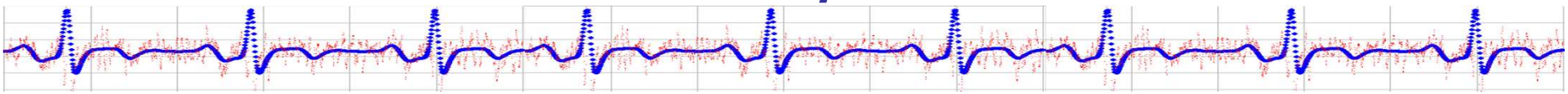
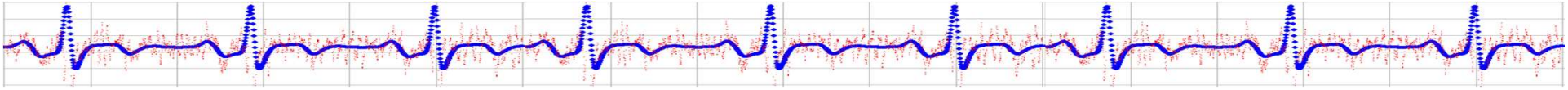
Empirical Research Methods in Information Science

IS 4800 / CS6350



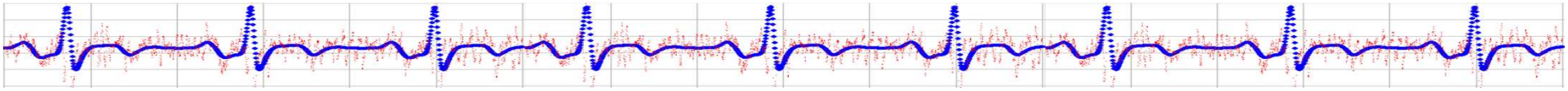
Midterm Review

Basic definitions



- Science (Describe, explain, predict)
- Non-Science
- Pseudoscience
- Protoscience (“fringe science”)
- Scientific explanation: empirical, testable, rational, general, parsimonious, tentative, rigorously evaluated

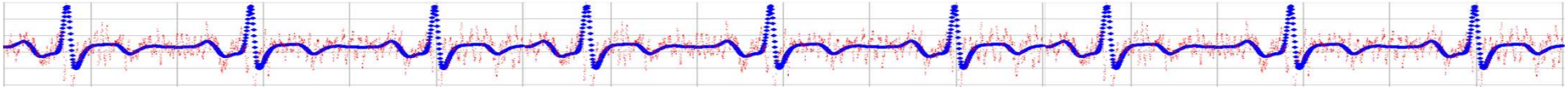
Research method



A strategy of inquiry, which moves from underlying philosophical assumptions to a research design

