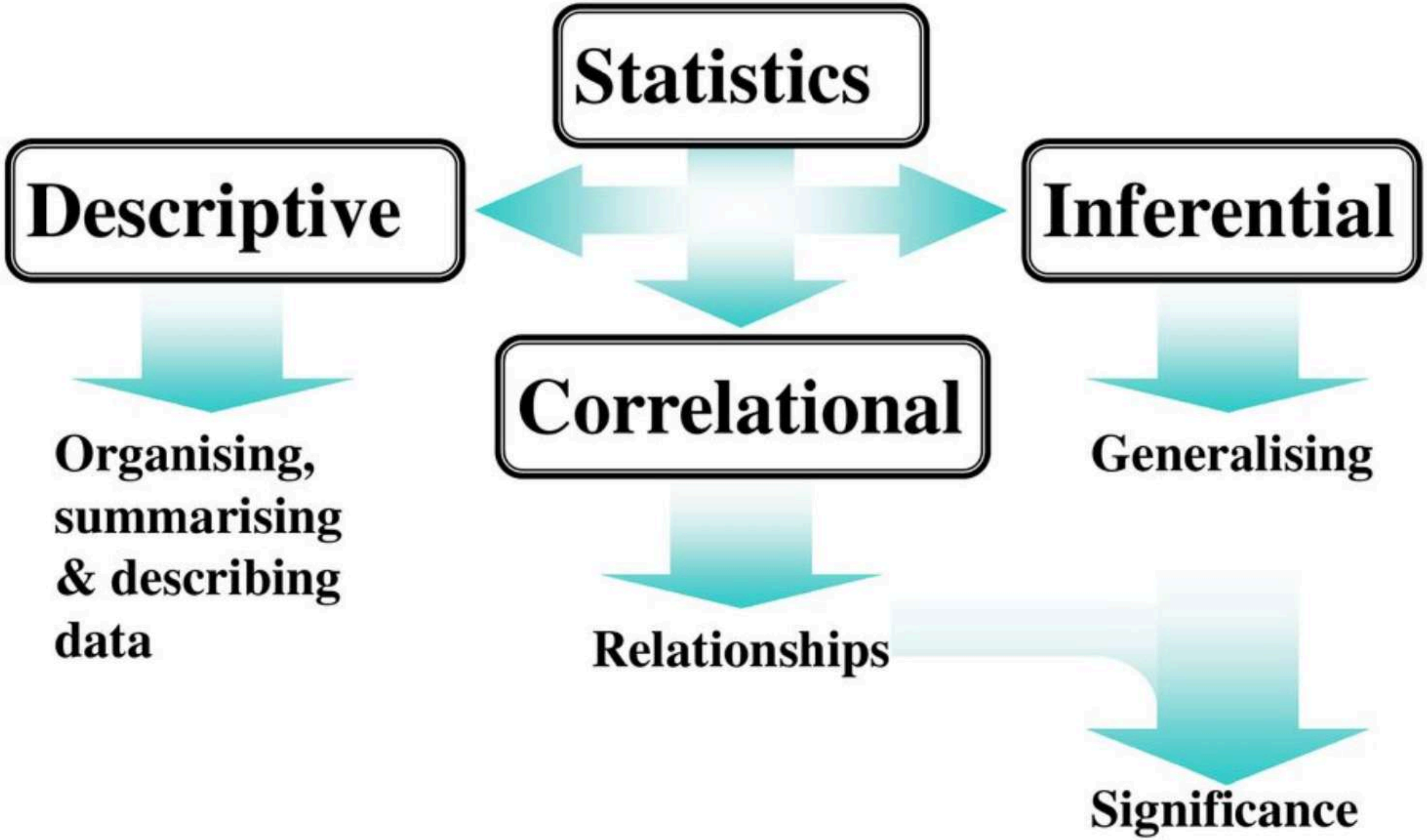
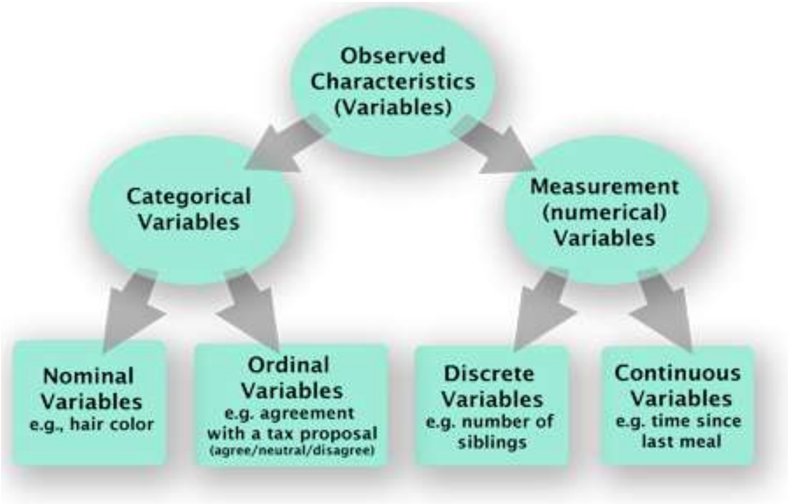
BRSM
Data Organization & Summarization

Vinoo Alluri



How do you start?

Data Organization



- identify variables (IV, DV) and respective types
- identify different levels of measurement
- missing data?
  - replace with mean
  - remove

20-25 years = 1
26-30 years = 2
31-35 years = 3
36-40 years = 4
41-45 years = 5
46 years and older= 6

Continuous → Categorical

Table format: XY		X	A		
		minutes	Test group A		
	x	X	A:Y1	A:Y2	A:Y3
1	Title	0	0.0	0.0	0.0
2	Title	2			
3	Title	4			
4	Title	6			
5	Title	8			
6	Title	10			
7	Title	12			
8	Title	14			
9	Title	16			
10	Title	18			
11	Title	20			

3	1	5.611248	4.1174493	5.4884806
4	2	5.5560017	3.9532921	5.2730727
5	3	4.5405	4.5603814	4.4634323
6	4	5.236287	3.8760467	5.198486
7	5	5.9417286	3.398312	5.965598
8	6	5.4199543	4.0421543	5.181767
9	7	4.4019384	3.394504	4.5349746
10	8	5.1843286	4.168893	4.9799395
11	9	5.3209386	3.9951186	5.300459
12	10	5.1961555	3.8243186	5.080454
13	11	5.5065527	3.938081	5.2821956
14	12	5.118871	3.8536696	5.1487226
15	13	5.4678555	3.9871855	5.308297
16	14	5.261652	3.4055495	5.6112976
17	15	5.9904175	4.116685	5.461459
18	16	3.838822	4.4964914	4.35598
19	17	5.68176	3.9998796	5.340737
20	18	4.433616	4.4853745	4.5518494
21	19	5.4475813	3.1434624	6.075916
22	20	5.3806806	3.8687606	5.3088202
23	21	5.417145	4.1244016	5.288509
24	22	5.8884277	4.254202	5.5335727

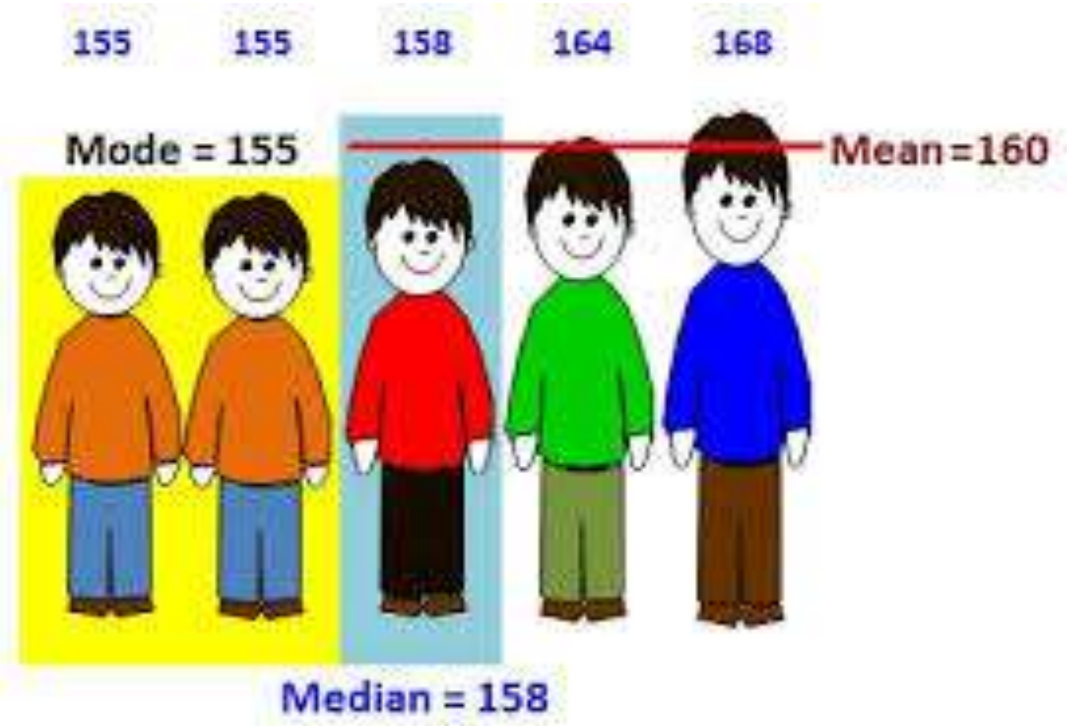
Summarize

to tell, in your own words, what has happened in the story

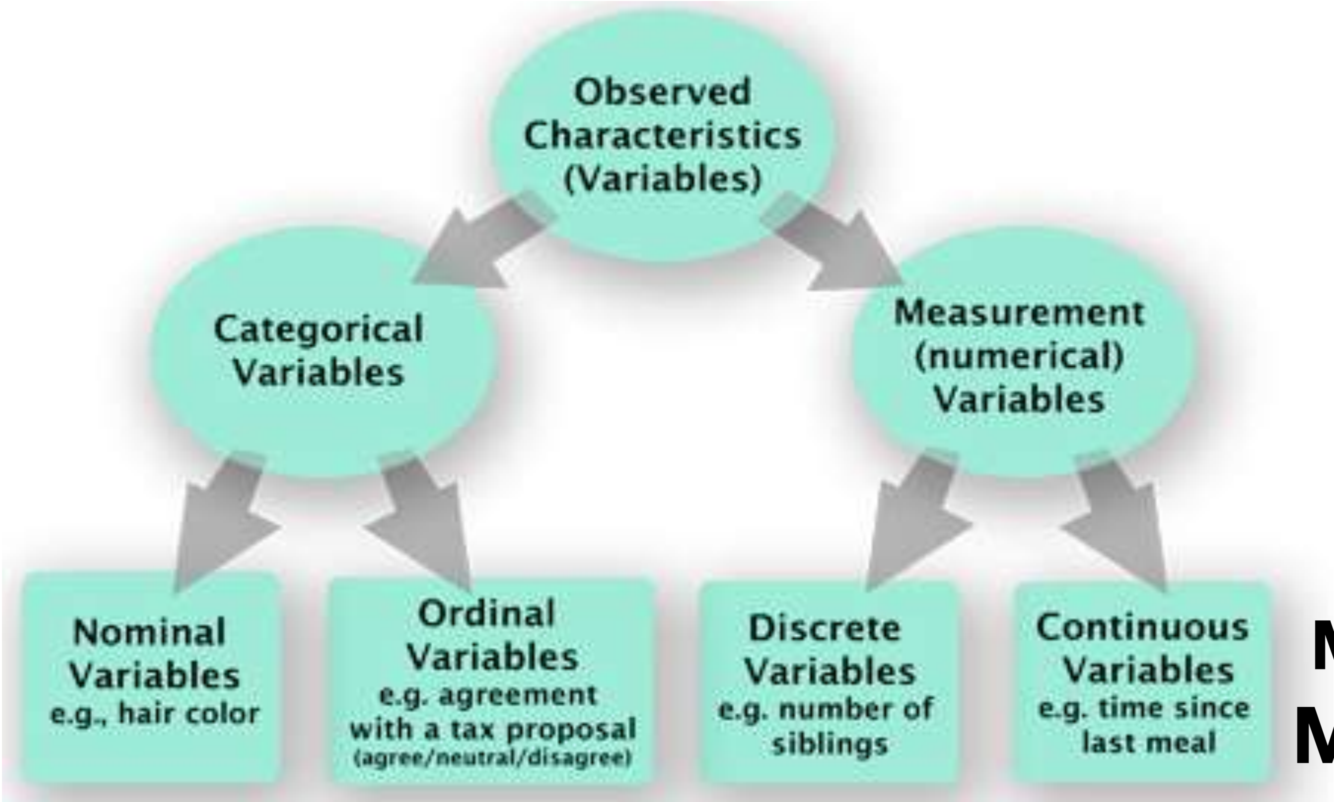
Descriptive Statistics

- 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



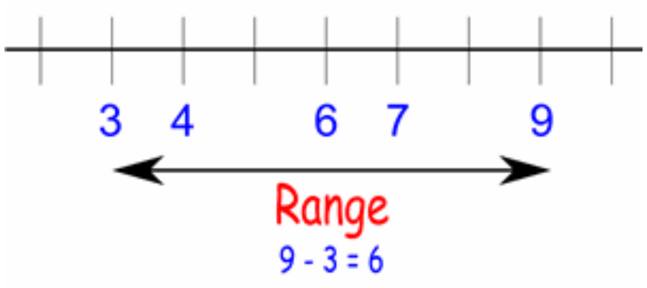
MODE
Median

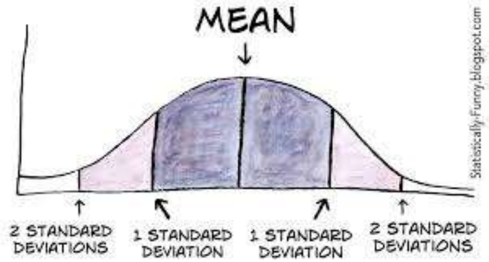
MODE **Median** **Mean** **Mean**

Measures of central tendency

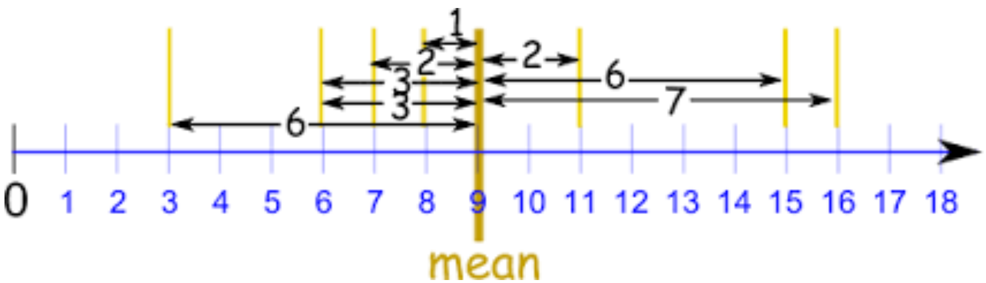
	Advantages	Disadvantages
Mean	A sensitive and exact measure of the centre point of a group of values	A single extreme value in one direction can seriously distort the mean
Median	Not as susceptible to extreme values as the mean	Can be unrepresentative if there are only a small number of values
MODE	Indicates the most important value Unaffected by extreme scores More informative than mean	Not useful for small data sets where several values occur equally frequently

Measures of dispersion/spread



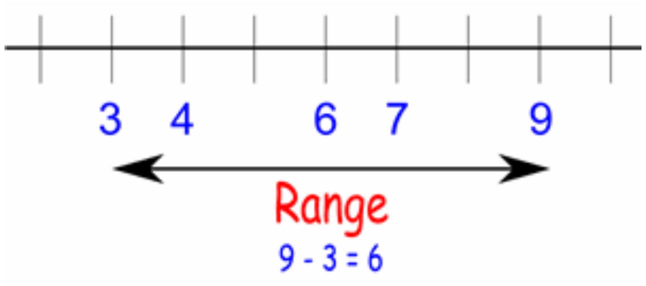
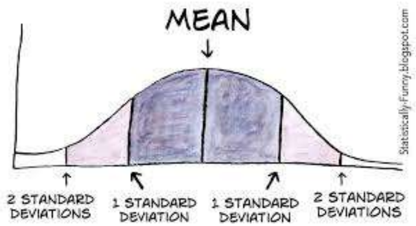


$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

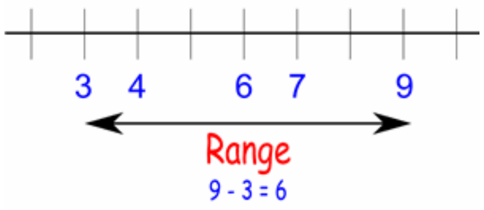


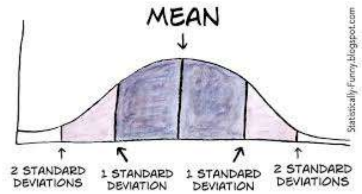
$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

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	——

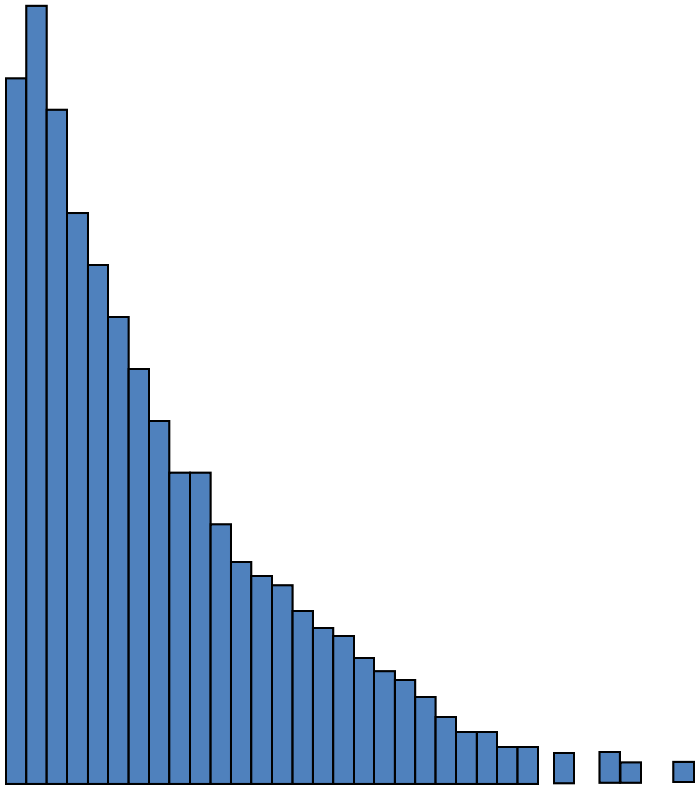
MEAN ? MODE MEDIAN





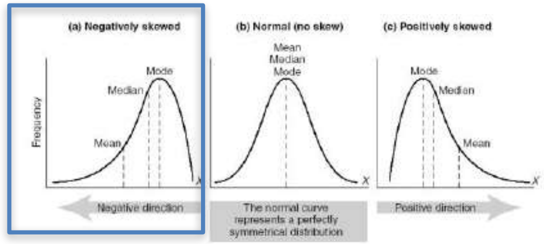
When do these measures fail to be representative ????

