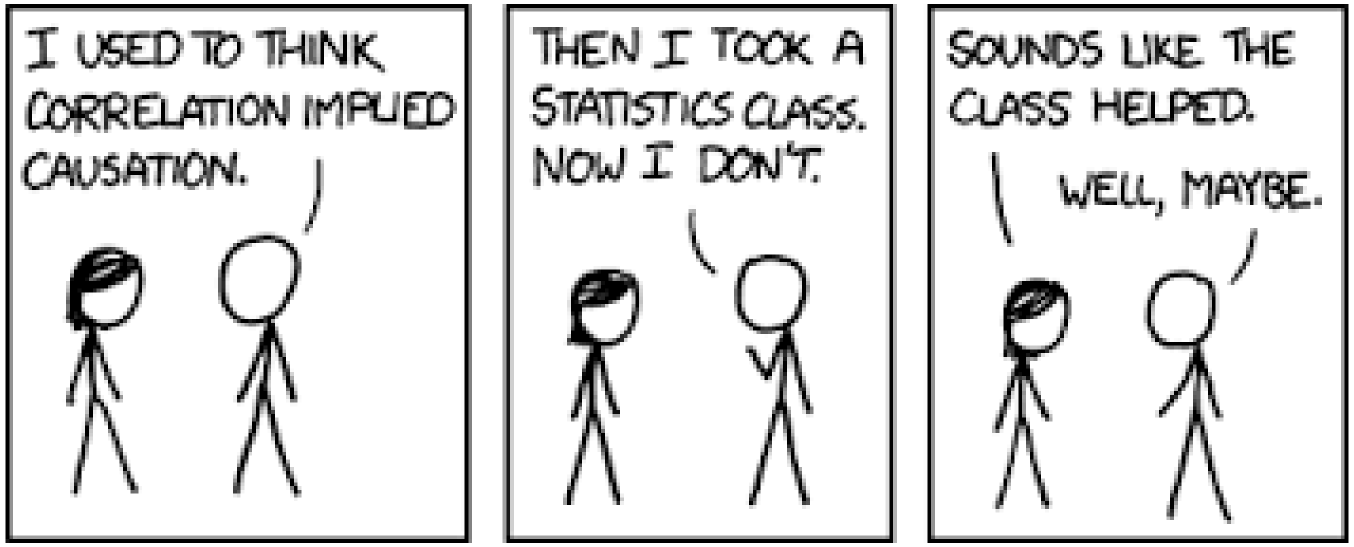




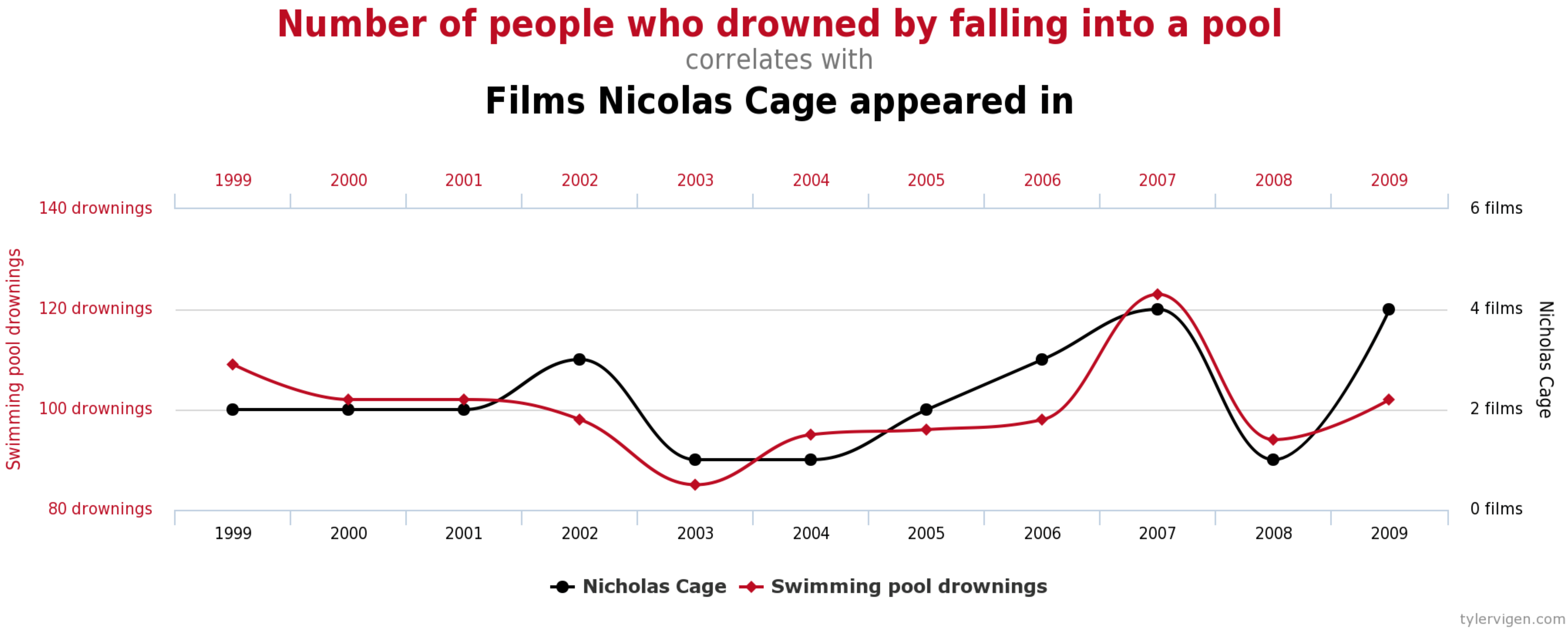
$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ (y_t^{\lambda} - 1)/\lambda & \text{otherwise.} \end{cases}$$


Statistics
Descriptive
Inferential
Correlational
Organising, summarising & describing data
Relationships
Generalising
Significance