Methods (community standards):
Ethnography, authority, rational, scientific

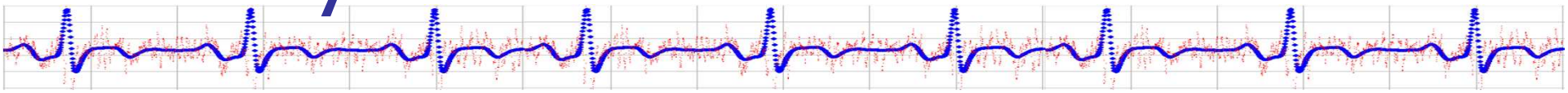
Research methods: The scientific method



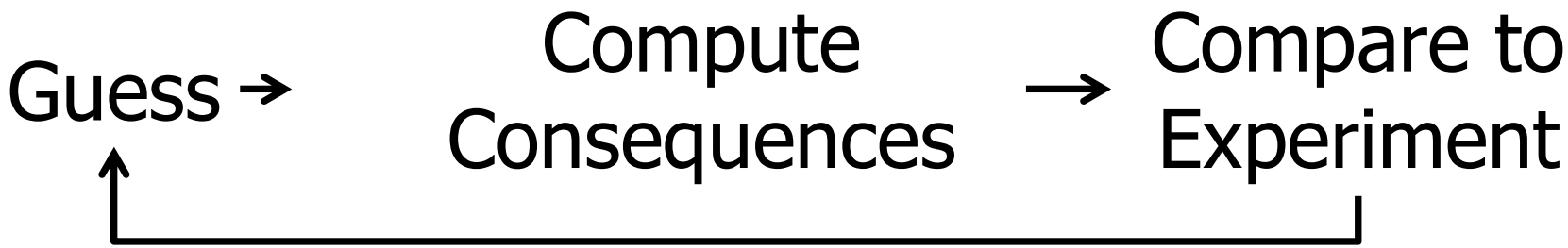
Steps?

1. Observe a phenomenon
2. Formulate testable explanations (hypotheses)
3. Further observe and experiment
4. Refine and retest explanations

Feynman's model



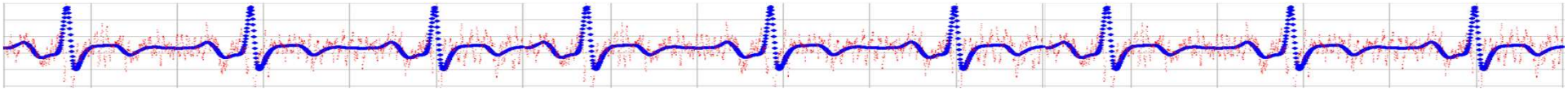
> "If it disagrees with experiment, it's wrong. In that simple statement is the key to science."
>
> — *Richard Feynman*