Skewed Distribution

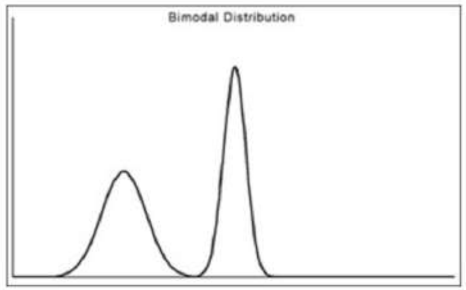


- Resembles an exponential distribution
- Lots of extreme values far from mean or mode
- Not straightforward to do useful statistical tests with this type of distribution

Skewed Distribution

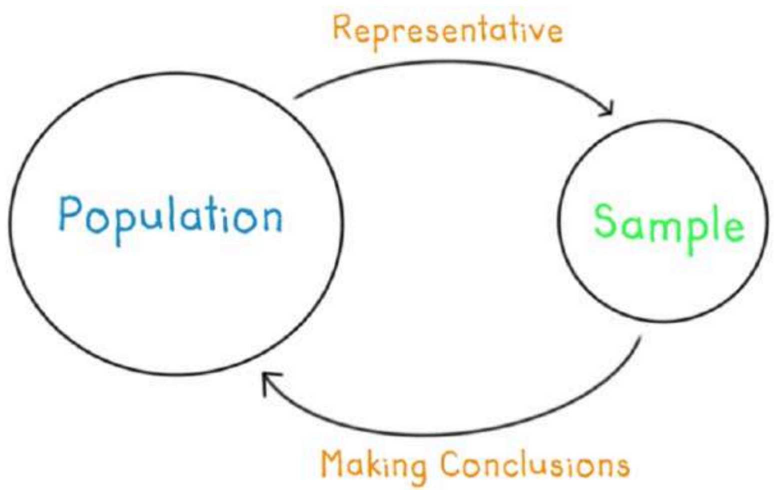


- **Negative skew**
  - Result from relatively easy tasks, due to a ceiling effect
- **Positive skew**
  - Results from tasks which are hard to improve upon, due to a floor effect (such as RT —reaction time)
- **Bimodal**
  - Two distinct peaks
  - probable indicator of groups
  - ex: completion time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

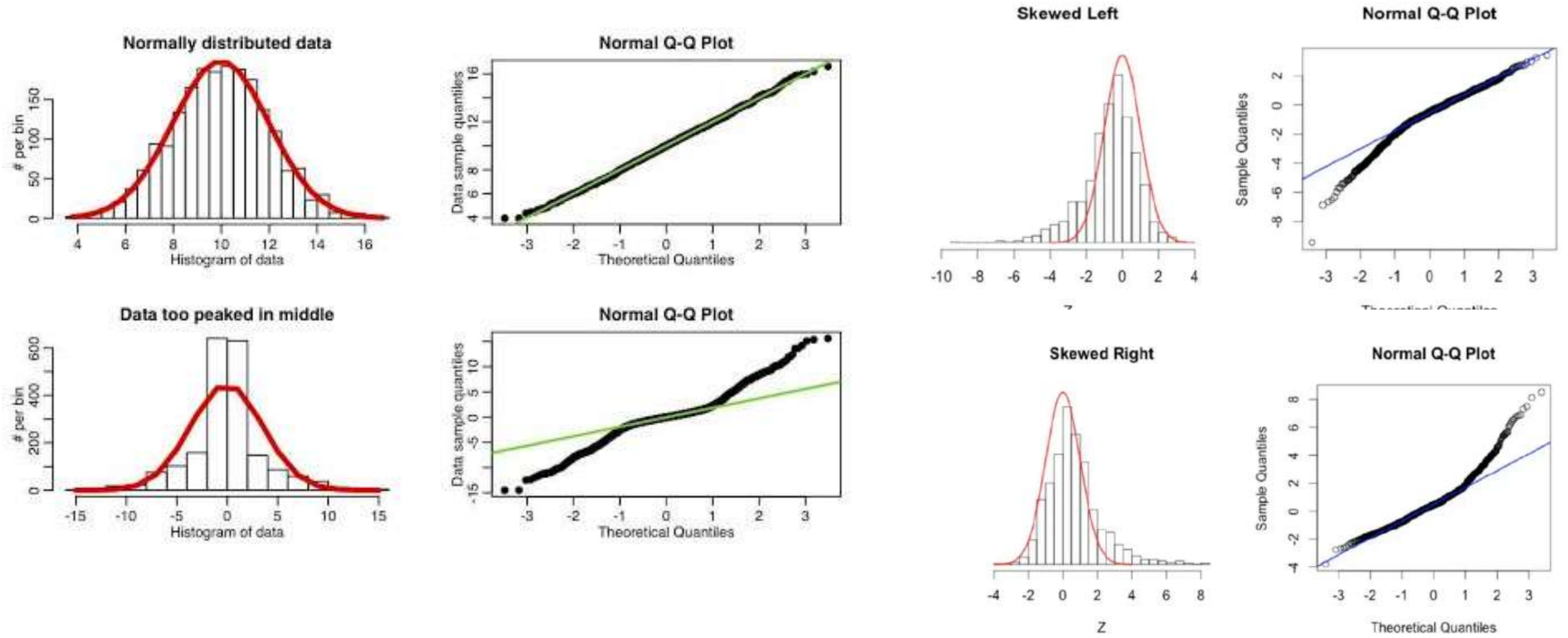


Normality in Real-World Data

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



Testing Normality

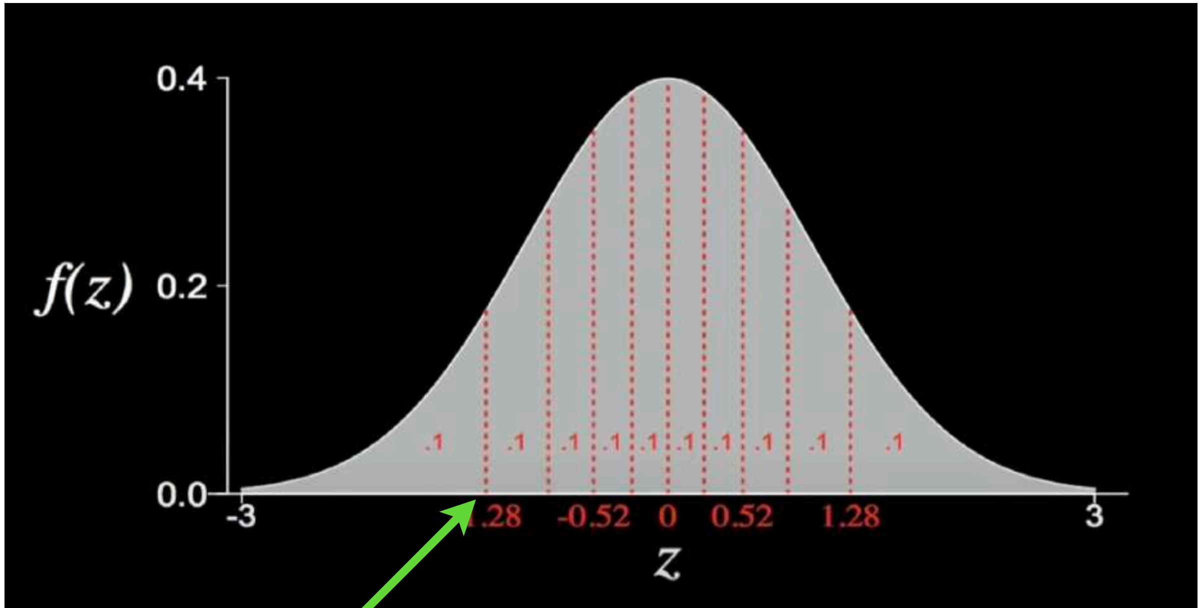


- Q-Q plot: graphical technique (can also use it to test any theoretical distribution)
- theoretical quantiles plotted on x-axis and sample quantiles plotted on y-axis

Example

- does this come from a normally distributed population?

3.89 4.75 6.33 4.75 7.21 5.78 5.80 5.20 6.64

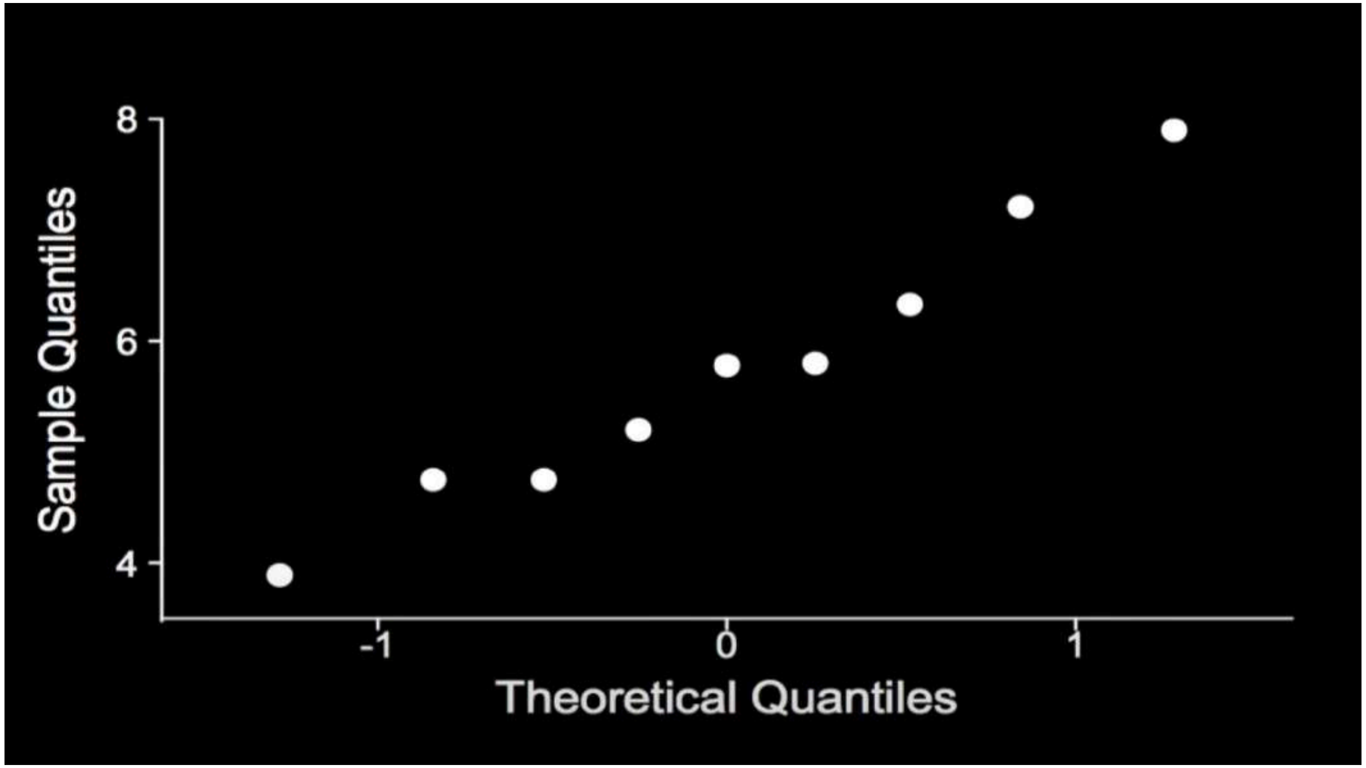


0.1th quantile or 10th percentile

Example

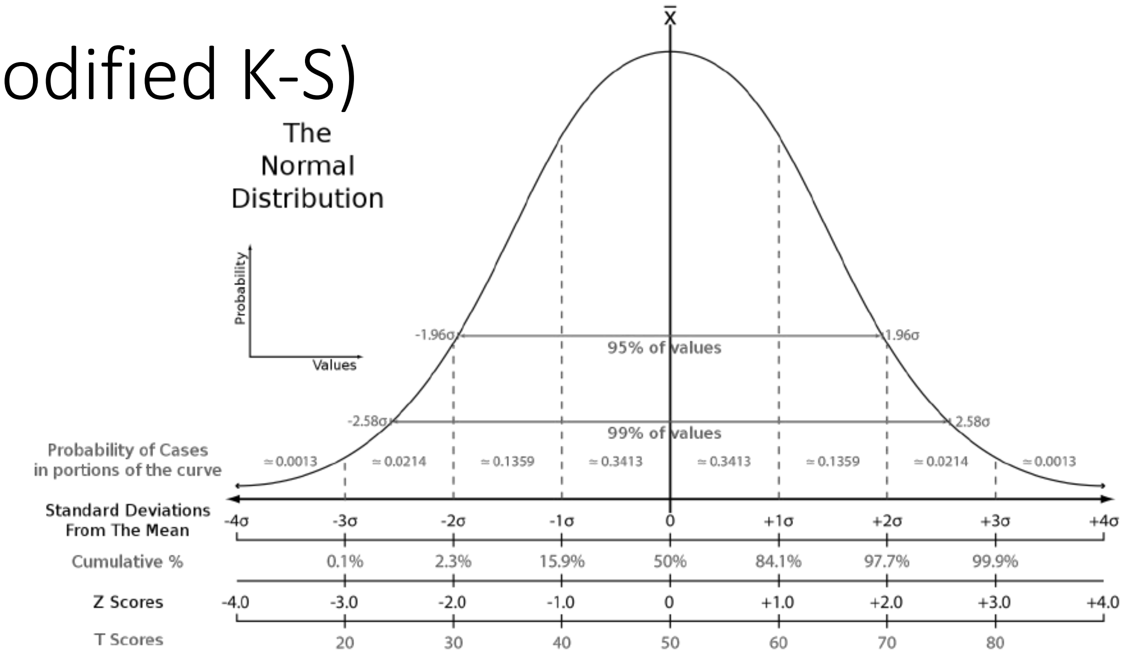
- does this come from a normally distributed population?

3.89 4.75 6.33 4.75 7.21 5.78 5.80 5.20 6.64



Testing Normality

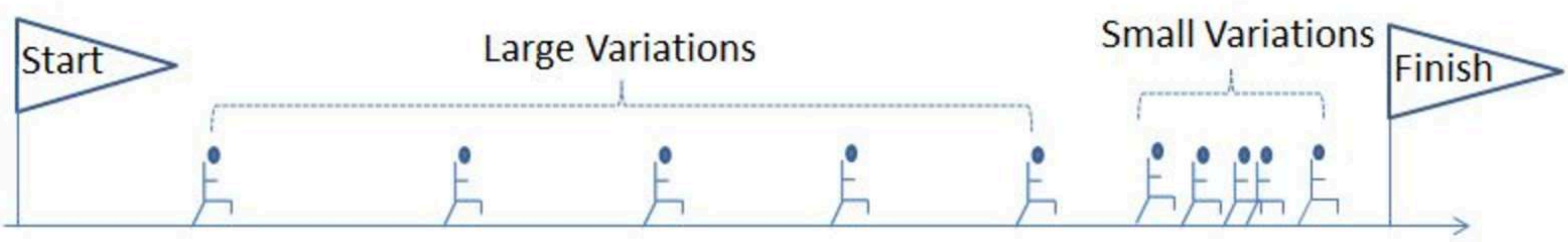
- Tests to assess normality (null hypothesis: data are sampled from a population that follows a normal distribution)
  - Kolmogorov-Smirnov (≥ 50)
  - Shapiro-Wilk (for smaller sample size, i.e. < 50)
  - Anderson-Darling (modified K-S)
  - Lilliefors test
  - Cramer-von Mises
  - etc..