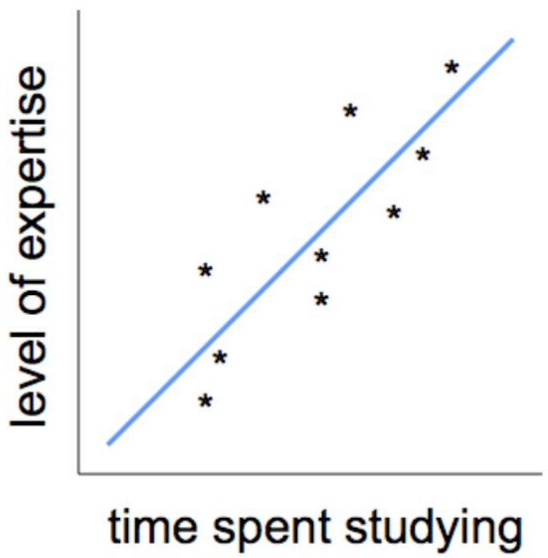
Correlation

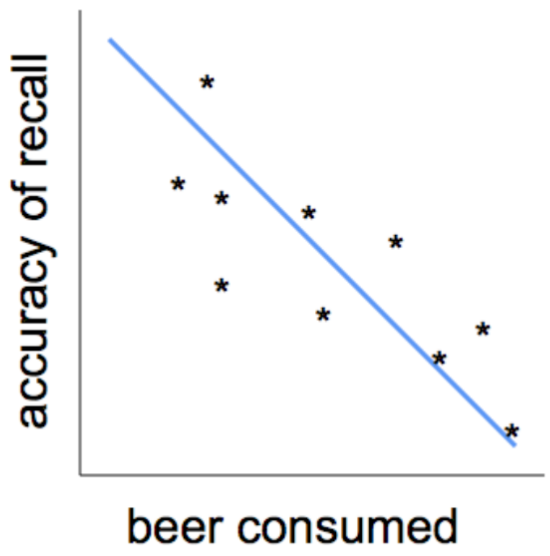


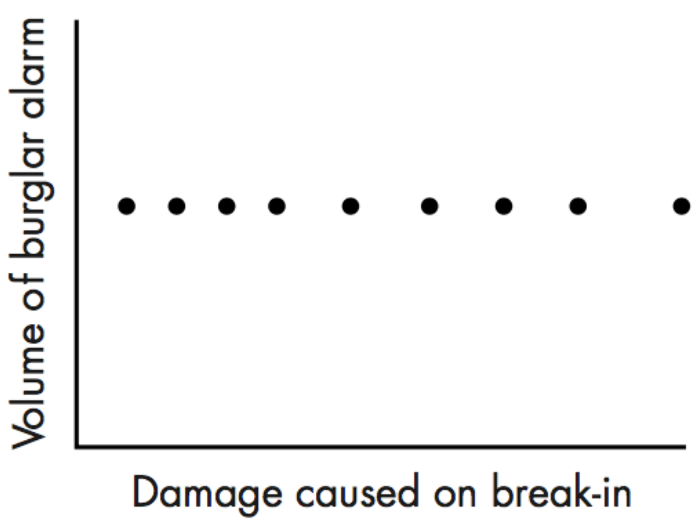
Not Causality

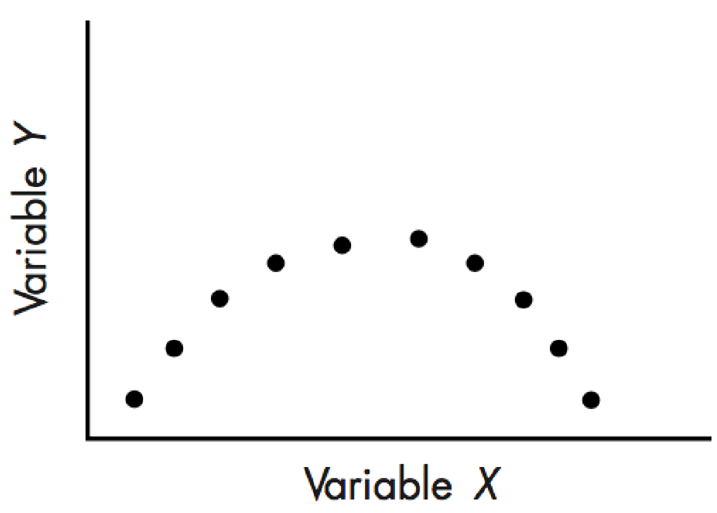


Correlation

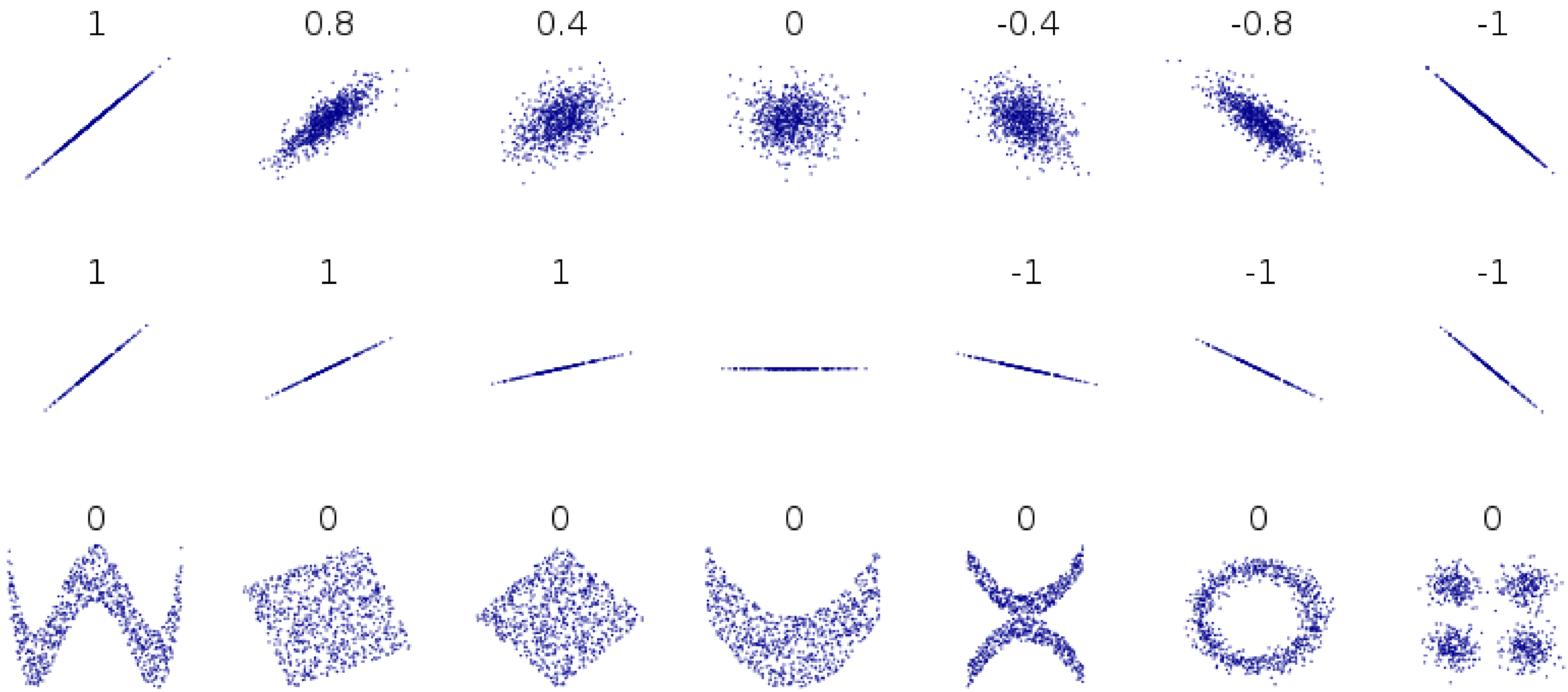








Pearson's r



$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Correlation

- calculation of correlation between two variables is a descriptive measure of the association
- testing the correlation for significance is an inferential procedure

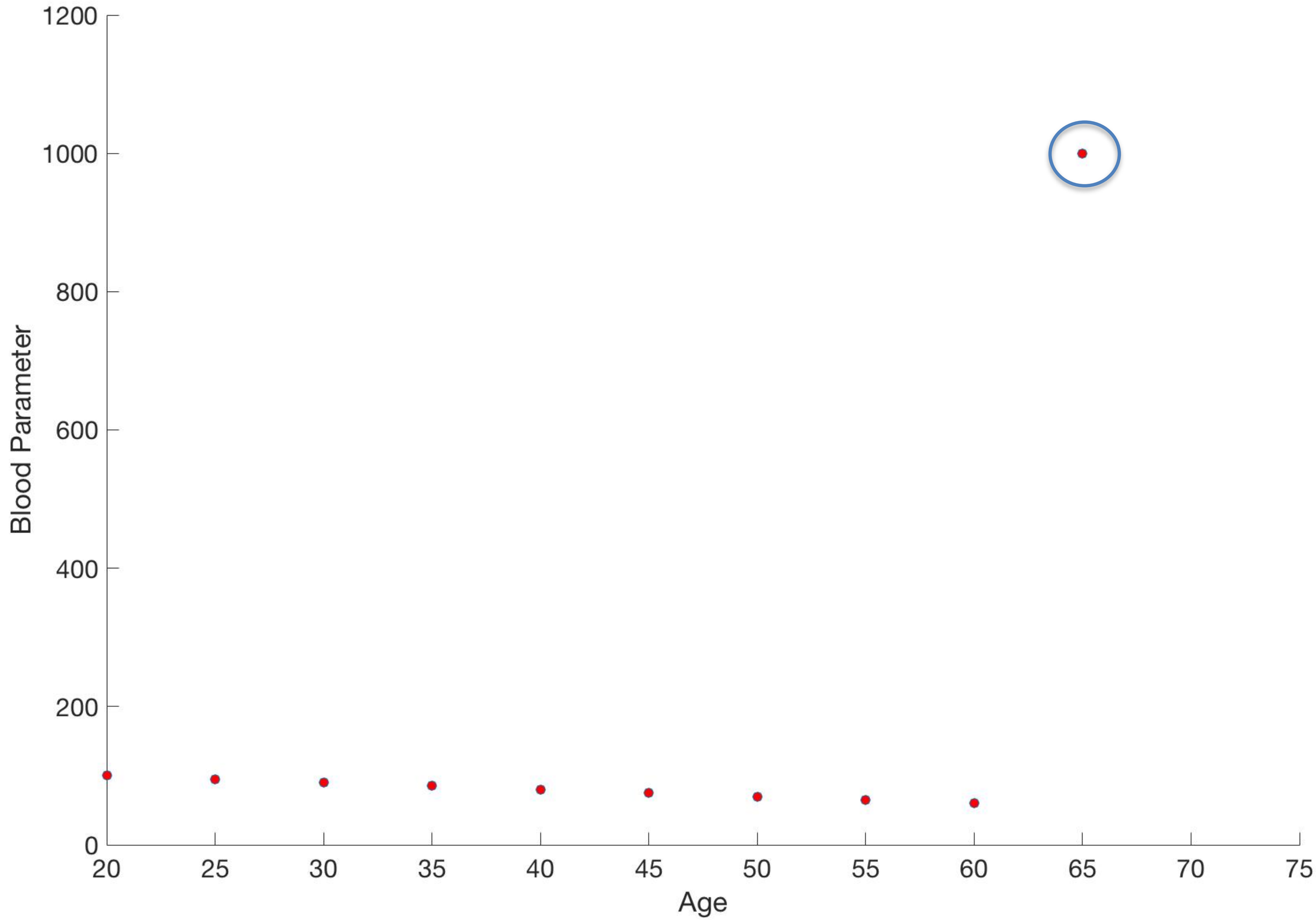
Variable Y\X	Quantitiative X	Ordinal X	Nominal X
Quantitative Y	Pearson r	Biserial r_b	Point Biserial r_{pb}
Ordinal Y	Biserial r_b	Spearman rho/Tetrachoric r_{tet}	Rank Biserial r_{rb}
Nominal Y	Point Biserial r_{pb}	Rank Bisereal r_{rb}	Phi, L, C, Lambda