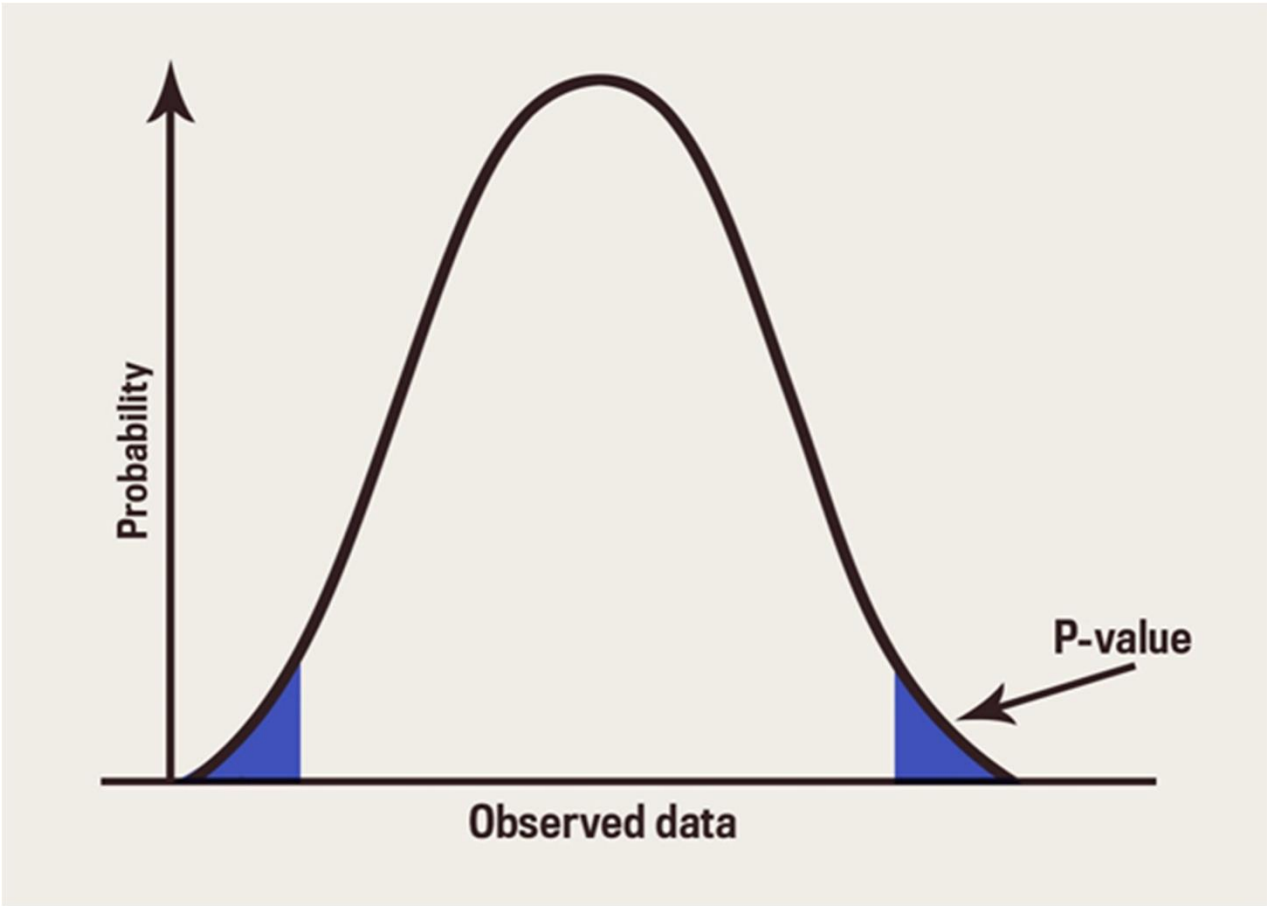


"falsifiability criterion"

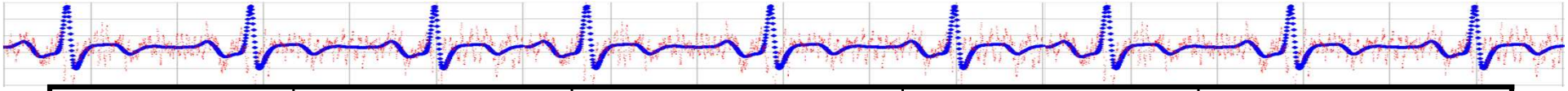
Watch Feynman: https://fs.blog/2009/12/mental-model-scientific-method/

Even with best practices, mistakes 5% of the time!



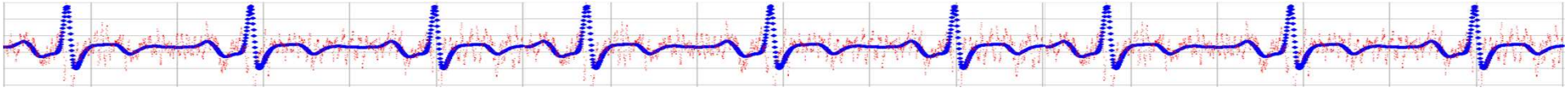


Sources of research information



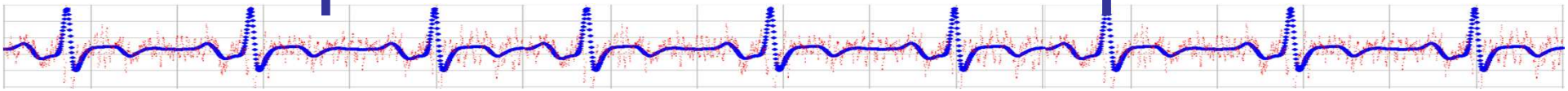
Example?	Scholarly	Substantive	Popular	Sensational
Appearance	Sober and serious	Attractive, with photographs	Attractive with many photos	Newspaper format
Reference Citations	Always provided	Sometimes cited	Rarely provided	Obscure references
Author	Scholar in the field	Scholar, editorial staff, freelance	Wide range of authors	Wide range
Language	Geared to scholars	For educated, no specialty	Simple, for less educated	Elementary for gullible audience
Content	Original research	No original research	Sources mentioned, may be obscure	Pseudoscientific sources
Publisher	Many by professional organizations	Commercial or professional organization	Commercial to entertain	Commercial to arouse curiosity