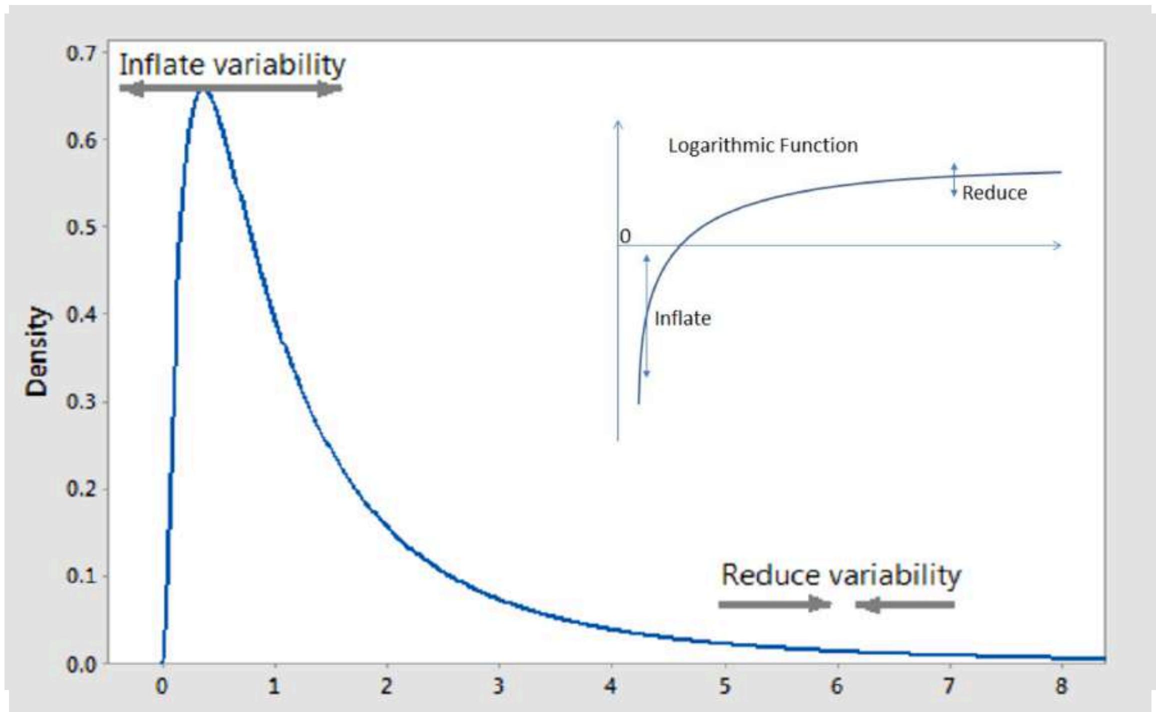


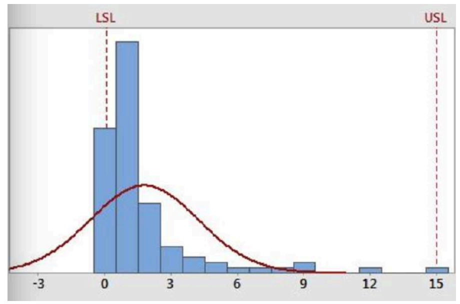
Testing Normality

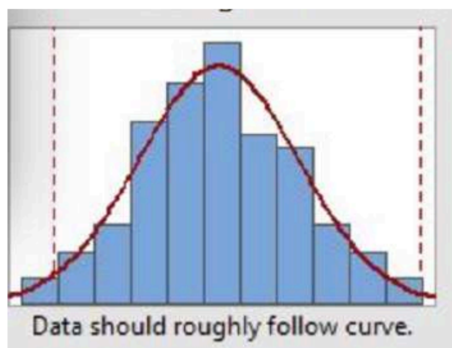
- For non-normal data
  - transform to normal distribution (eg: sqrt, log)
    - if it works - use parametric tests
    - if still not normal - use non-parametric tests
  - if you have groups of data, you **MUST** test each group for normality.

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

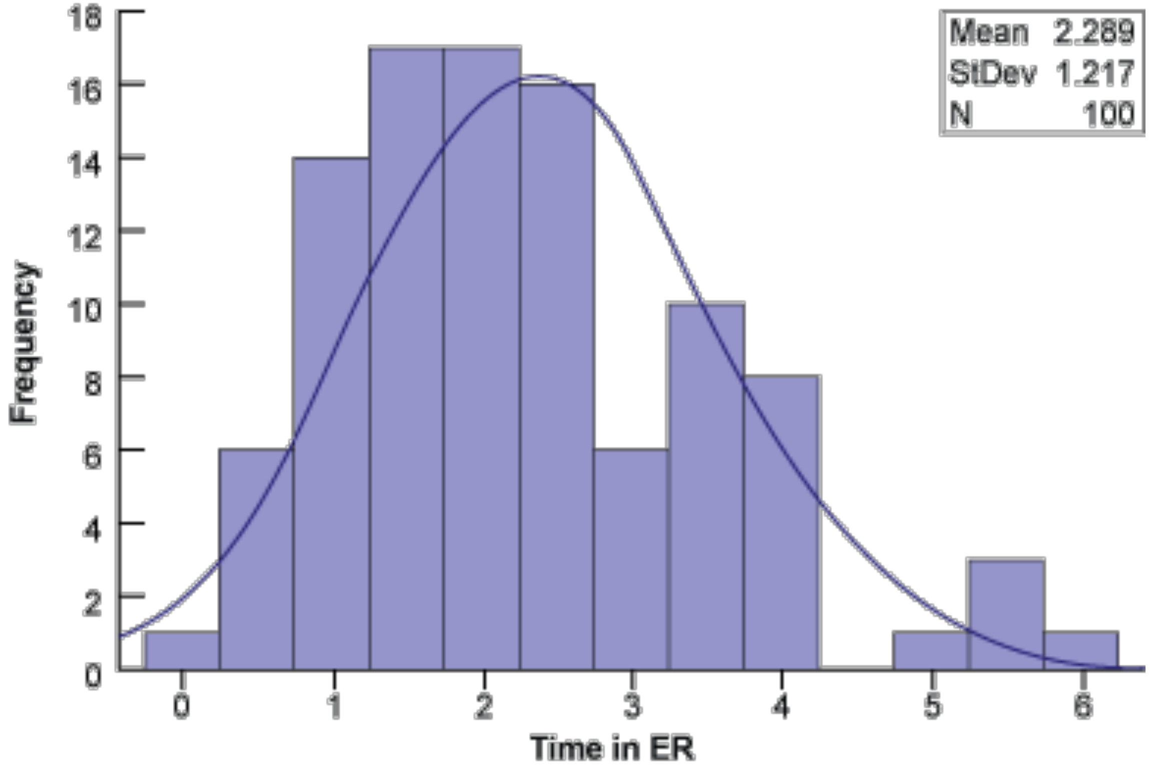
Moderately positive skewness	sqrt(X)
Substantially positive skewness	$\log_{10}X$
Substantially positive skewness (with zero values)	$\log_{10}(X + C)$
Moderately negative skewness	sqrt(K-X)
Substantially negative skewness	$\log_{10}(K\text{-}X)$

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

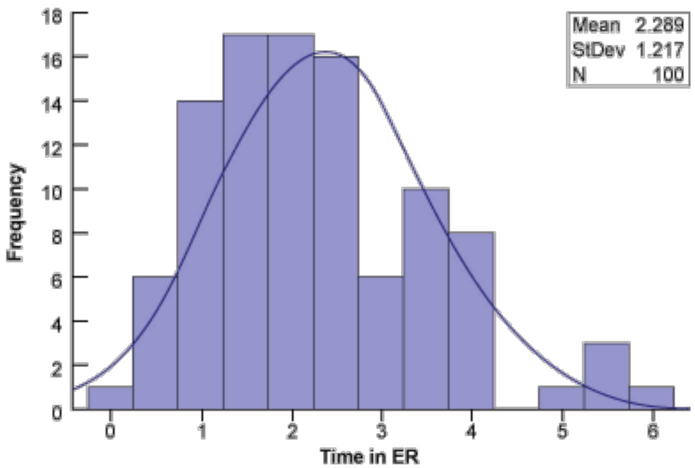


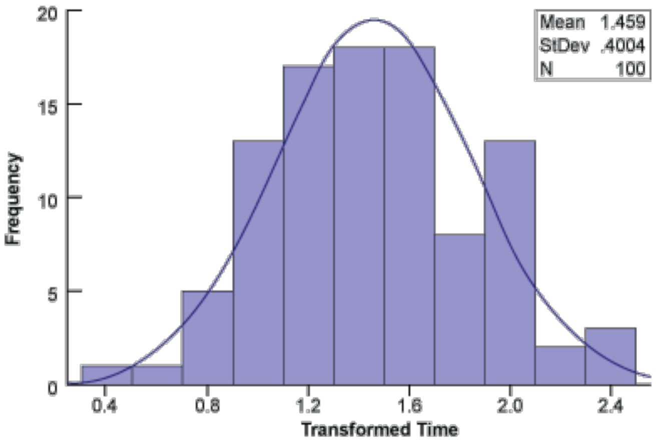
typically it is four hours or less

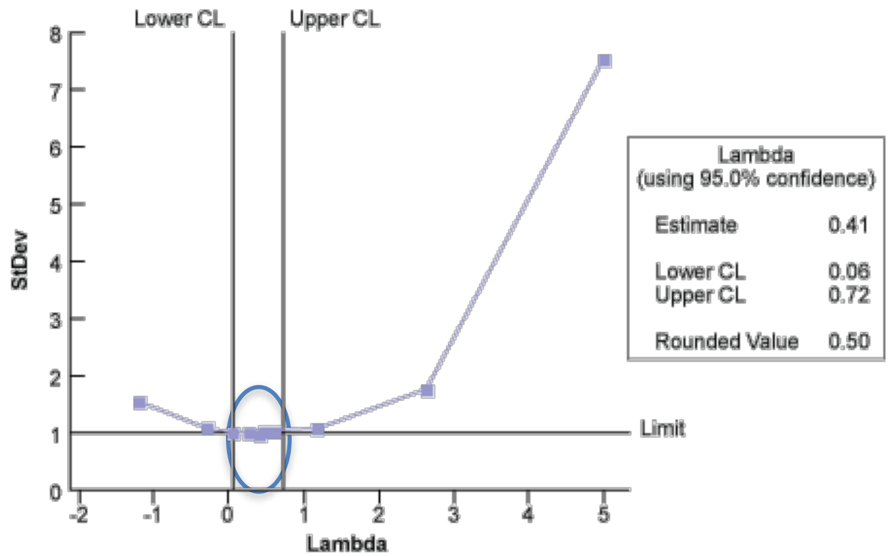
EXAMPLE

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

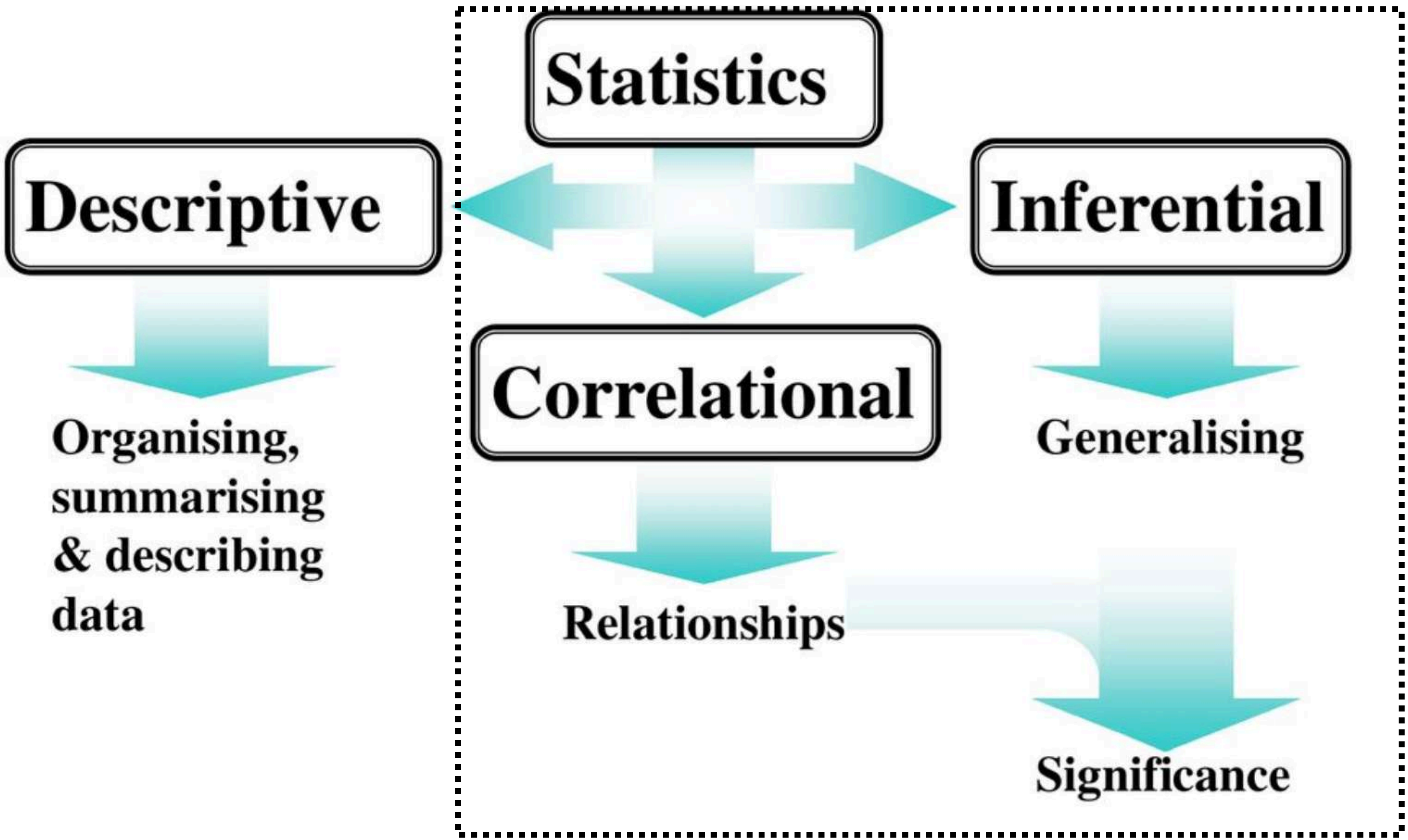
the "optimal value" is the one which results in the best approximation of a normal distribution curve



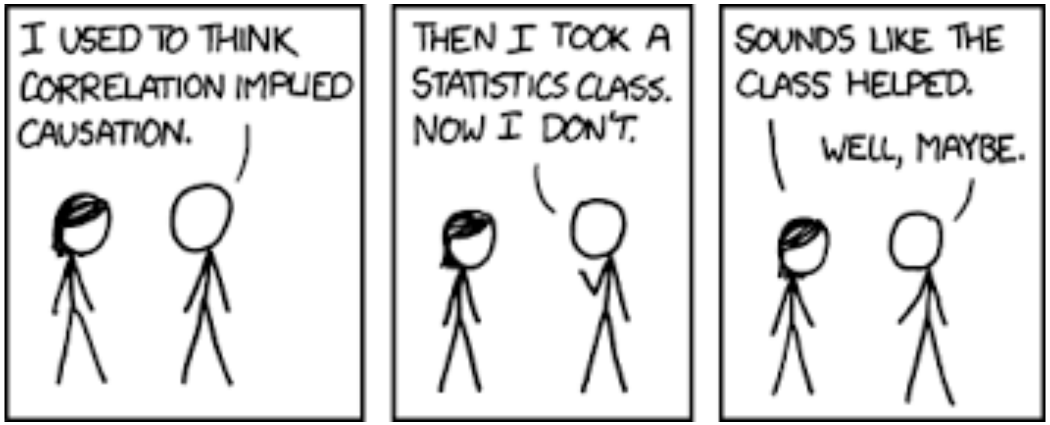




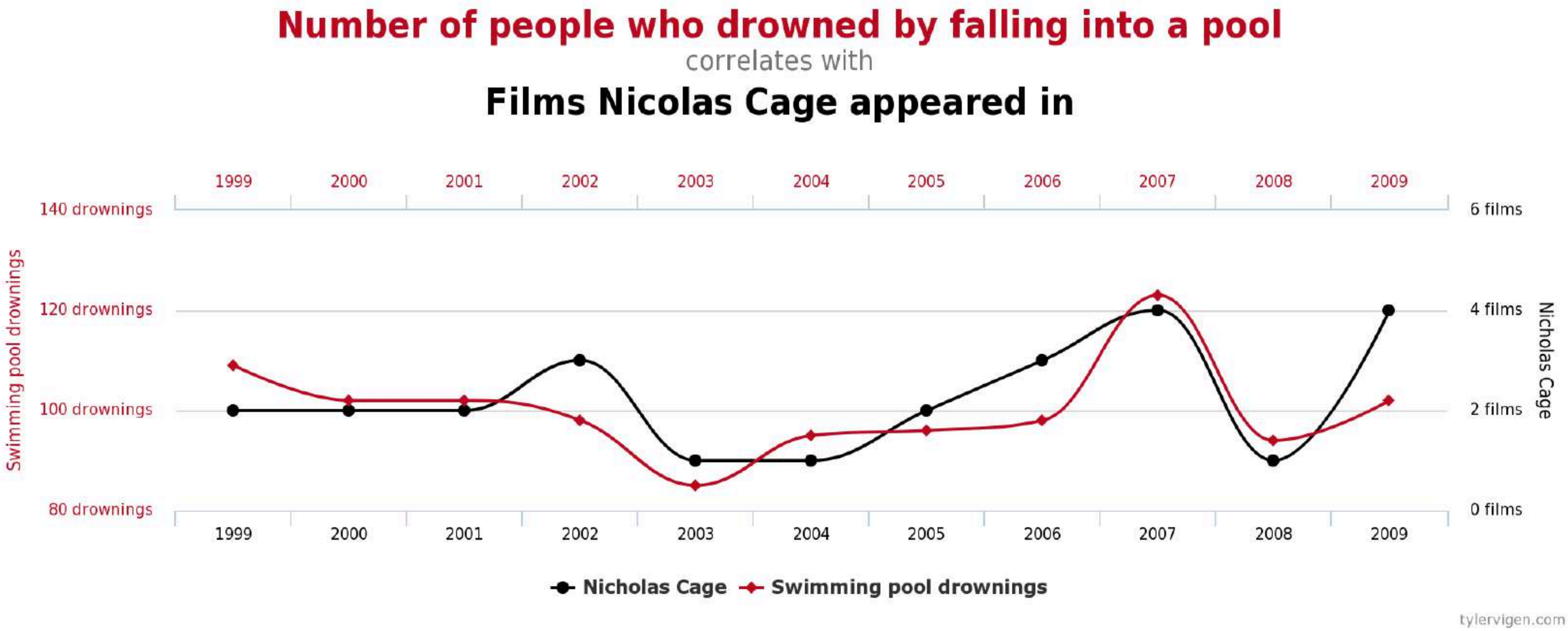
$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ (y_t^\lambda - 1)/\lambda & \text{otherwise.} \end{cases}$$