r = correlation coefficient

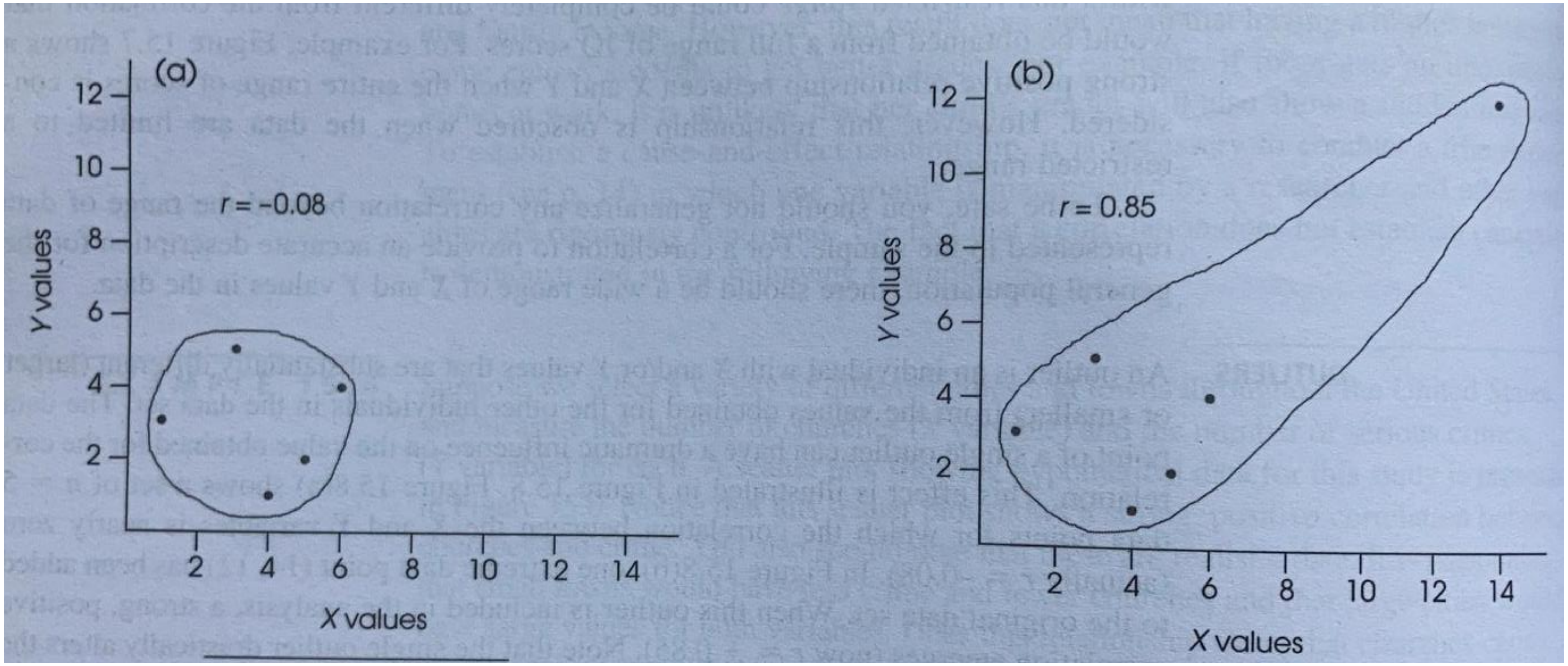
r^2 = coefficient of determination

r = ?

Pearson's ***r*** = .48

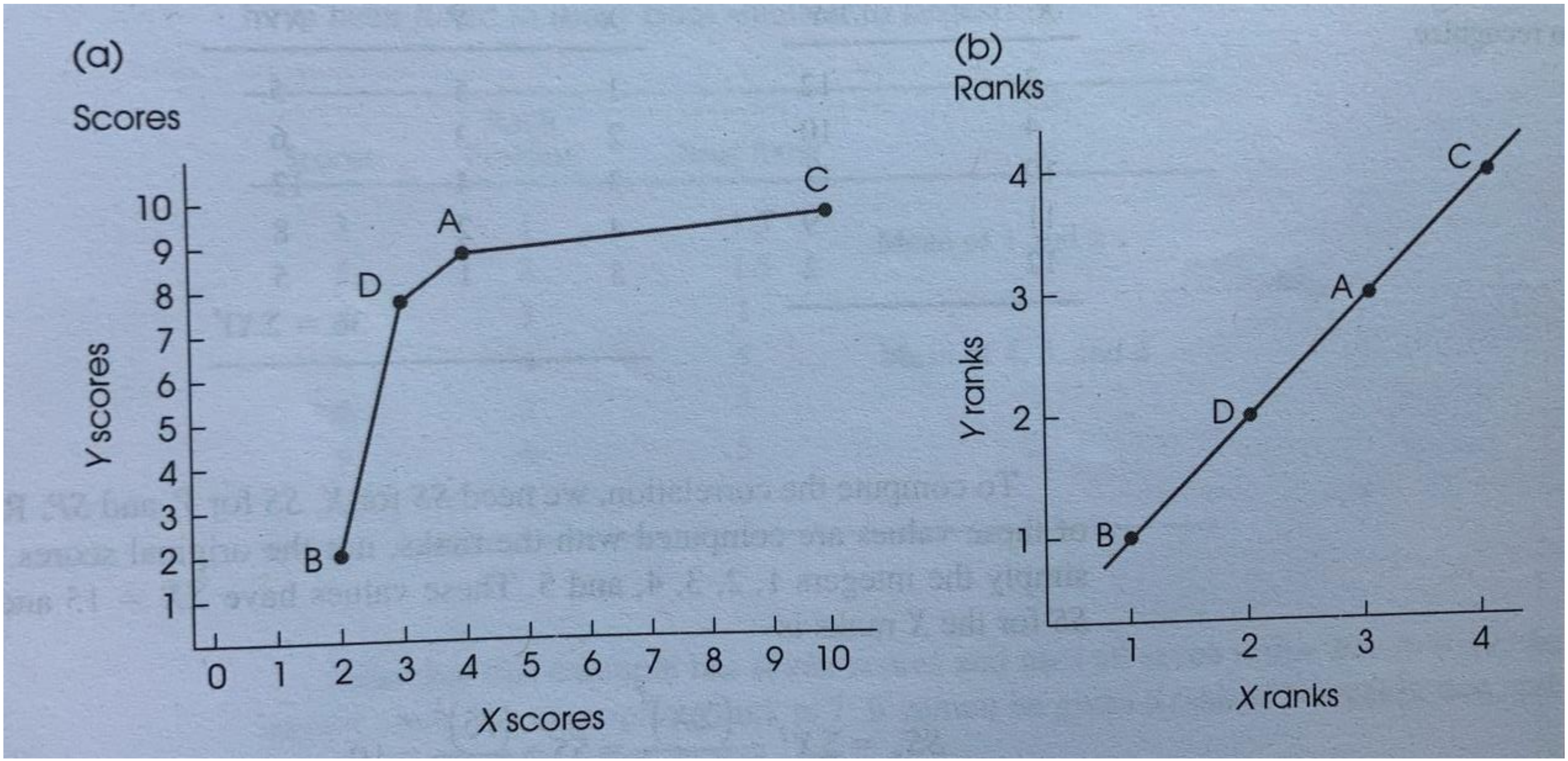


Pearson's r



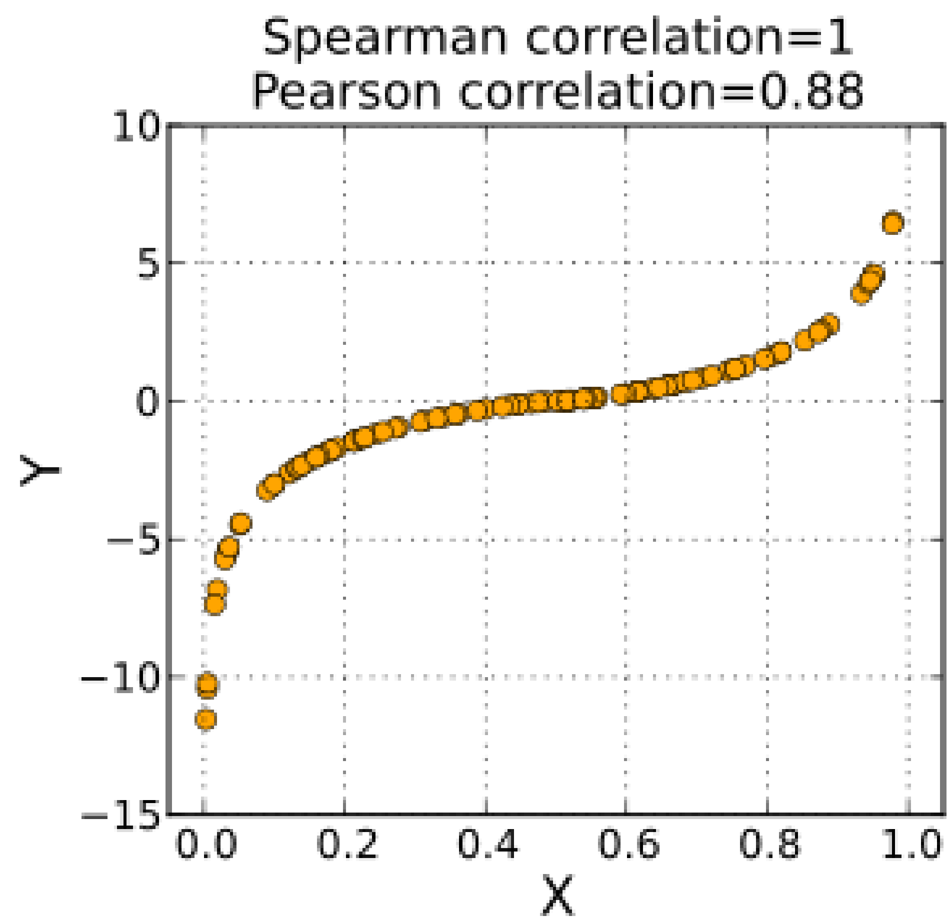
sensitive to outliers

Spearman's *rho*



Spearman's rho

- Pearson's correlation coefficient on the ranks of the data
- deals with ordinal data
- If there are no repeated values, a perfect Spearman's correlation occurs when each of the variables is a perfect monotone function of the other