Factors affecting the quality of research information



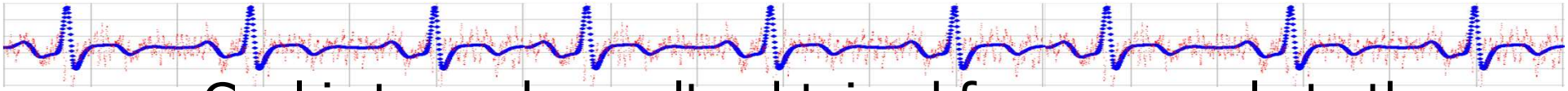
- Statistical Significance
  - Journals typically do not publish findings that do not meet the minimum .05 level of statistical significance
  - *File drawer* phenomenon: Findings that don't reach significance at .05 end up in the file drawer
    - If 100 articles about a phenomena are studied…
    - How to prevent?
  - "Effect Size" important when interpreting significance

Populations and Samples



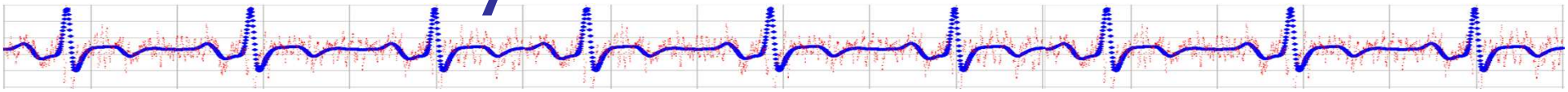
- *Population*
  - Large group including all potential subjects
  - May be defined in many ways
    - All children in day care
    - Children in day care in a particular city
- *Sample*
  - Small subgroup of subjects chosen from the population

Sampling and Generalization



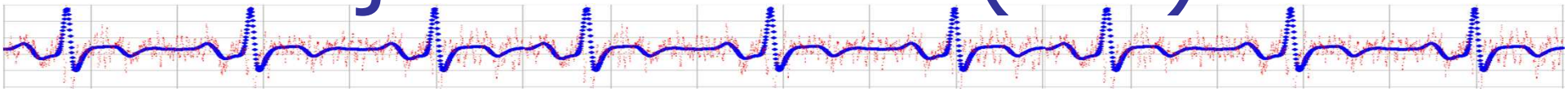
- Goal is to apply results obtained from a sample to the population
- *Generalization* is the ability to apply findings from a sample to the population
  - Aka "External Validity" of a study
- *Random sample:* A sample in which every member of the population has an equal chance of being chosen
  - Ideal that is not often met
- *Nonrandom sample:* A sample from a specialized population (e.g., college students)

Internal vs. External Validity of a study..



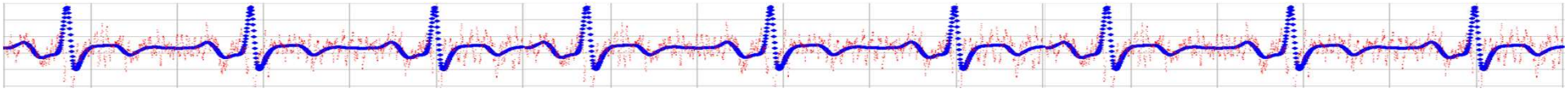
- Internal:
  - ability to prove/disprove hypotheses
  - appropriate methods (well designed)
  - conducted properly
  - data analyzed correctly
  - correct inference
  - replicability: could someone else conduct your study and get the same result?
- External:
  - generalizability

Ethical principals in human subjects research (IRB)



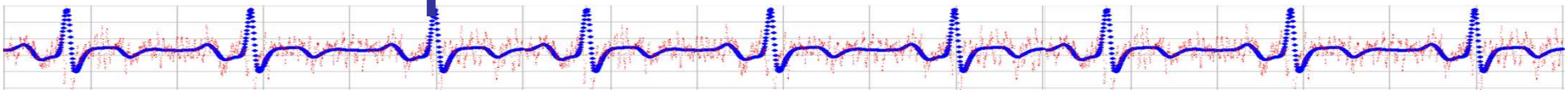
- Respect for persons (autonomous, special protections (minors, diminished capacity,…)
- Beneficence (maximize benefits, minimize harm)
- Justice (fair, risks/benefits distributed across society)
- Privacy (control over sharing of info)
- Confidentiality (treatment of info disclosed with trust)