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

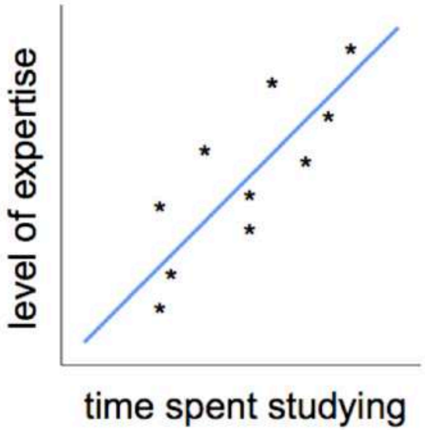
Correlation

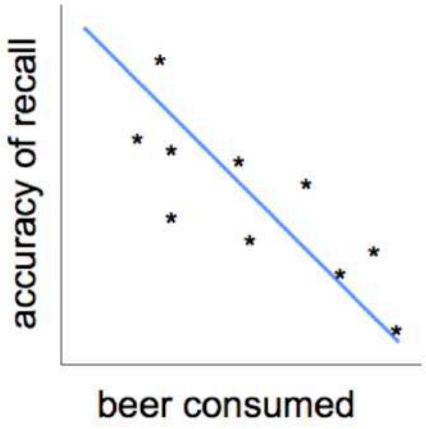


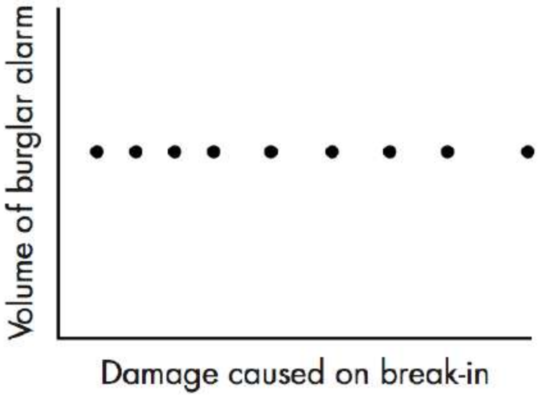
Not Causality

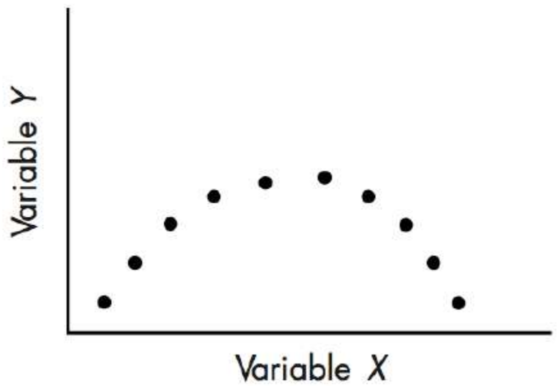


Correlation

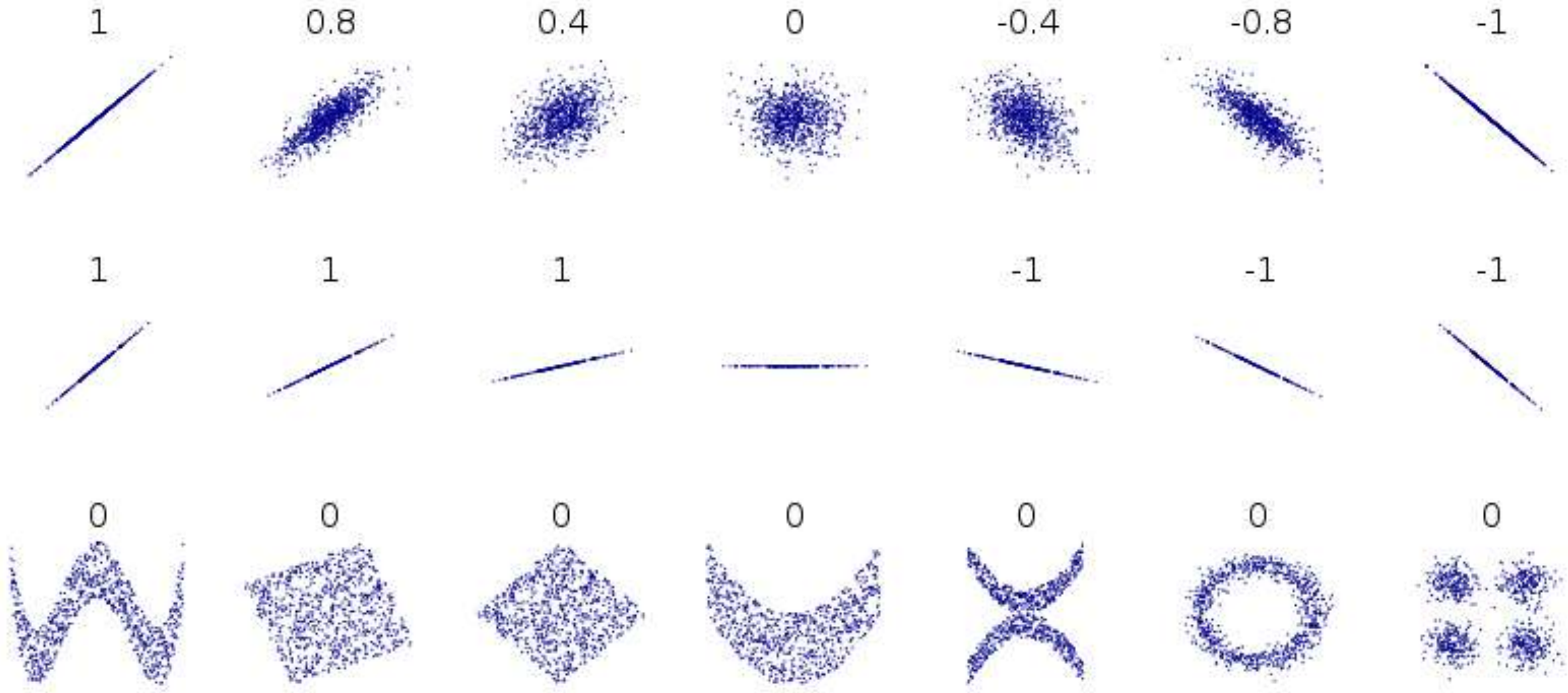








Pearson's r



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

Correlation

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

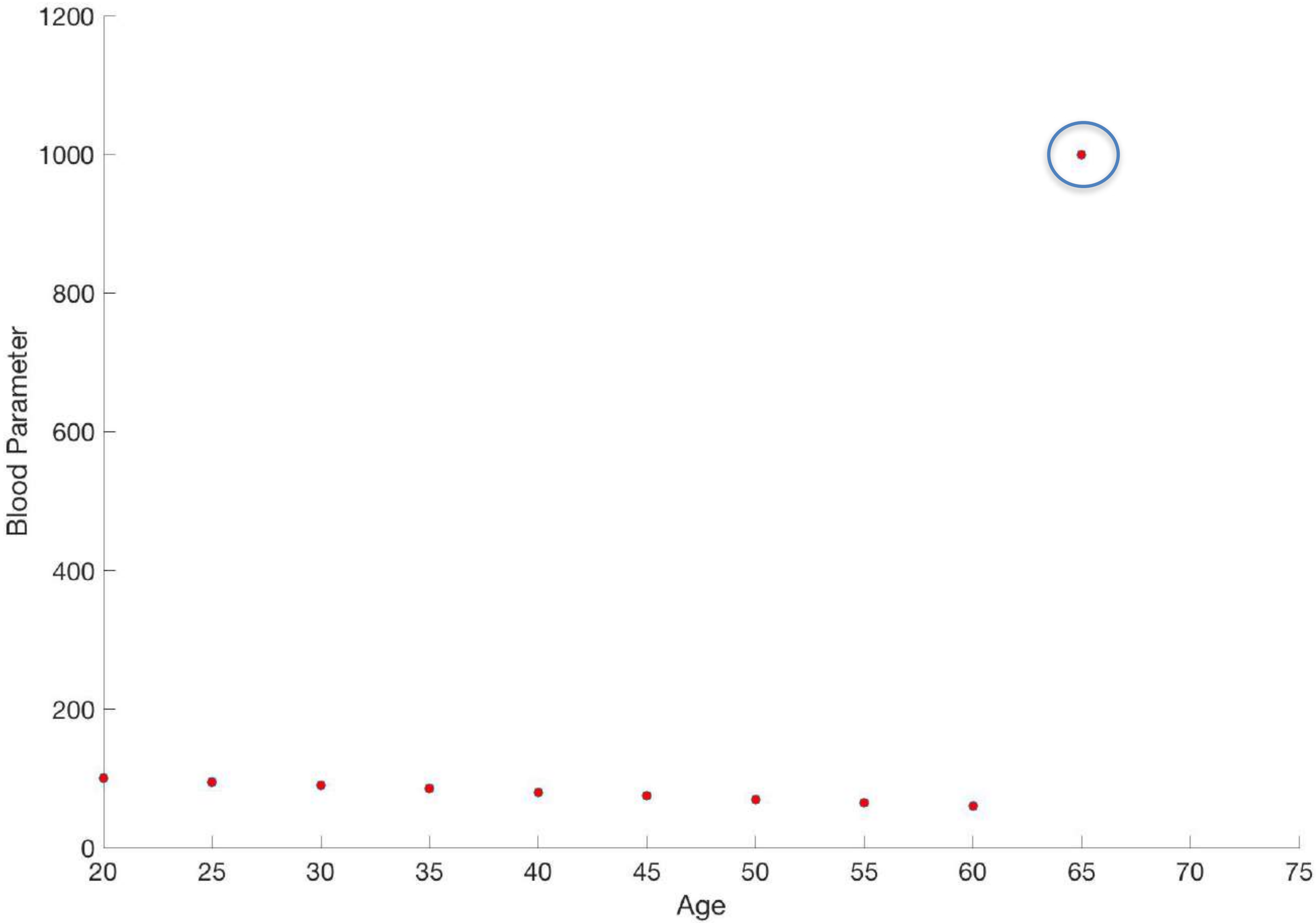
Variable Y\X	Quantitiative X	Ordinal X	Nominal X
Quantitative Y	Pearson r	Biserial r_b	Point Biserial r_{pb}
Ordinal Y	Biserial r_b	Spearman rho/Tetrachoric r_{tet}	Rank Biserial r_{rb}
Nominal Y	Point Biserial r_{pb}	Rank Bisereal r_{rb}	Phi, L, C, Lambda

r = correlation coefficient

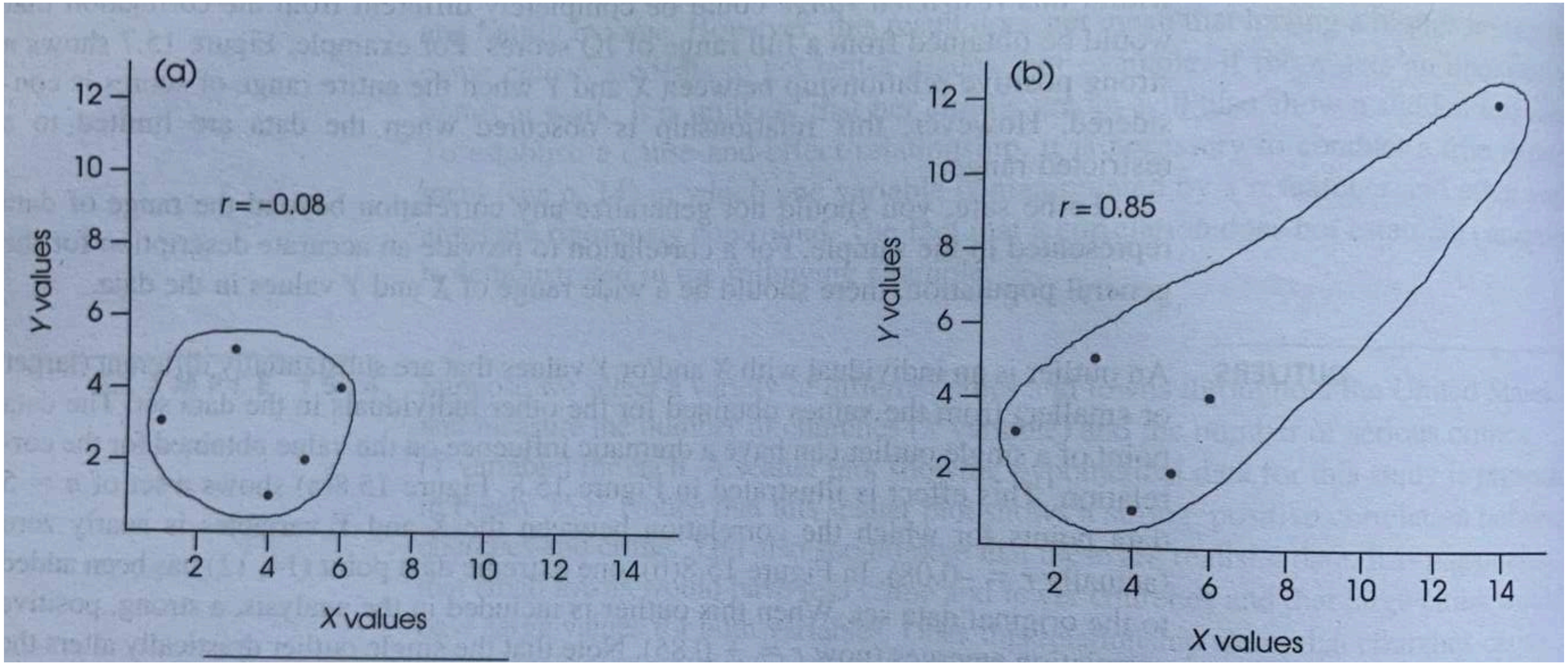
$\boldsymbol{r^2}$ = coefficient of determination

r = ?