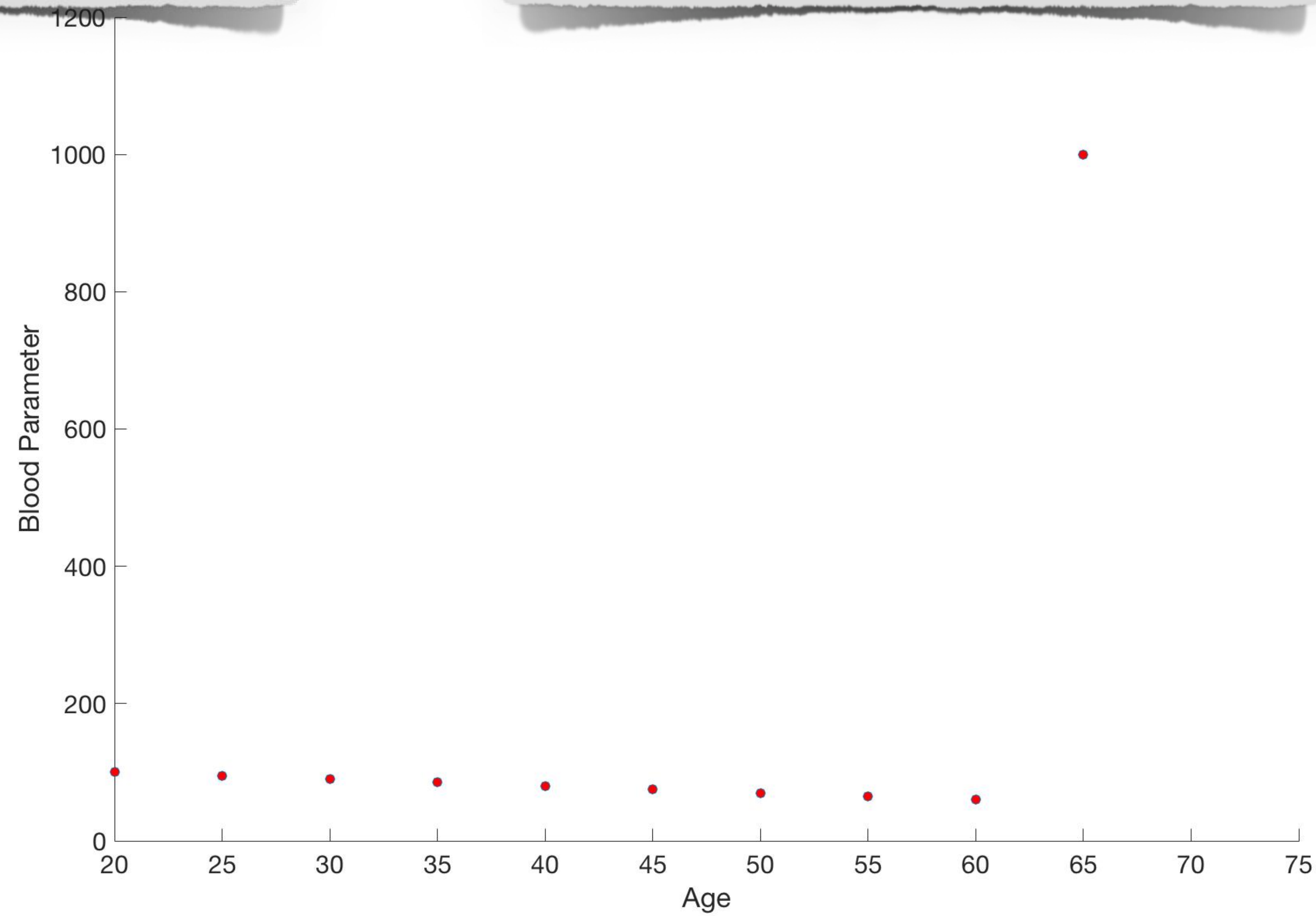
Pearson's *r* vs Spearman's *rho*

- Pearson's sensitive to outliers

Pearson's ***r*** = .48

Spearman's ***r*** = -.45

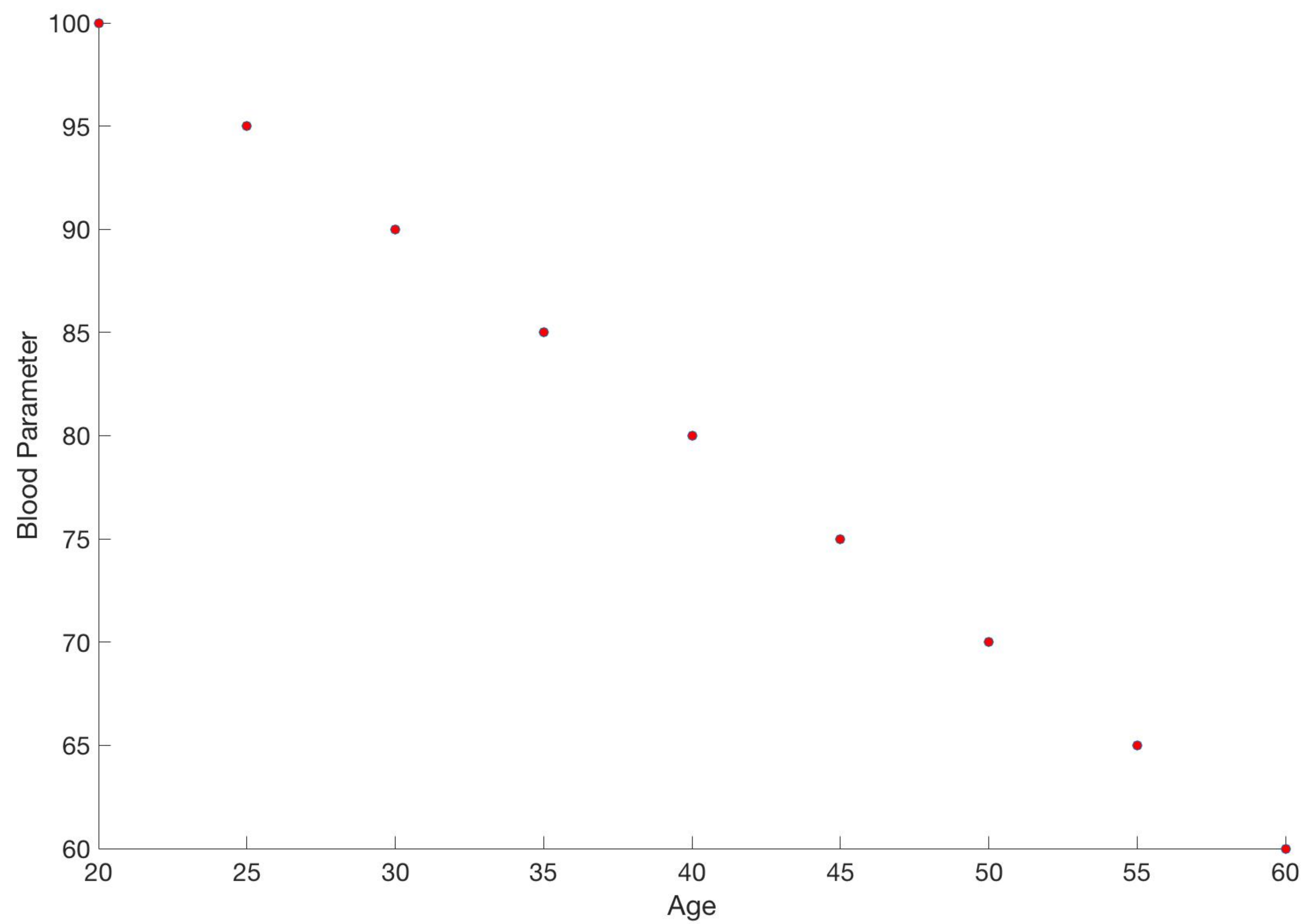
r = ?