Pearson's r = .48

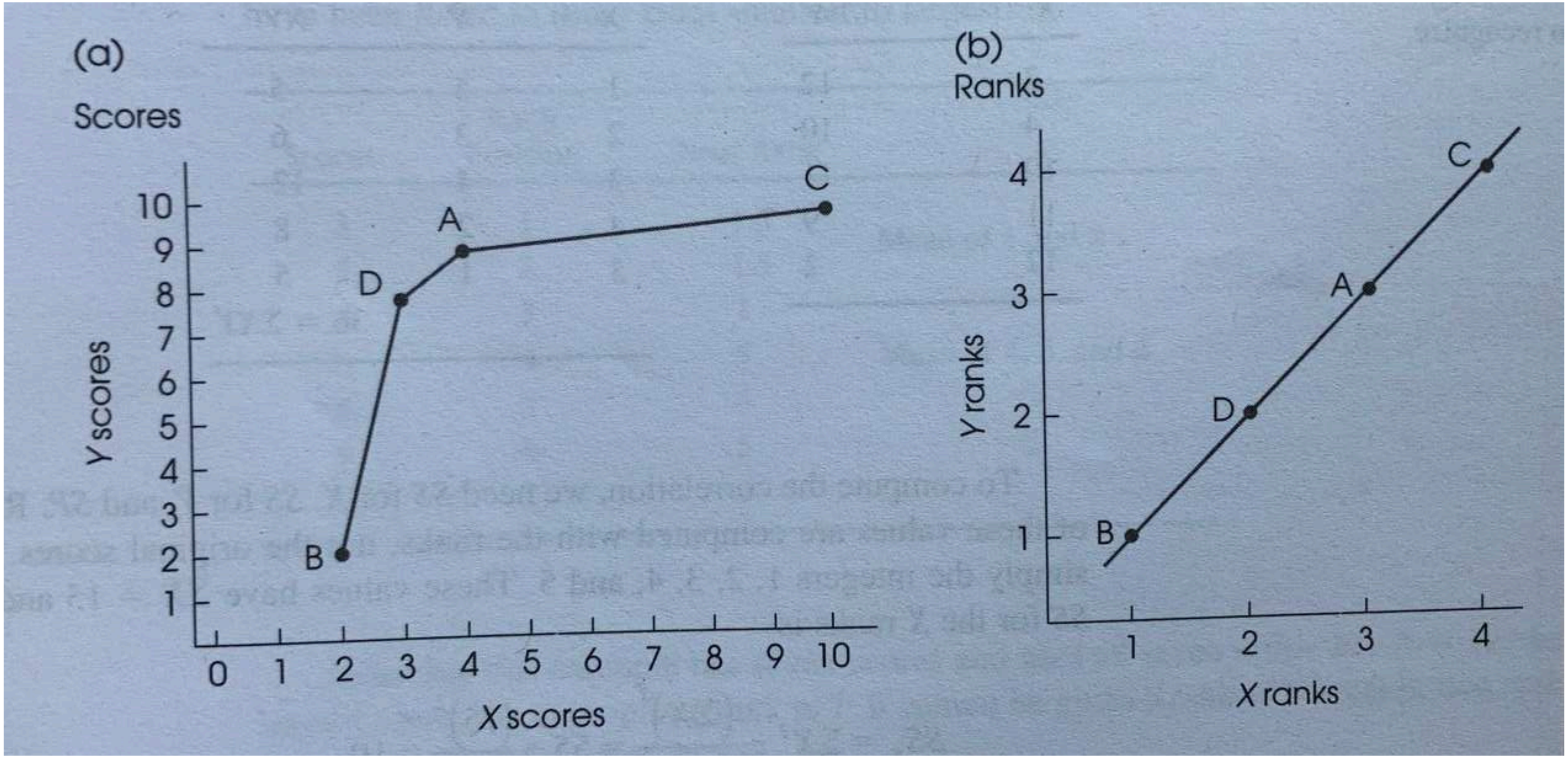


Pearson's r



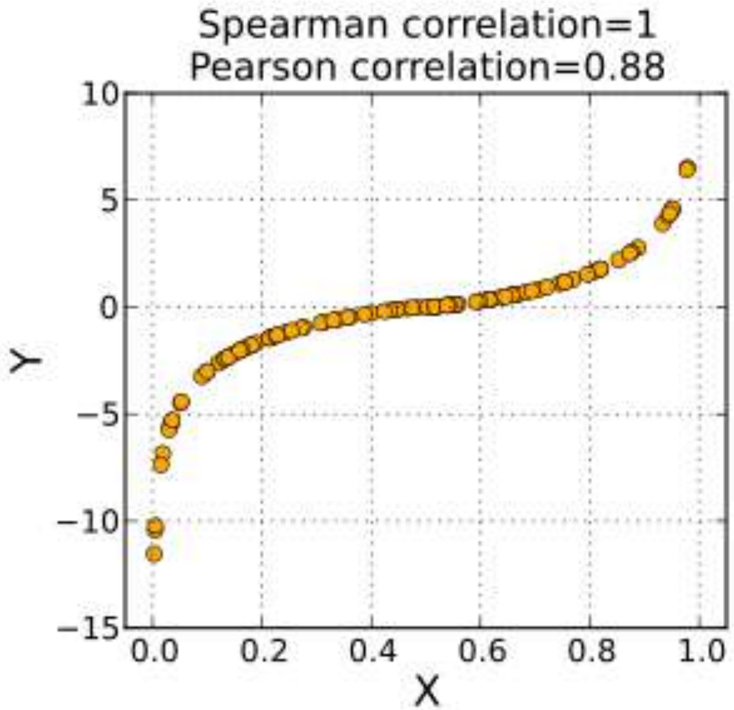
sensitive to outliers

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



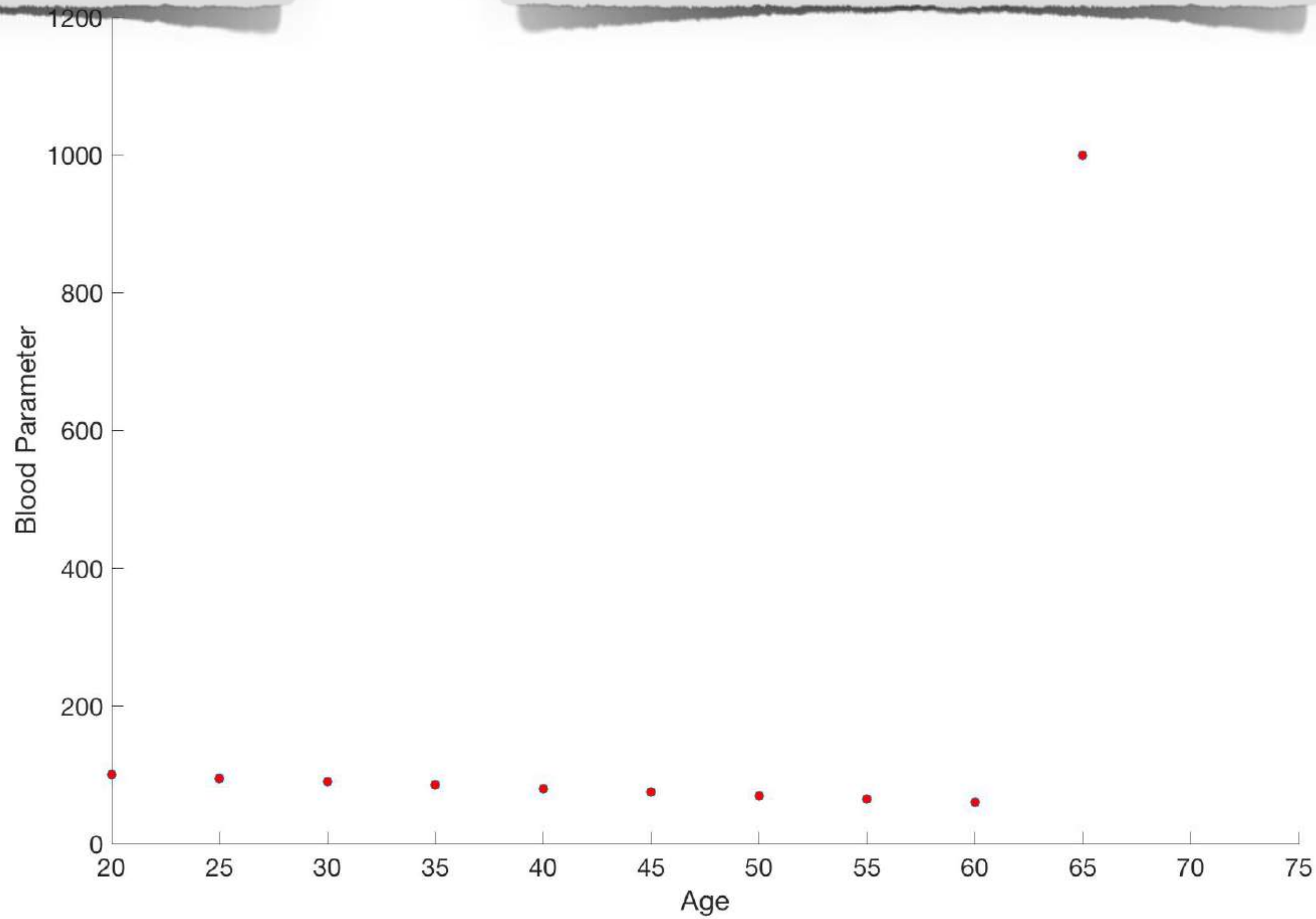
Pearson's *r* vs Spearman's *rho*

- Pearson's sensitive to outliers

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

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

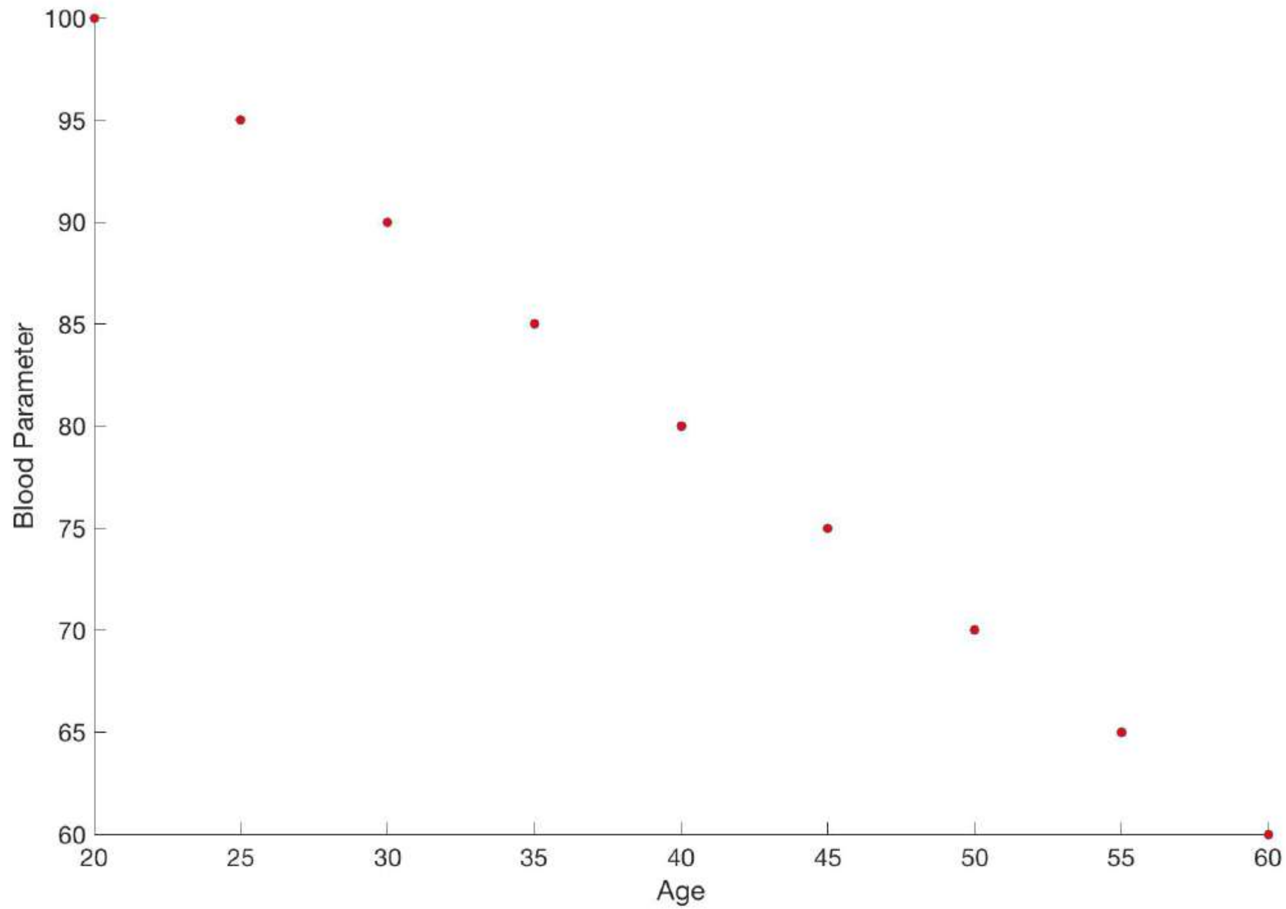
r = ?



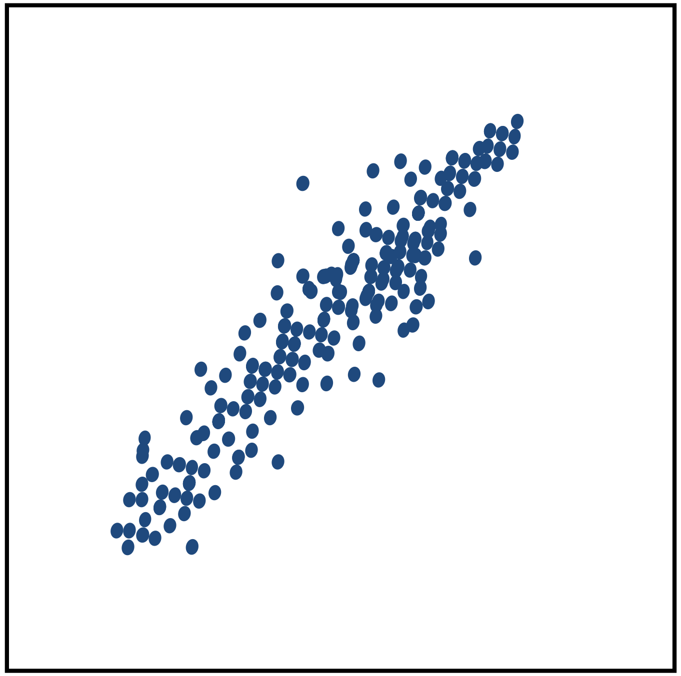
Pearson's *r* vs Spearman's *rho*

Pearson's ***r*** = -1

Spearman's ***r*** = -1

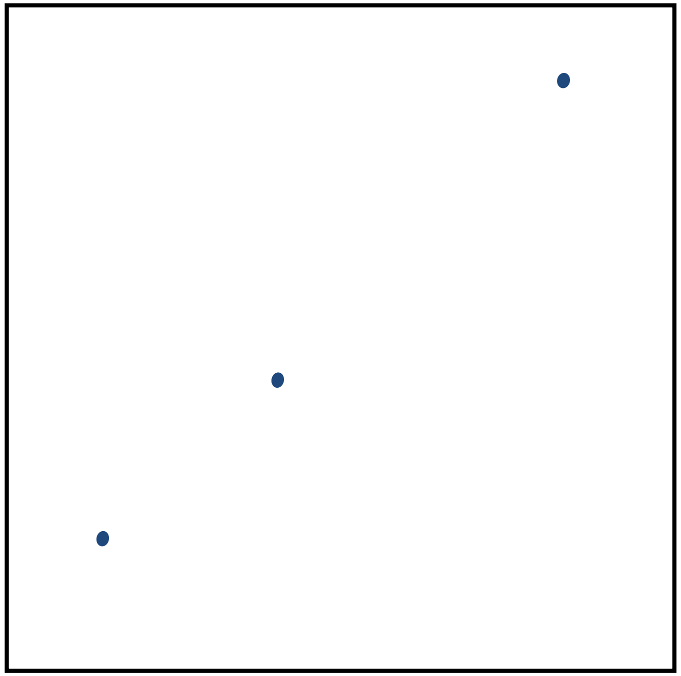


Significance of Correlation



r = 0.85

Is this significant?

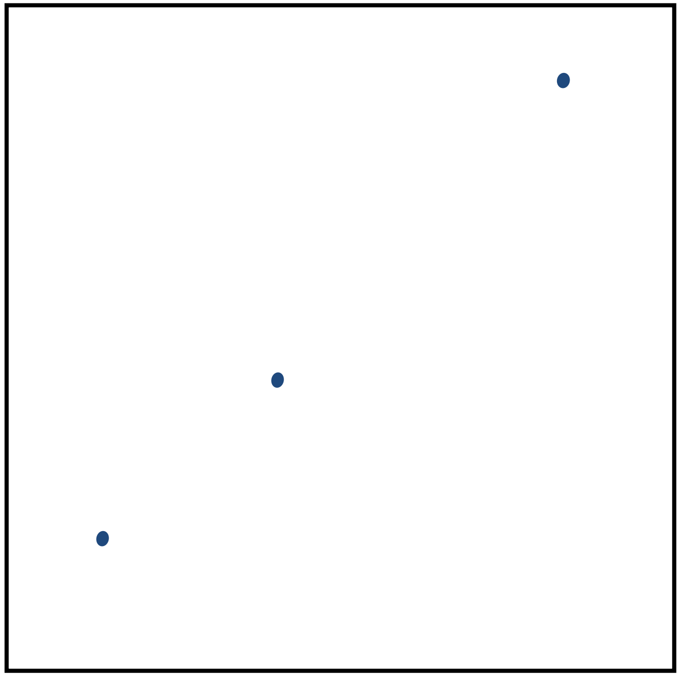


r = 0.99

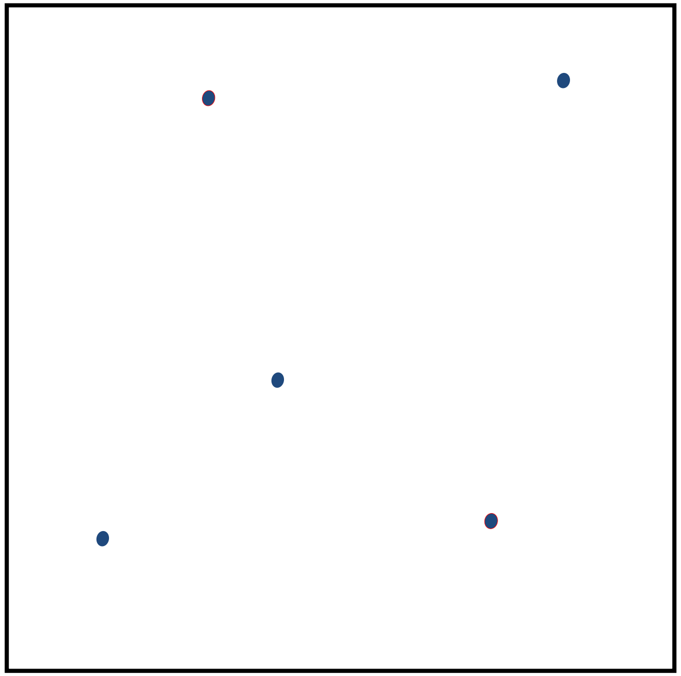
Is this significant?

Significance of Correlation

Add 2 more points to the plot



r = 0.99

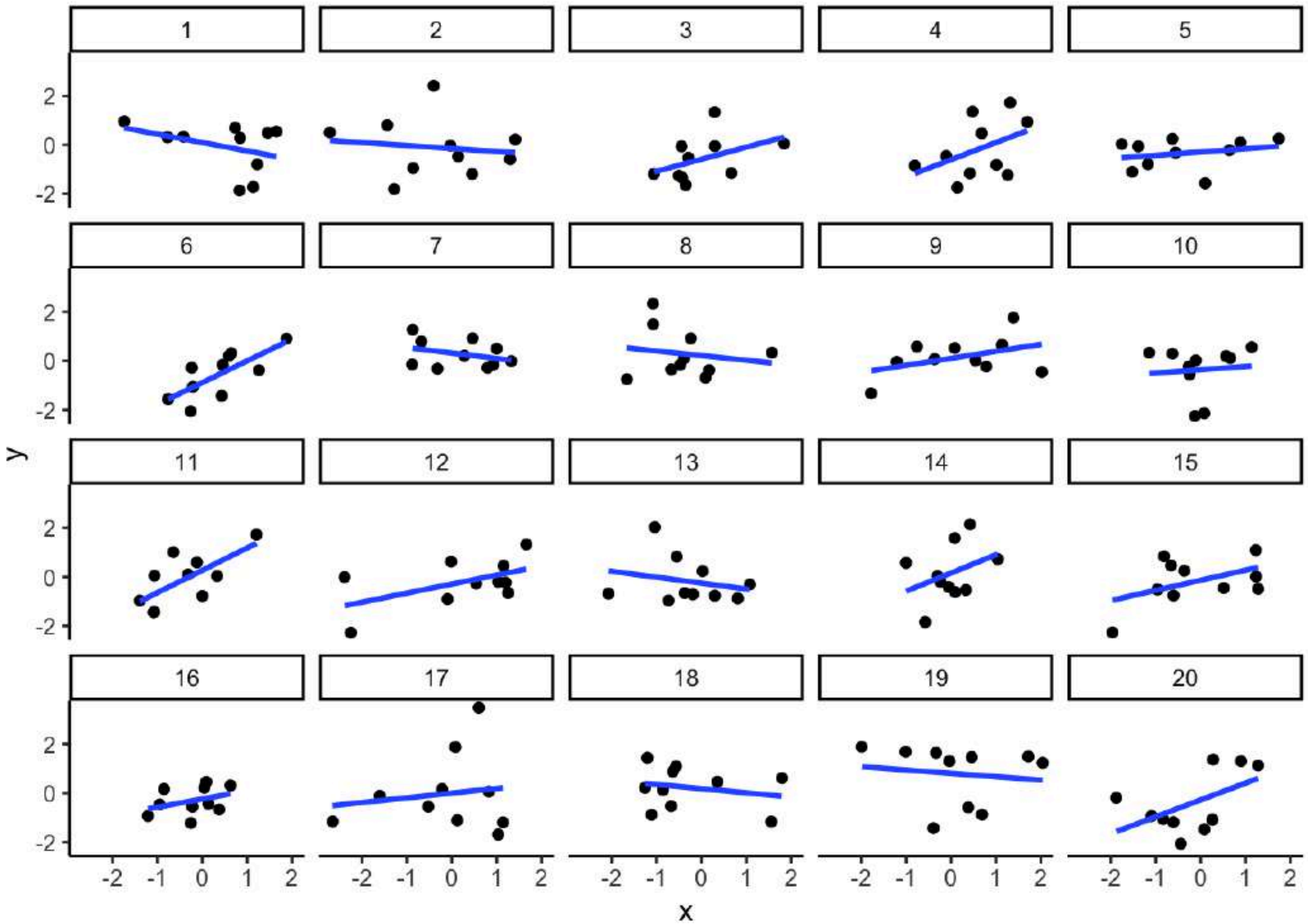


r = 0.05

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.

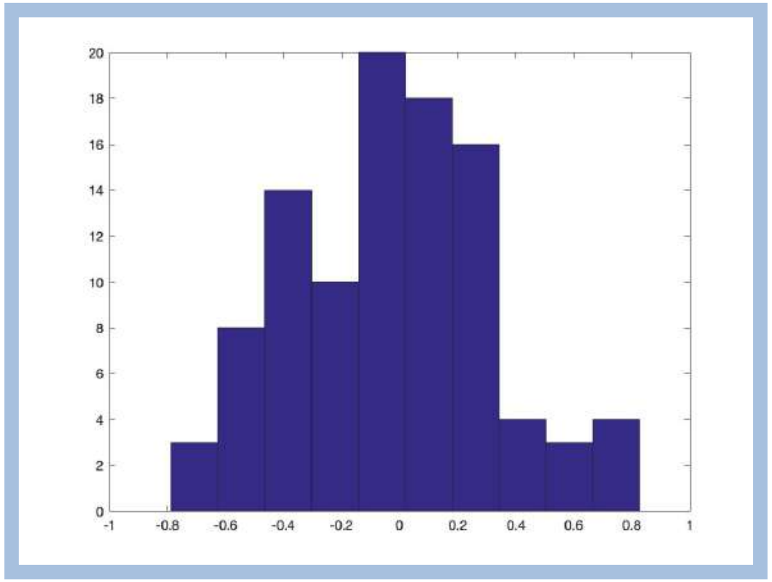
Let's Simulate

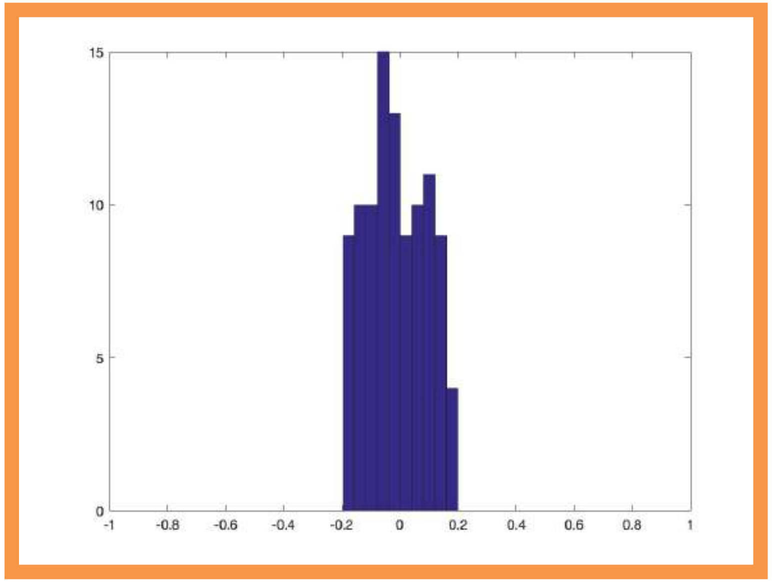
How would the distributions of *r* look like for the following:

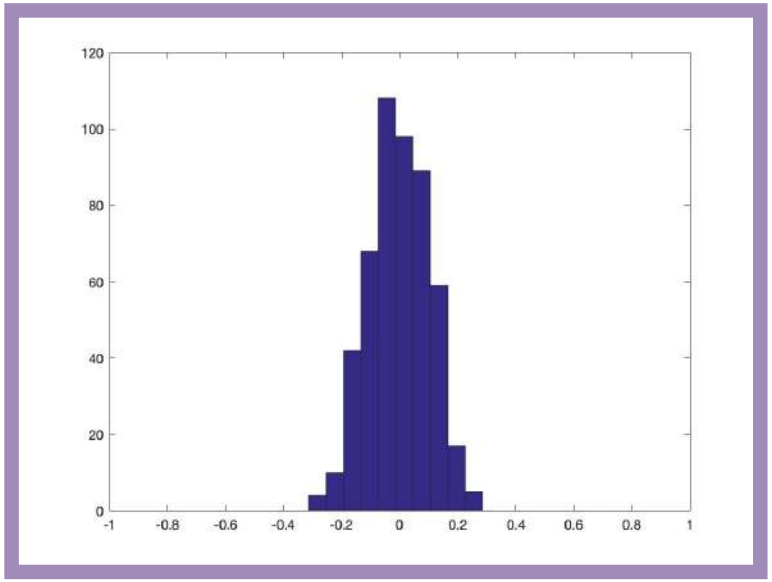
i) sample size = 10, iterations = 100

ii) sample size = 100, iterations = 100

iii) sample size = 100, iterations = 500





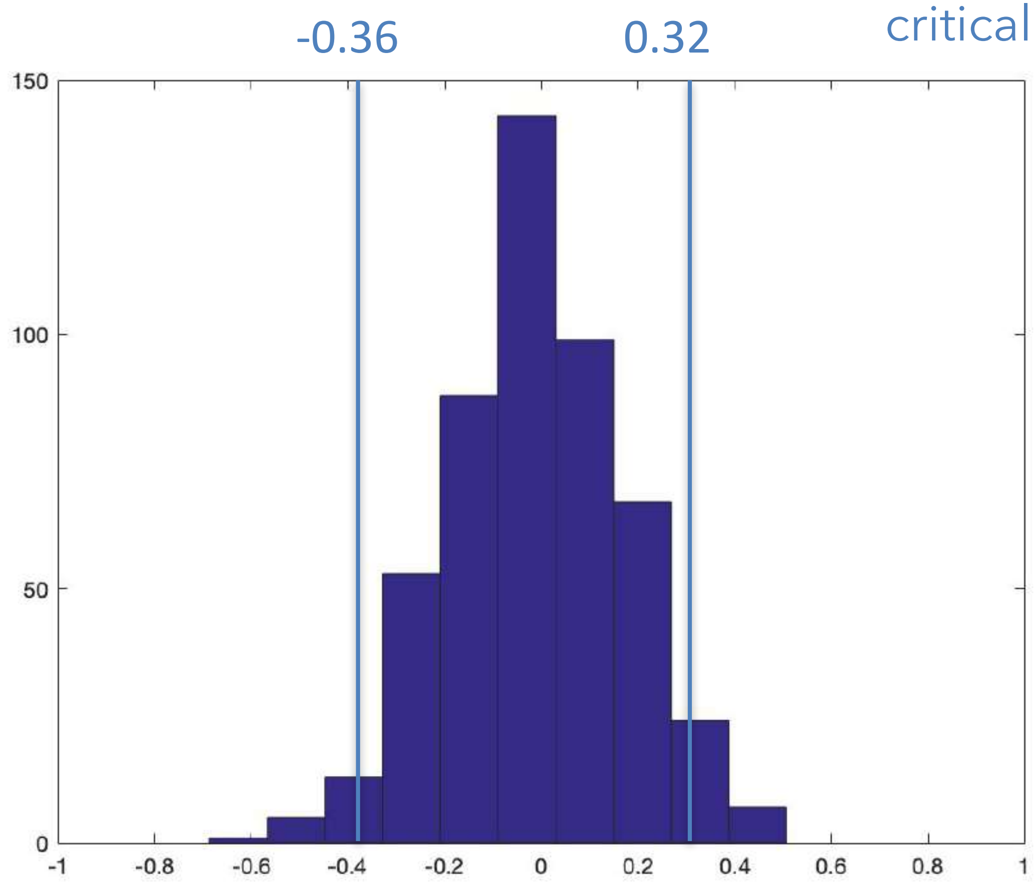


Let's Simulate

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

i) n = 30, iterations = 500

critical values (α < .05)



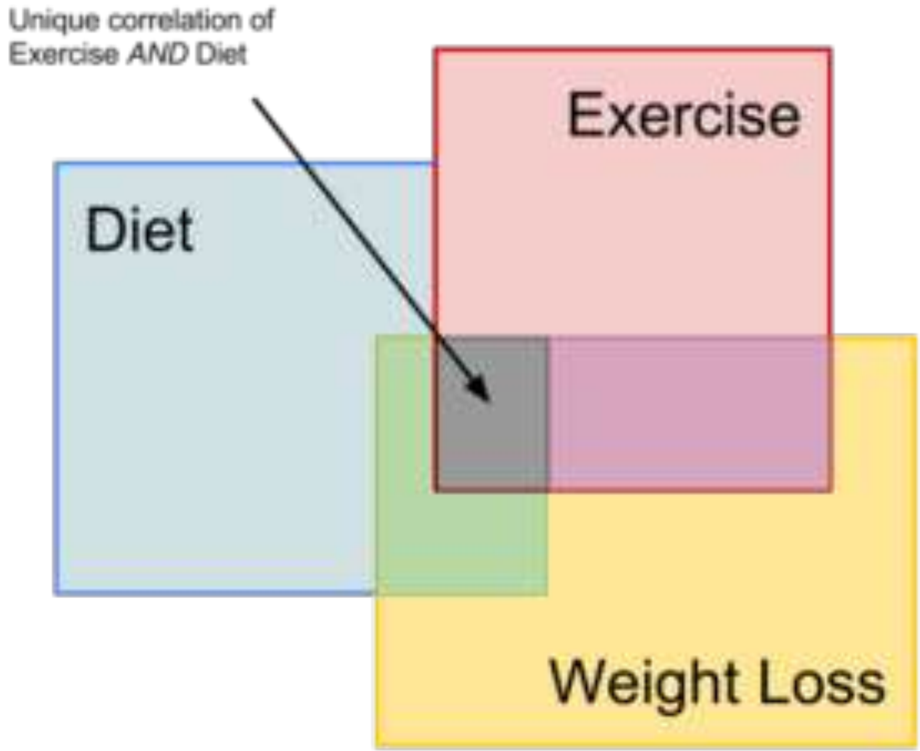
two-tailed

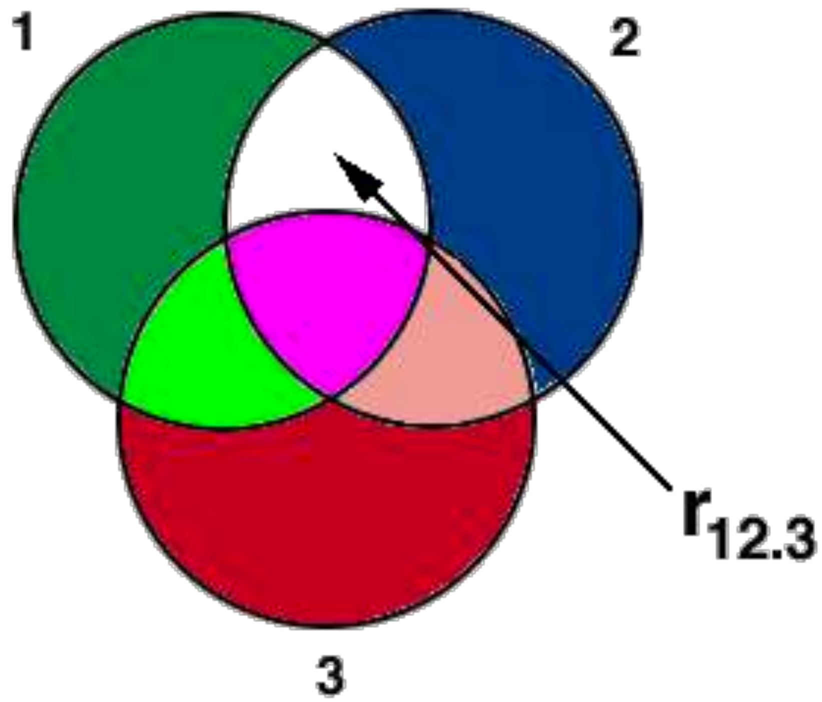
vs

one-tailed

How would the significance of the correlation change if you correlated time-series?

Partial Correlation





Partial Correlation

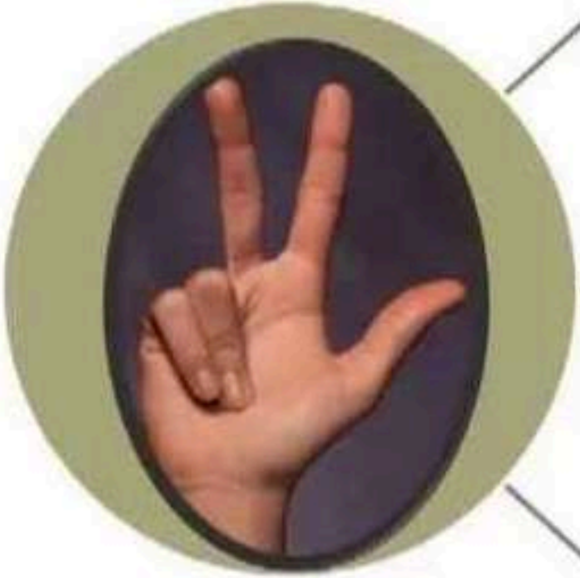
- measure of association between two variables, while controlling or adjusting the effect of one or more additional variables
  - What is the relationship between test scores and IQ scores after controlling for no. of hours of study?

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

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.