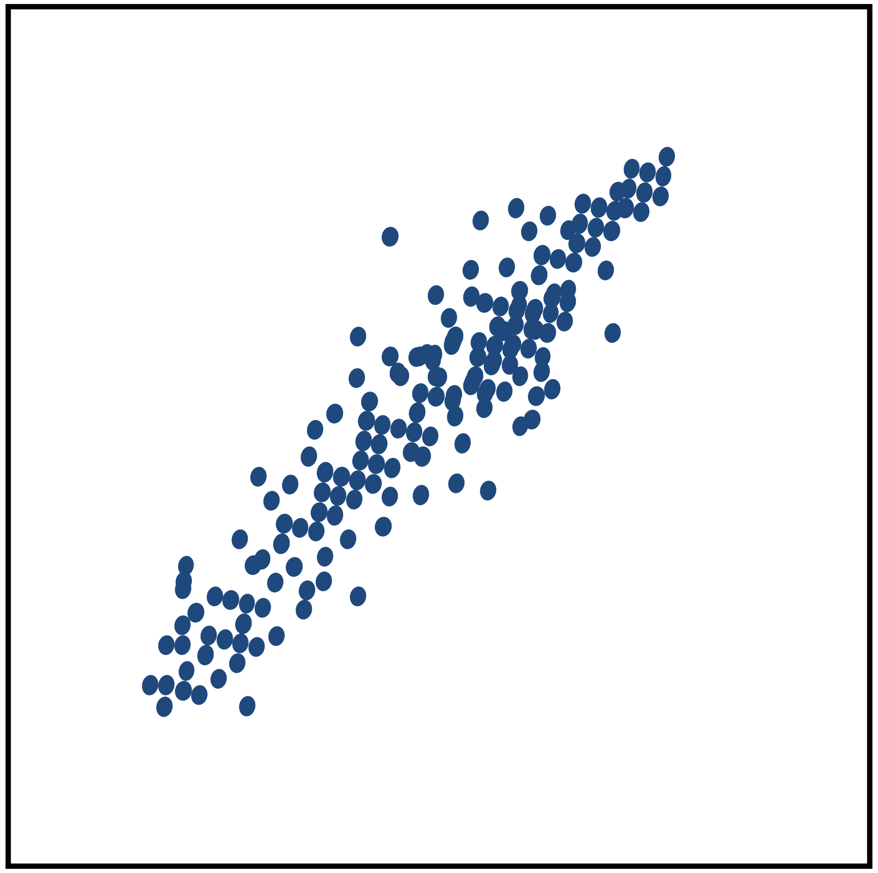
Pearson's *r* vs Spearman's *rho*

Pearson's *r* = -1

Spearman's *r* = -1

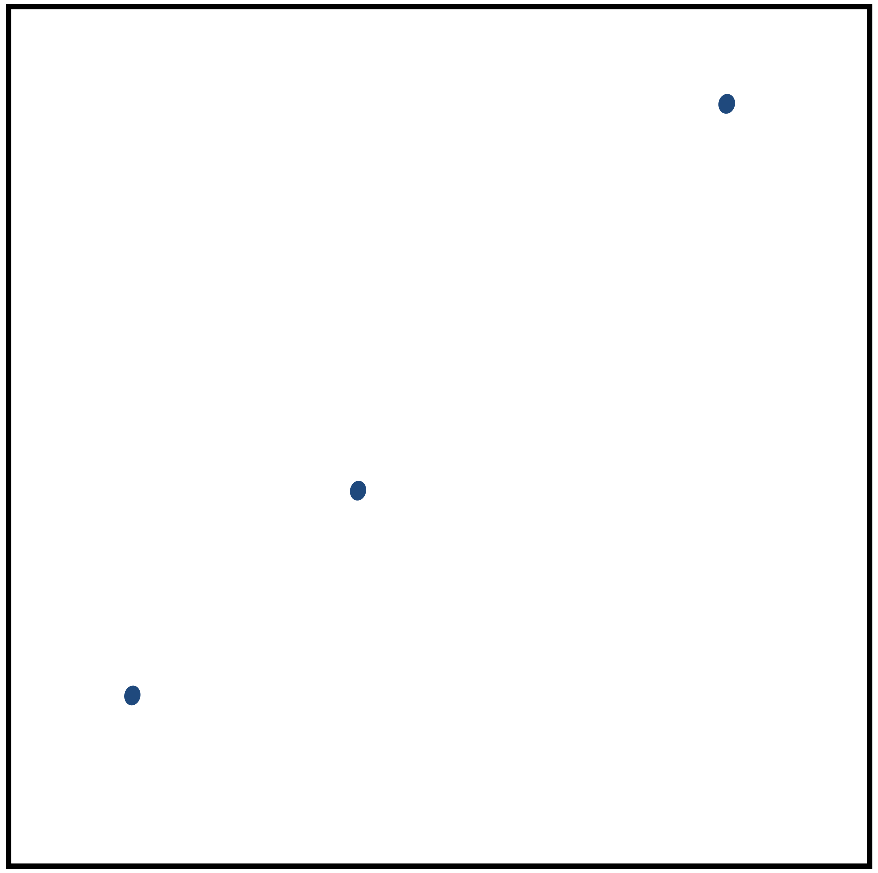


Significance of Correlation



r = 0.85

Is this significant?

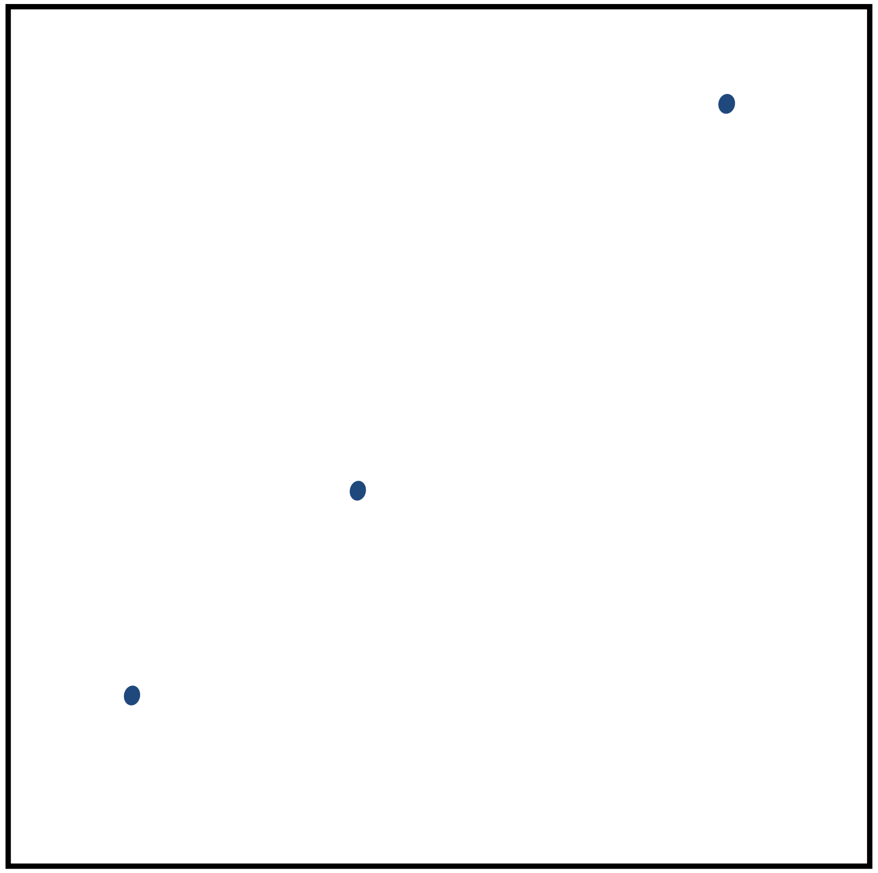


r = 0.99

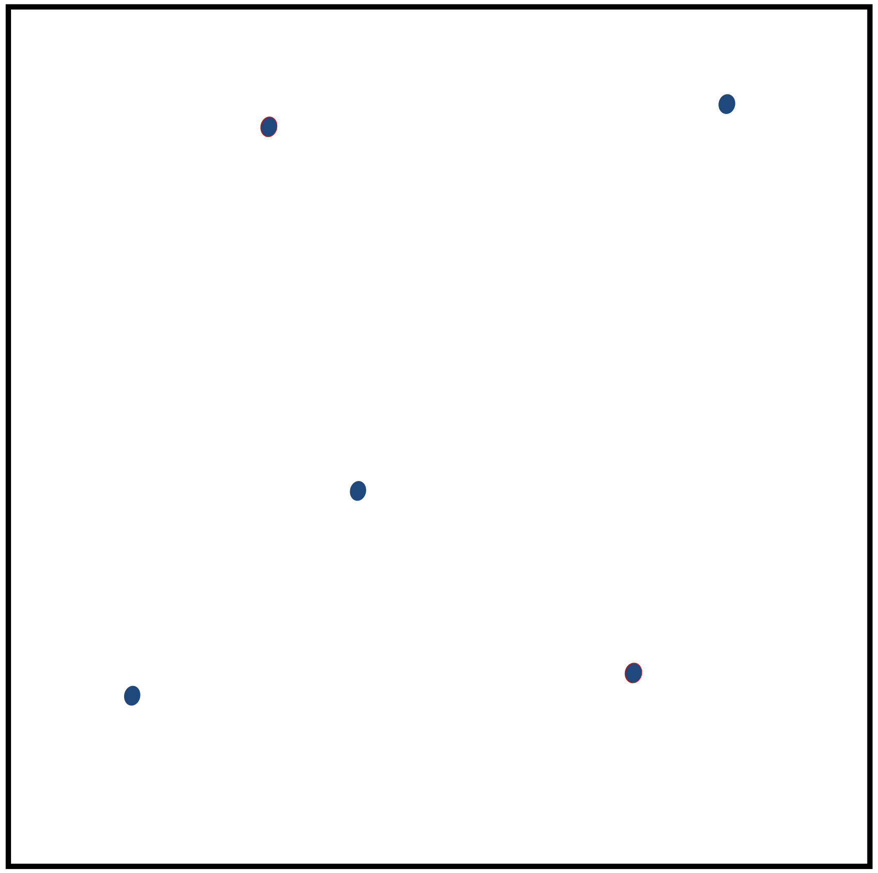
Is this significant?