Reliability

Vinoo Alluri

Reliability

- **consistency** and **stability** of a research instrument (ex: measure or score or person)

- any measure we use in research should be reliable, otherwise it's useless

- **repeatability** of a method/test or research findings

Kinds of Reliability

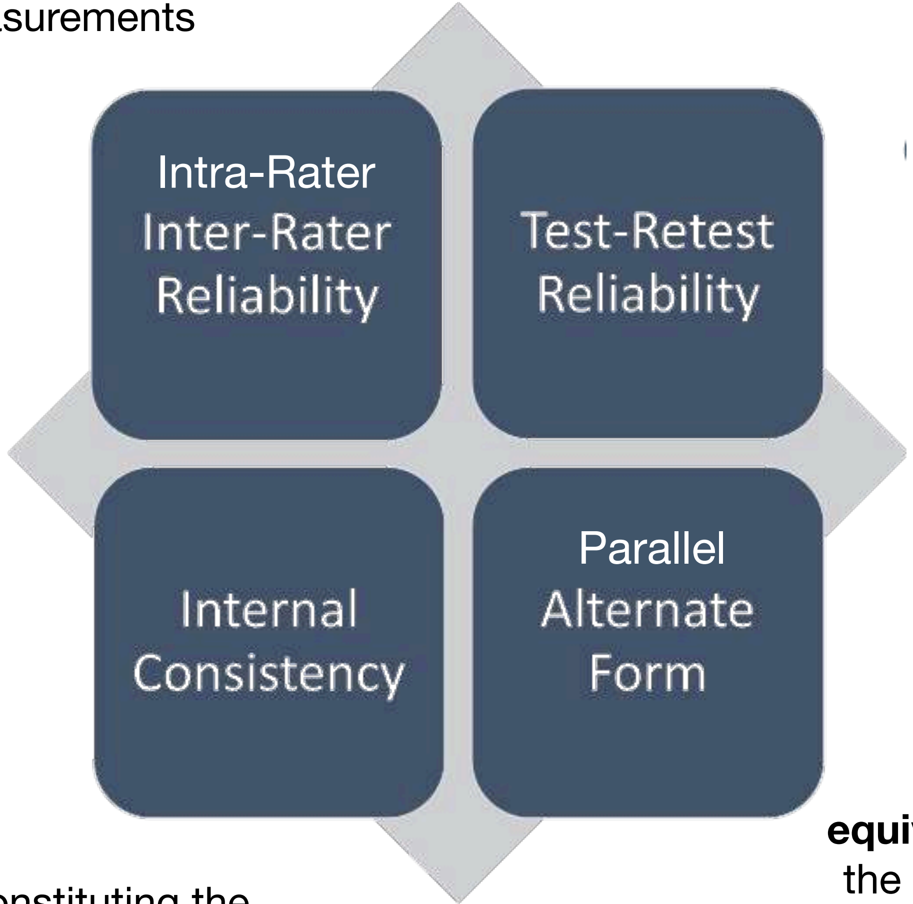
- Tools/methods or measuring device

- People

Kinds of Reliability

stability and **degree of agreement** between **people** during measurements

stability and **consistency** of method/tool/apparatus over time/repeated measurements

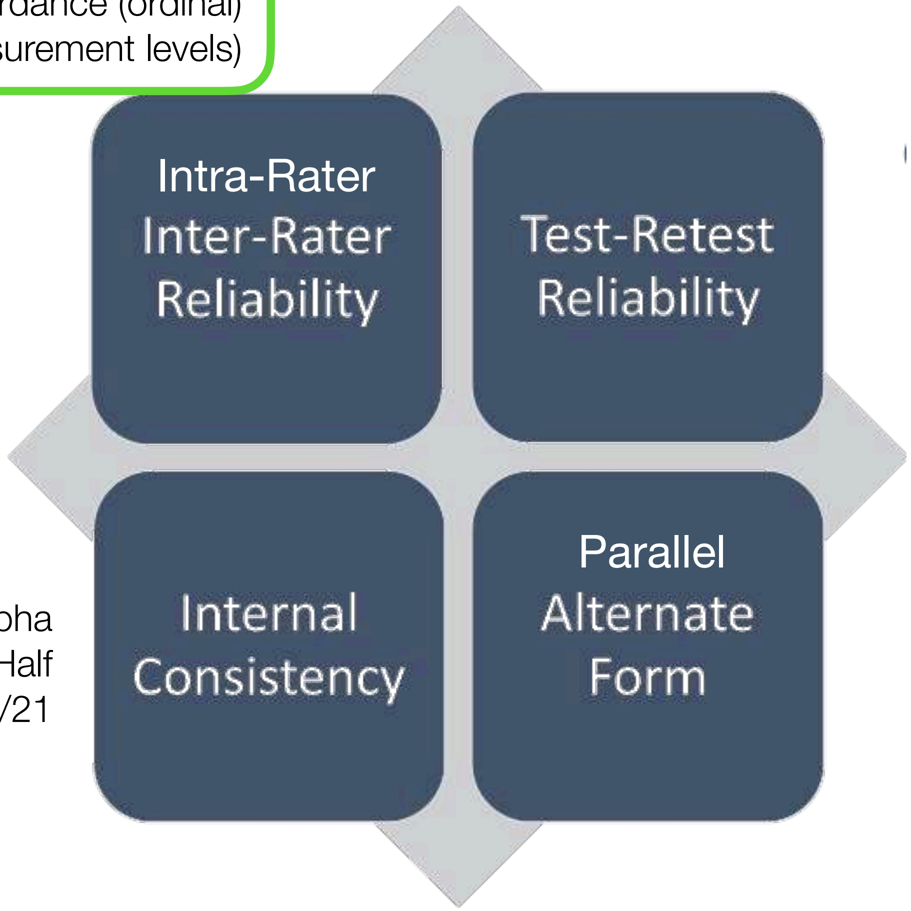


coherence of attributes constituting the method/tool/apparatus

equivalence of two versions of the method/tool/apparatus to compare results

Kinds of Reliability

Cohen’s Kappa (nominal; 2 raters)
Fleiss’ Kappa(nominal; >2 raters)
Kendall’s coefficient of concordance (ordinal)
Krippendorff’s Alpha (all measurement levels)



Pearson’s correlation

Cronbach Alpha
Split-Half
Kuder Richardson-20/21

Reliability

– **Internal consistency**: Is the measurement device consistently measuring what you want it to measure?

- Average inter-item correlation finds the average of all correlations between pairs of questions
- Split Half Reliability: all items that measure the same thing are randomly split into two. The two halves of the test are given to a group of people and find the correlation between the two. The split-half reliability is the correlation between the two sets of scores.
- Kuder-Richardson 20: average correlation for all the possible split half combinations in a test.

Reliability

- **Internal consistency**: Is the measurement device consistently measuring what you want it to measure?
  - ***Cronbach's alpha***:
    - was developed in 1951 by Cronbach Lee to meet the need of finding an objective way of measuring the internal consistency reliability of an instrument used in a research work
    - mostly used when the research being carried out has multiple-item measures of a concept
    - typically used in questionnaires/surveys (self-reported)

Reliability

– **Internal consistency**: Is the measurement device consistently measuring what you want it to measure?

  - ***Cronbach's alpha***:

$$\alpha = \frac{k\bar{r}}{(1+(k-1)\bar{r})}$$

  - *$r \equiv$ mean inter-indicator correlation*
  - *k=number of indicators or number of items*

Reliability

EXAMPLE

– **Internal consistency**:
  – we have a 5 item scale showing data collected from 100 respondents

0 = Never 1 = Almost Never 2 = Sometimes 3 = Fairly Often 4 = Very Often

1. In the last month, how often have you been upset because of something that happened unexpectedly?	**0**	**1**	**2**	**3**	**4**
2. In the last month, how often have you felt that you were unable to control the important things in your life?	**0**	**1**	**2**	**3**	**4**
3. In the last month, how often have you felt nervous and “stressed”?	**0**	**1**	**2**	**3**	**4**
4. In the last month, how often have you felt confident about your ability to handle your personal problems?	**0**	**1**	**2**	**3**	**4**
5. In the last month, how often have you felt that things were going your way?	**0**	**1**	**2**	**3**	**4**

Reliability

EXAMPLE

– **Internal consistency**:
  – we have a 5 item scale showing data collected from 100 respondents
  – Correlate 100 responses x 5 items matrix

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1.0				
Item 2	.35	1.0			
Item 3	.42	.31	1.0		
Item 4	.25	.38	.41	1.0	
Item 5	.21	.36	.46	.31	1.0

$$\alpha = \frac{k\bar{r}}{(1+(k-1)\bar{r})} = .73$$

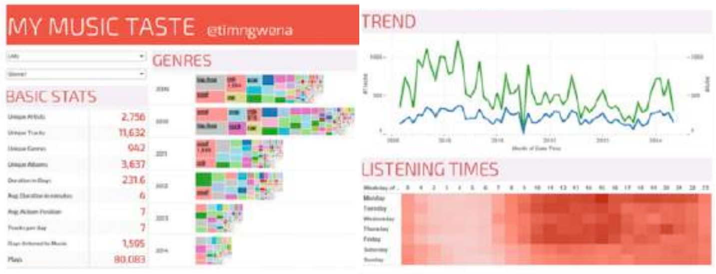
Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Reliability

EXAMPLE

– **internal consistency**

lost·fm



Users

r = 0.55, df=539, p<0.001

Psychological Distress Score

HEALTHY-UNHEALTHY MUSIC SCALE

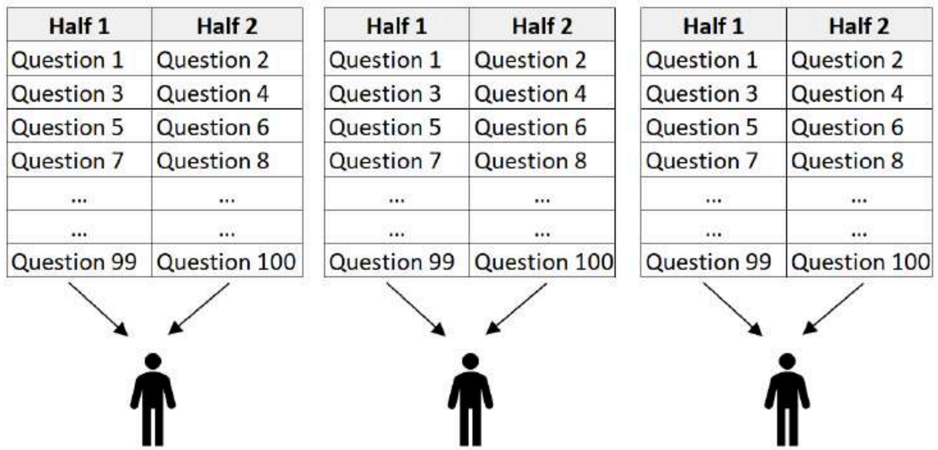
Cronbach α = .91

Cronbach α = .80

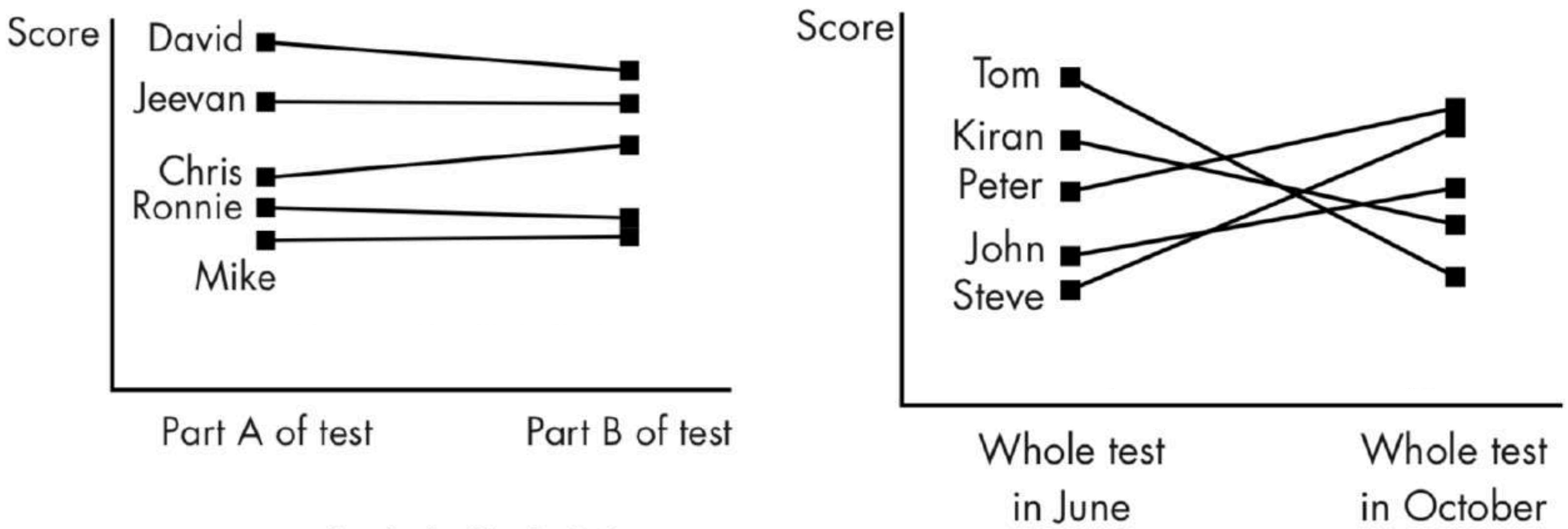
Reliability

– **Internal consistency**: Is the measurement device consistently measuring what you want it to measure?

- ***Split-half*** :
  - uses only some of available correlations;
  - compare results of one half to the other half.
  - If the test is reliable each half should be



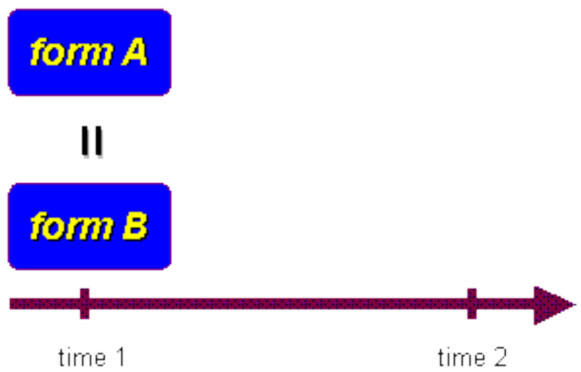
Reliability

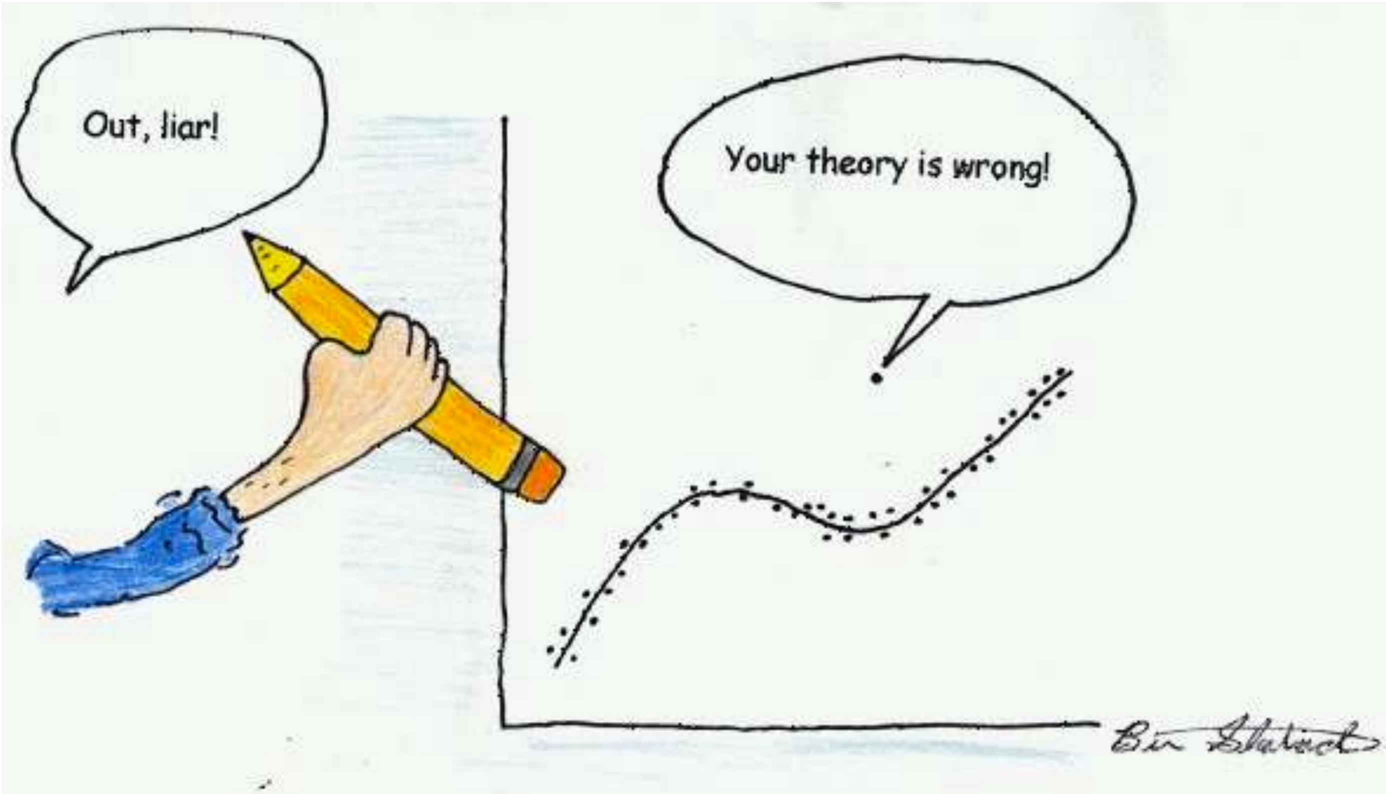


What kind of reliability and how good/bad is it?