Significance of Correlation

Add 2 more points to the plot



r = 0.99

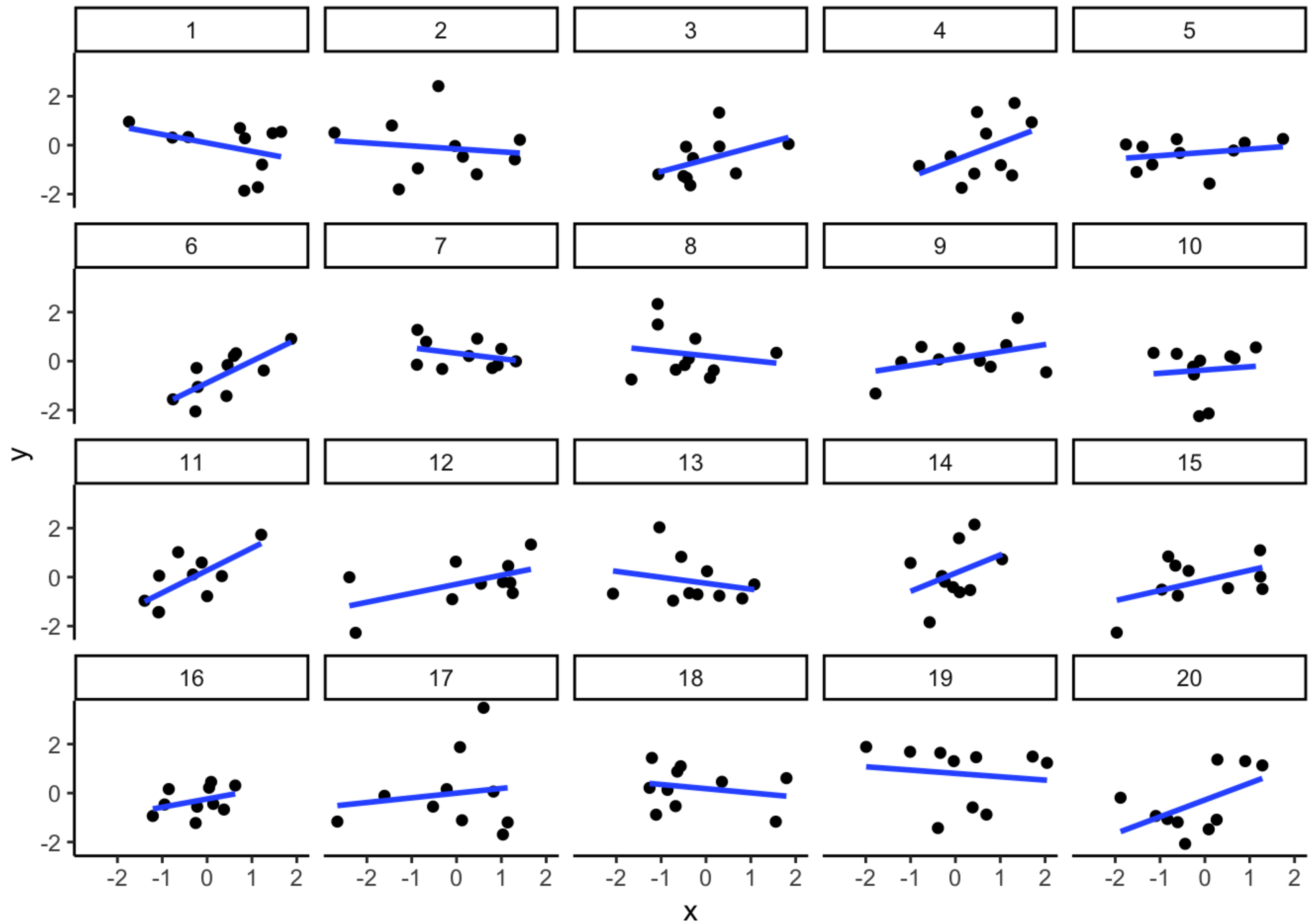


r = 0.05

Strength & Significance

- Strong relationship shown by correlation coefficient close to +/-1
  - apparently ‘strong’ relationships may not be statistically significant
  - e.g., sample size - when *n* is low, the odds are high that a ‘good’ correlation will occur by chance

Let's Simulate



Let's make fake data: 20 draws/iterations of random numbers for two variables
For each, sample size will be 10 and scatter plot them.

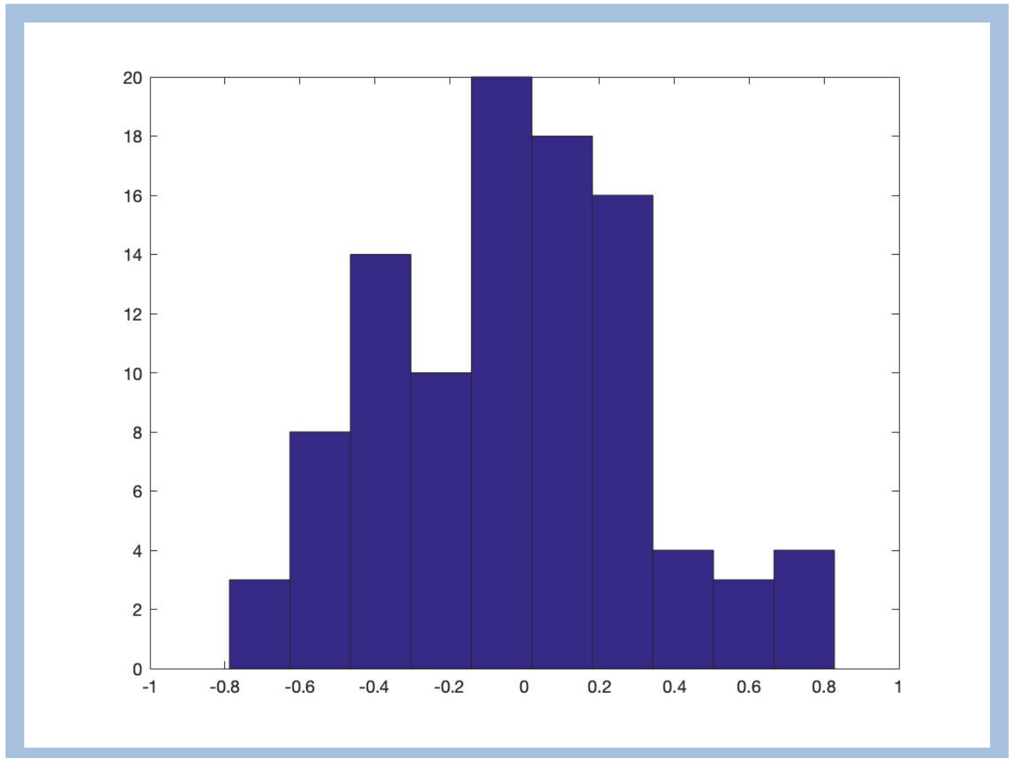
Let’s Simulate

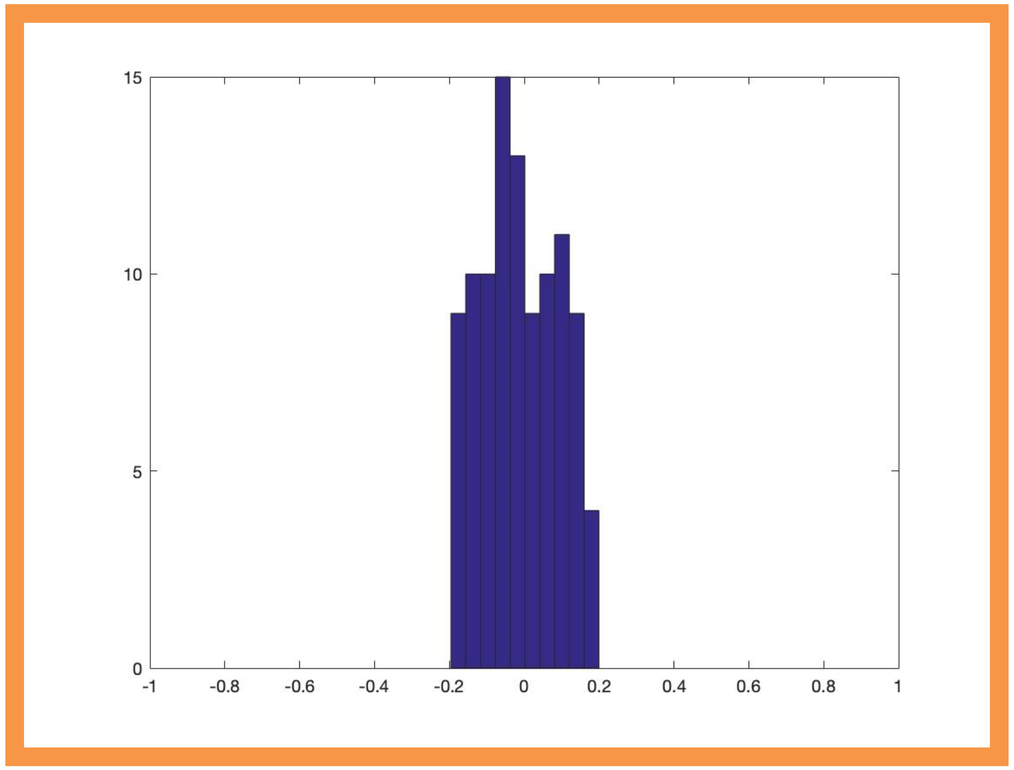
How would the distributions of *r* look like for the following:

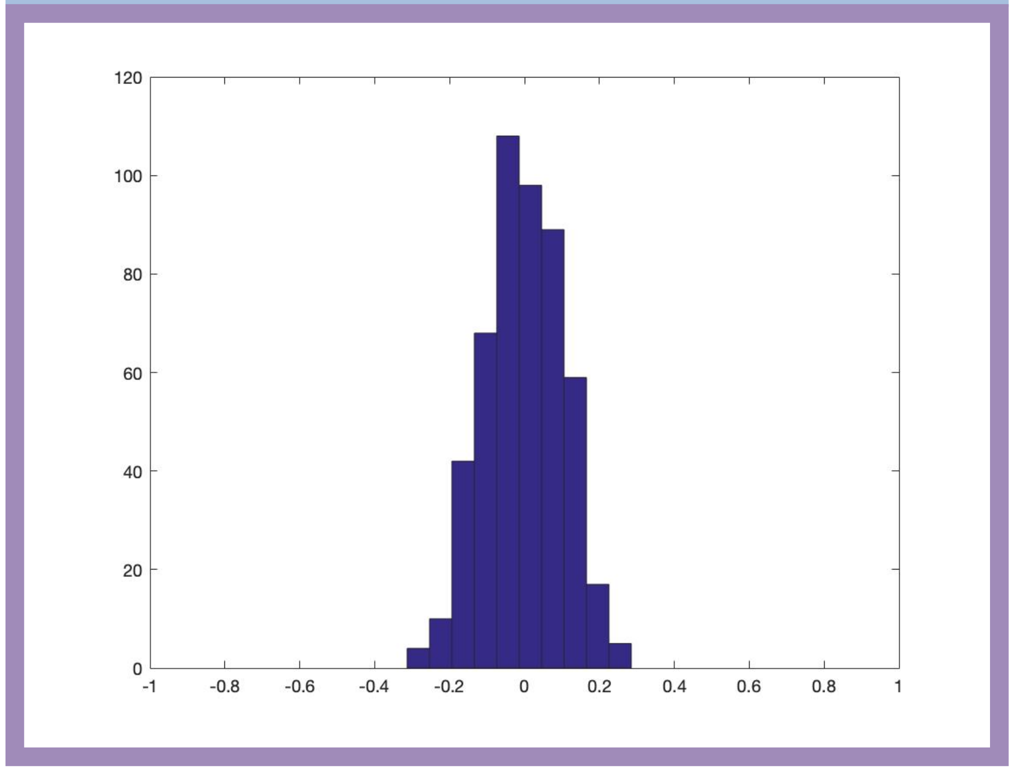
i) sample size = 10, iterations = 100

ii) sample size = 100, iterations = 100

iii) sample size = 100, iterations = 500





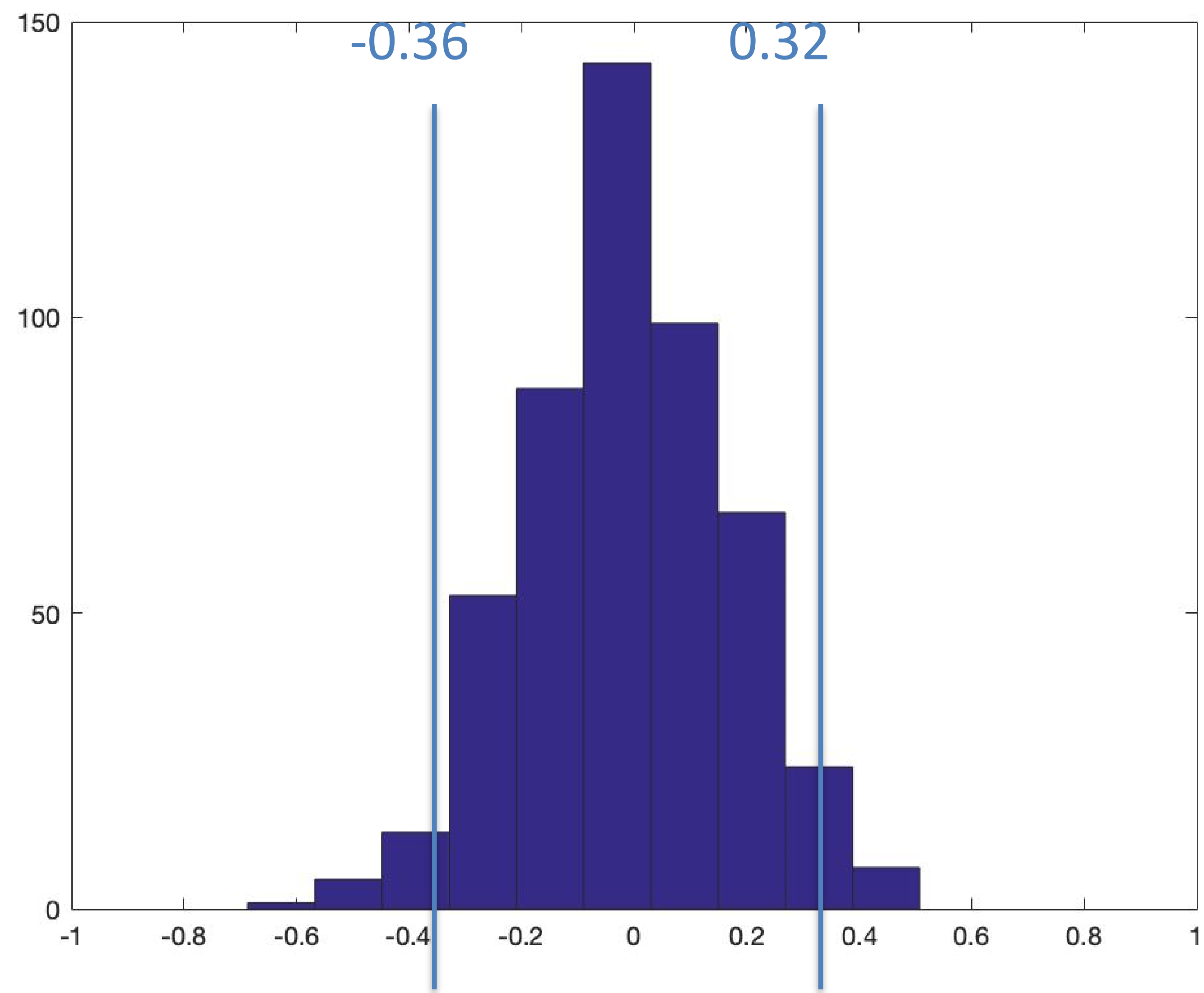


Let’s Simulate

What would the critical *r* values be for a sample size of 30?