Reliability

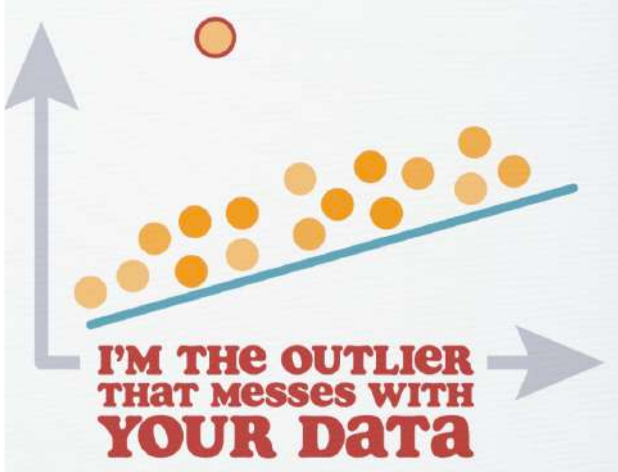
- **parallel forms:**
  - measure of reliability obtained by administering different versions of an assessment tool (both versions must contain items that probe the same construct, skill, knowledge base, etc.) to the same group of individuals
  - can avoid some problems inherent with test-resting





To have or not to have

Outliers



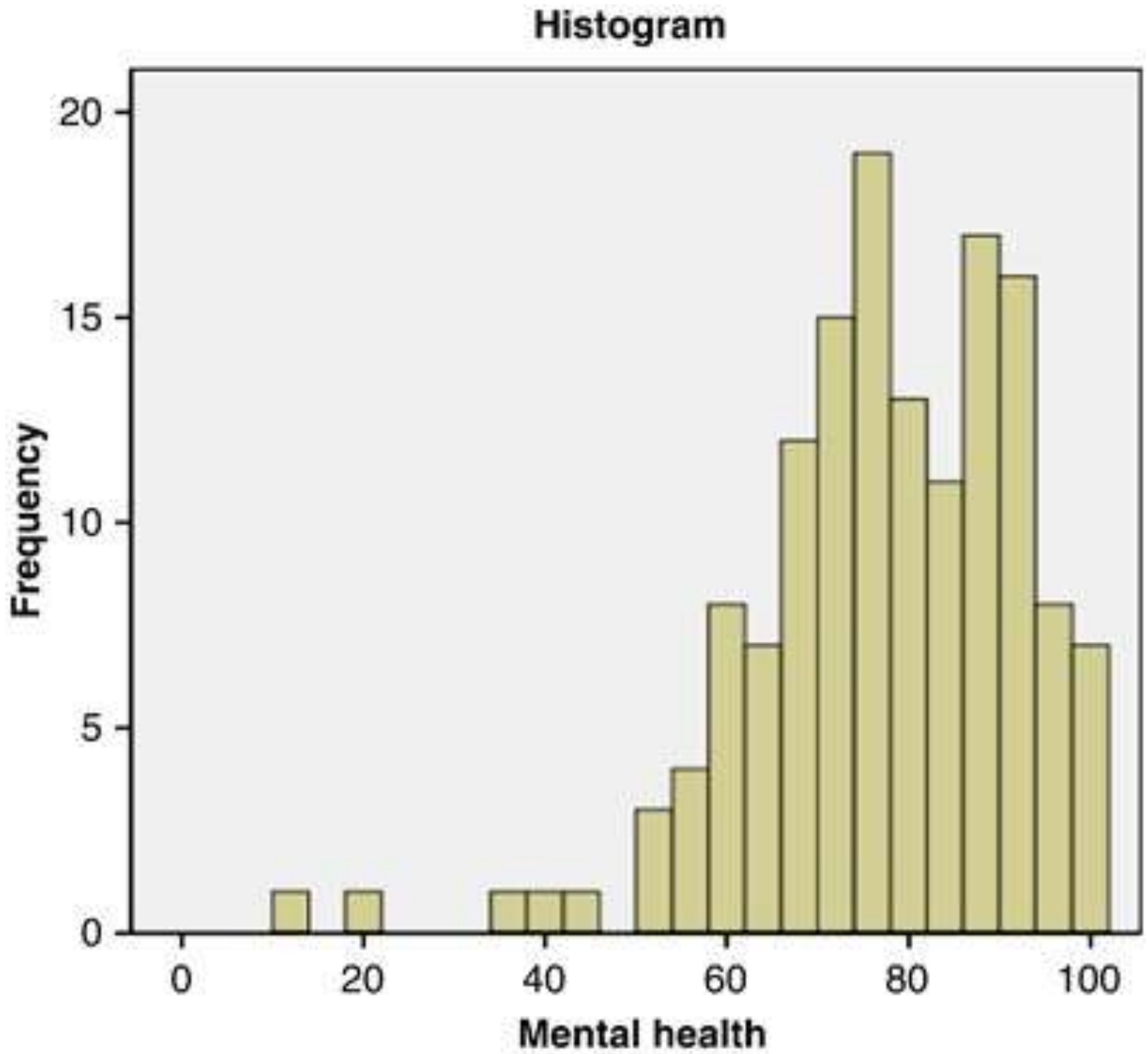
- detecting outliers is of major importance for almost any quantitative discipline (ie: Physics, Economy, Finance, Machine Learning, Cyber Security, Cognitive Science)
- not as common when sample size is low
  - ex: neuroimaging, qualitative studies involving interviews
- individual vs item/scale/stimulus

Dealing with Outliers

- omit
- replace (ex: with mean)
- using different analysis methods (ex: non-parametric tests)
- valuing the outliers
- data transformation

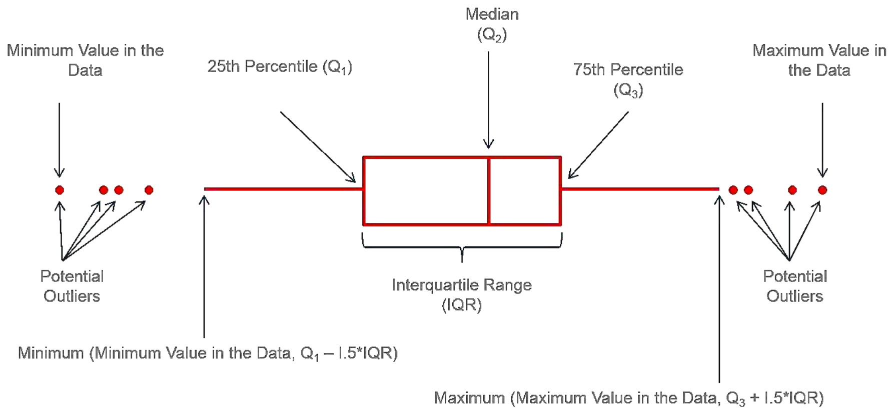
EXAMPLE

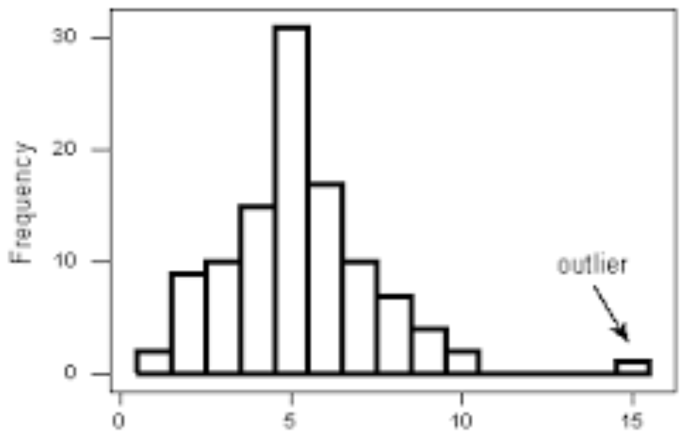
Natural Outliers

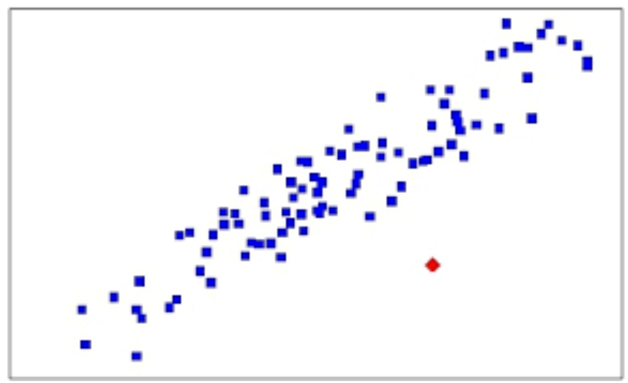


Outlier Detection

- graphical representations help (eg: scatter plot, box plot, histogram)







Outlier Detection

Intuitive way of detecting outliers (esp. in a perceptual experiment or survey)?

Outlier Detection

- graphical representations help (scatter plot, box plot, histogram)
- >1.5 x InterQuartile Range
- 2/3 SDs from mean (depending on the nature of data)
- Grubbs’ test (single), Tietjen-Moore test (multiple), etc..

EXAMPLE

Outlier (individual) Detection

- 2/3 SDs from mean (depending on the nature of data)
  - check individual 2SDs away from mean rating of each

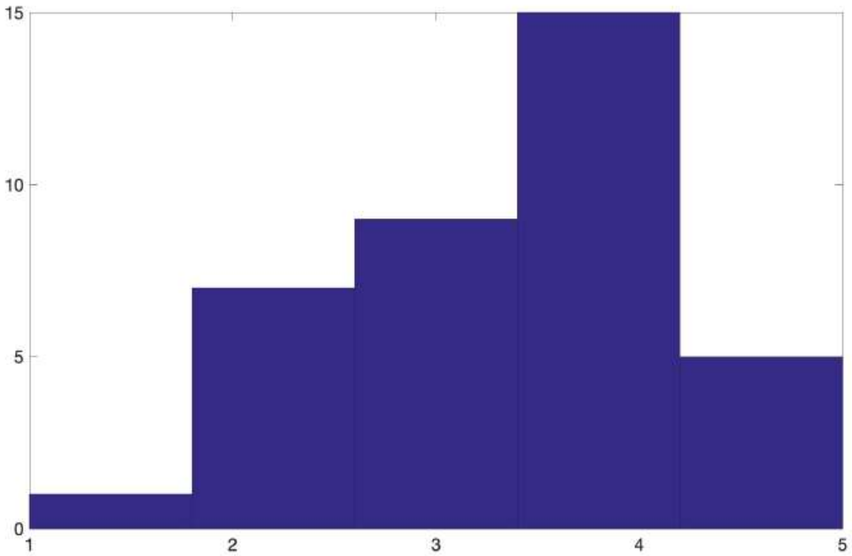
37 participants

37 x 100 Arousal ratings

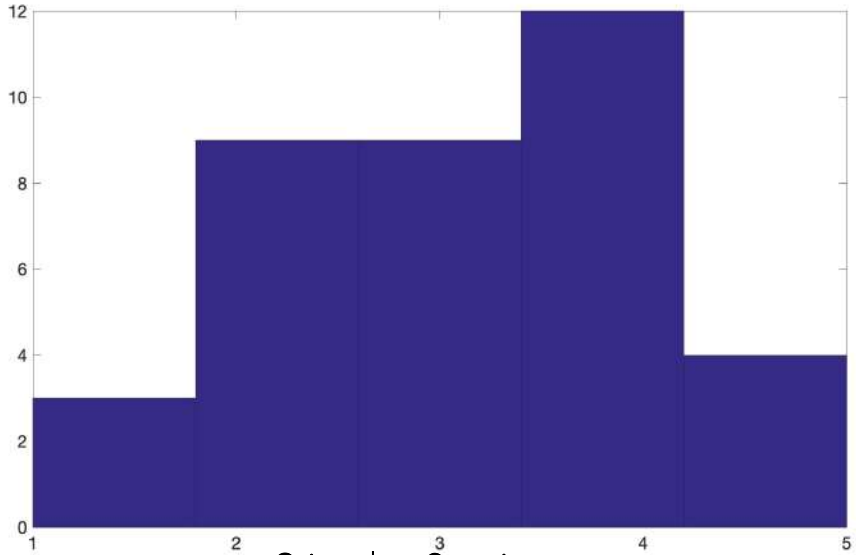
Rate **Arousal (Energy)** on a 5-point Likert scale
of
100 musical excerpts

EXAMPLE

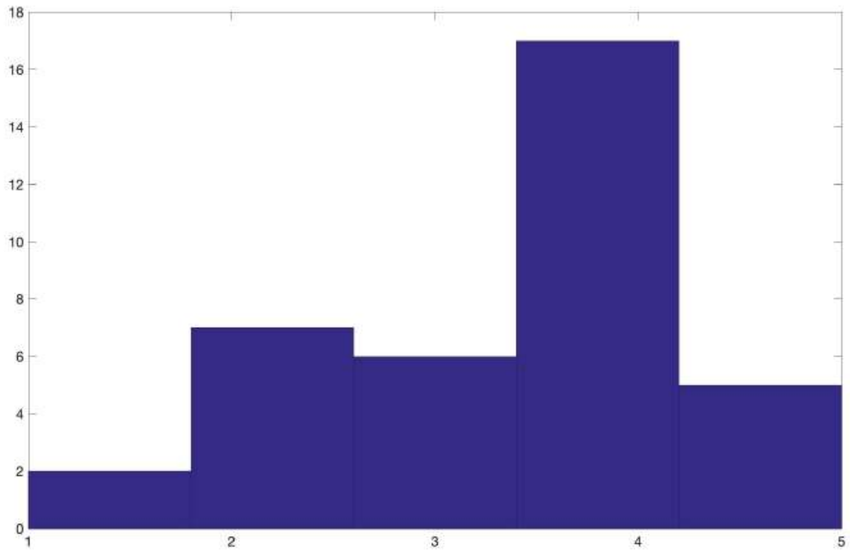
Outlier Detection



Stimulus 1 ratings



Stimulus 2 ratings



Stimulus 3 ratings

1 = low energy
5 = high energy

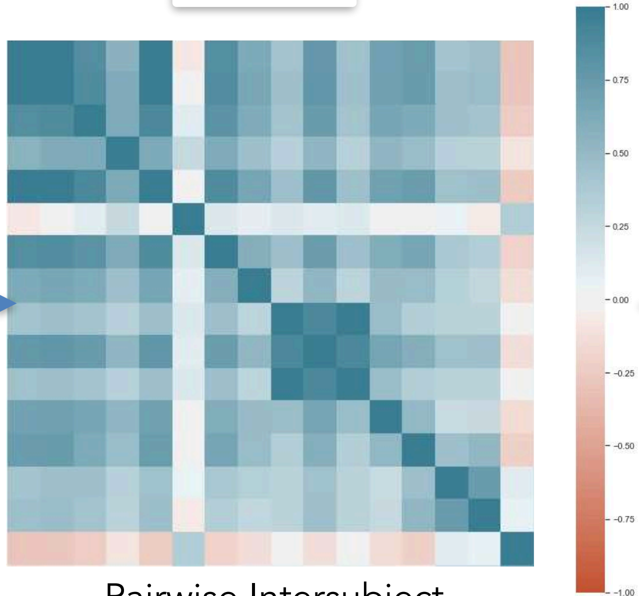
EXAMPLE

Outlier (individual) Detection

- 2SDs away from mean rating of each

37 x 37

37 x 100 Arousal ratings



Pairwise Intersubject Correlation matrix

Mean ISC

EXAMPLE

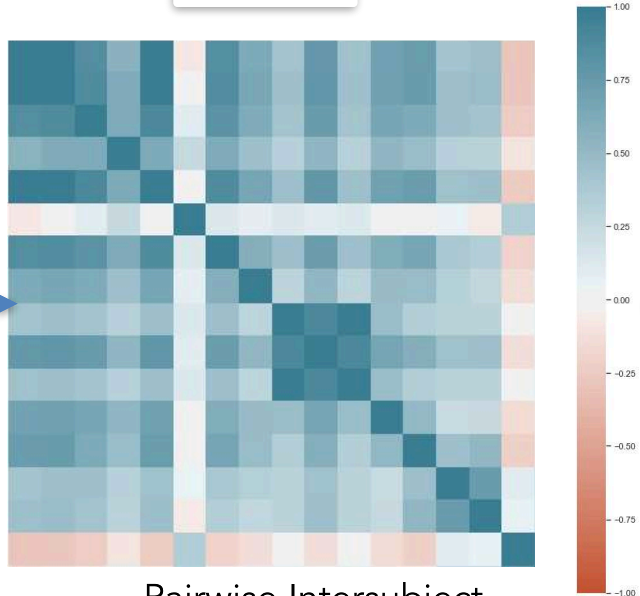
Outlier (individual) Detection

- 2SDs away from mean rating of each

29 participants

Rate **Adventurous/ Distressing** of 100 advertisements

29 x 29



Pairwise Intersubject Correlation matrix

Mean ISC

not always suitable (especially for subjective ratings)!