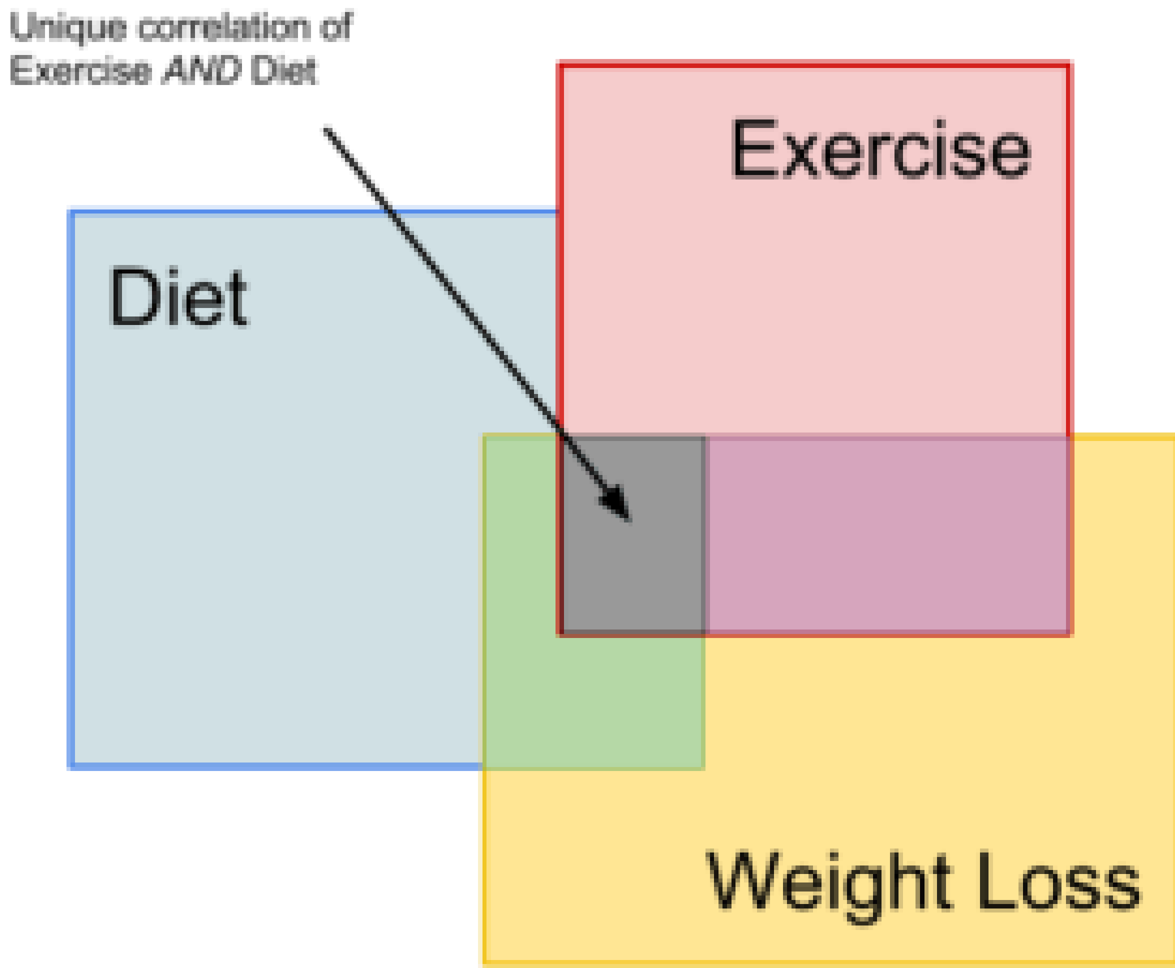
i) *n* = 30, iterations = 500

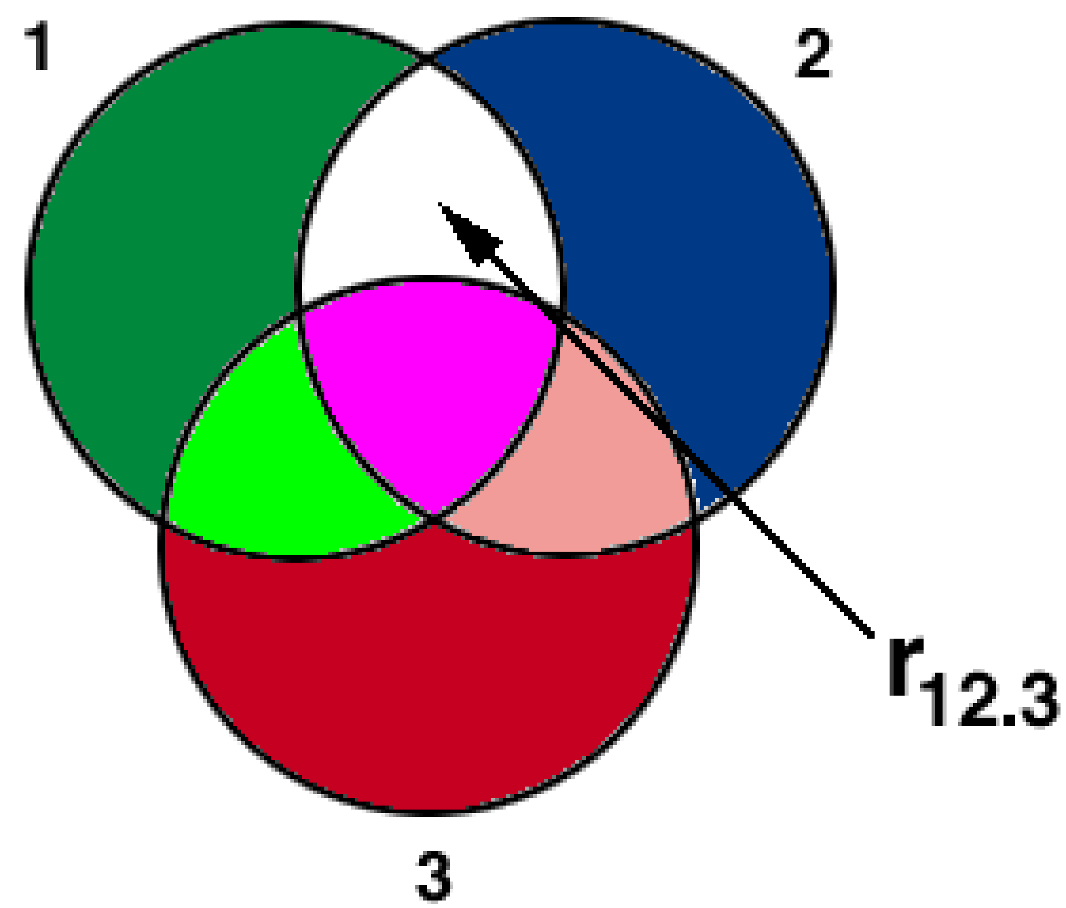
critical values (α < .05)



two-tailed
vs
one-tailed

Partial Correlation





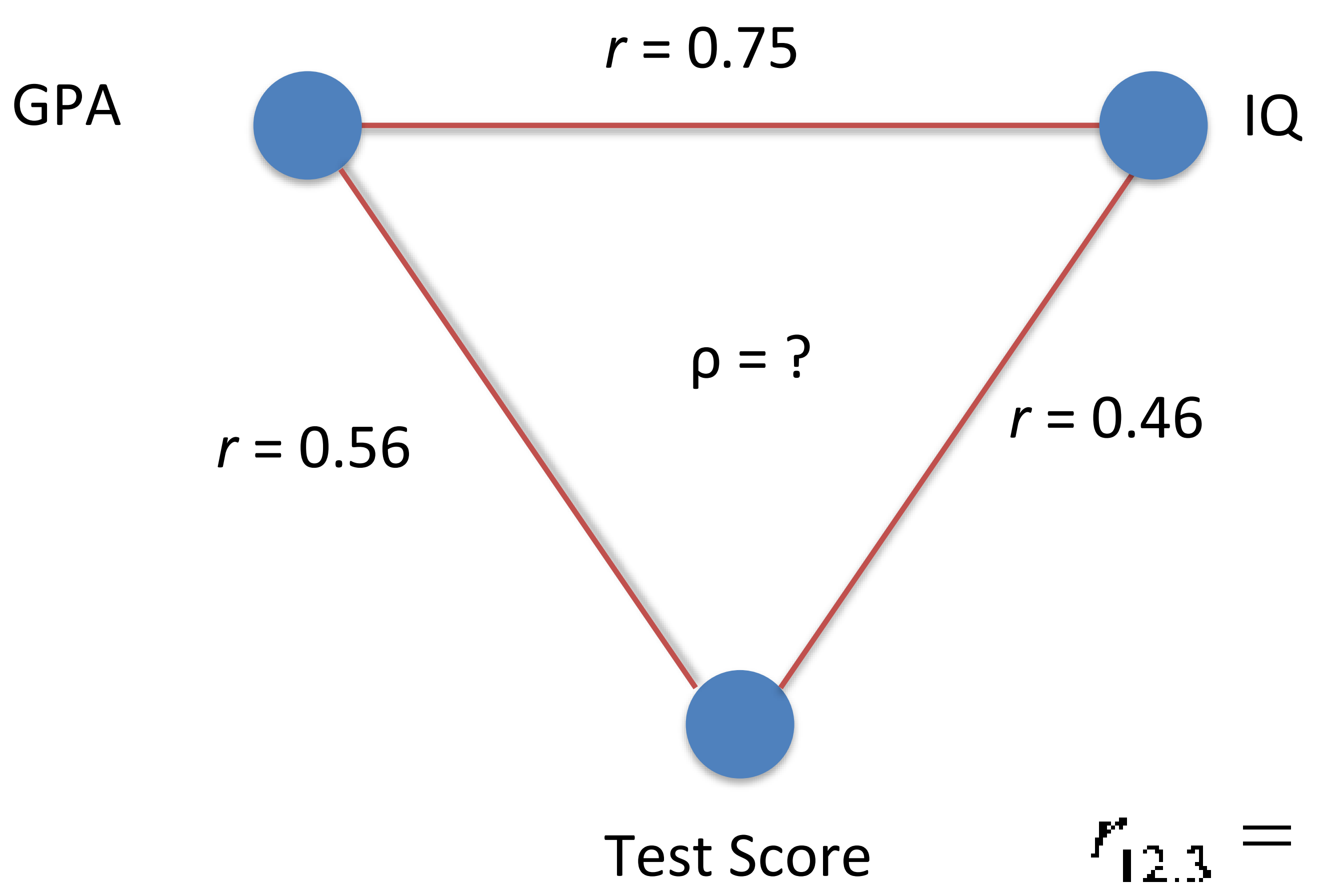
Partial Correlation

- measure of association between two variables, while controlling or adjusting the effect of one or more additional variables
  - What is the relationship between test scores and IQ scores after controlling for no. of hours of study?

Partial Correlation

- assumptions (Pearson)
  - all pairs of variables have a linear relationship
  - points are independent of each other
  - pairs of variables are bivariate normal (typically each variable is normally distributed)
  - non-parametric version for non-linear and or non-normal data

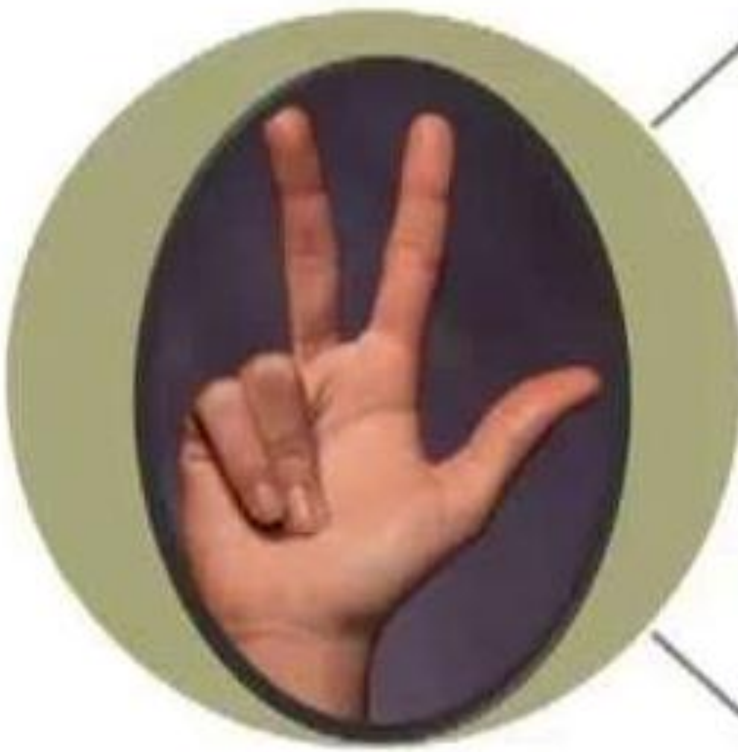
Activity/Assignment: Partial Correlation



$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Semi-Partial Correlation

- measure of association between two variables, while controlling or adjusting the effect of one or more additional variables **only on one of the two variables**
  - eg: you are interested in understanding the relationship between study time, tutoring, and exam scores while considering the potential confounding effect of study time on the relationship between tutoring and exam scores
  - how would you proceed?



Zero-Order Correlation

- This is the relationship between two variables.

Partial Correlations

- This is the relationship between two variables after removing the overlap of a third

Part (Semi-Partial) Correlations

- This is the relationship between two variables after removing a third variable from just the one variable.

$$r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^2}} \; and \; r_{2(1.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}}$$