Other IRB issues



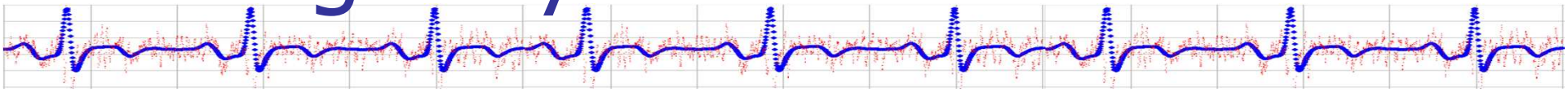
- Sensitive topics
  (sexual attitudes, drugs, …)
- Obtaining consent
- Fraud in research

Deception



- Active
- Passive (Conditioning, provoking negative behavior, …)
- Why is it a problem?
- Solutions (role playing, prior consent, debriefing)

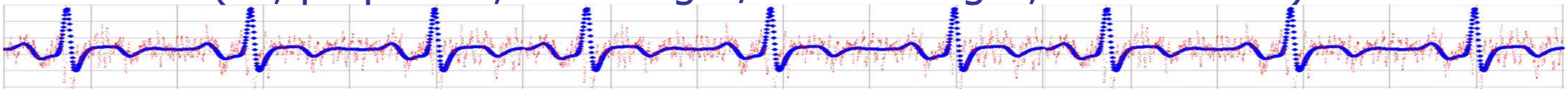
Eligibility



- State precise study population
- Eligibility (inclusion) criteria
  - Required to used in study (on recruitment)
- Exclusion criteria
  - Tests administered after informed consent
- Estimate eligibility given some information about the population

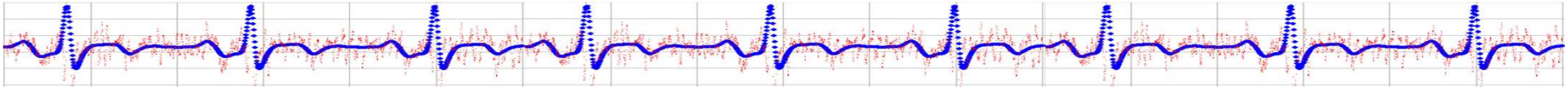
Types of studies

(ID, properties, advantages/disadvantages, when to use)



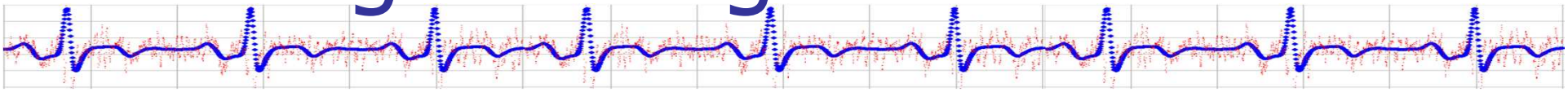
- Quantitative
  - Descriptive
  - Correlational
  - Experimental
  - Demonstration
- Qualitative
  - Ethnography (+ sociometry, content analyses, meta-analyses)
- Mixed

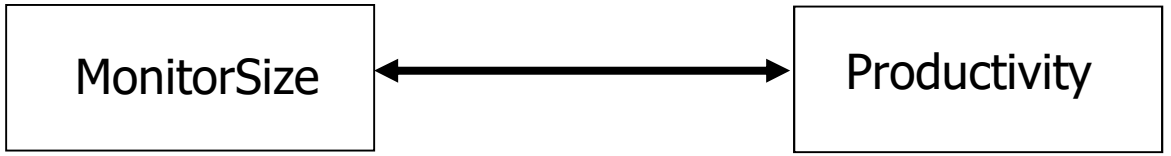
Research model variables

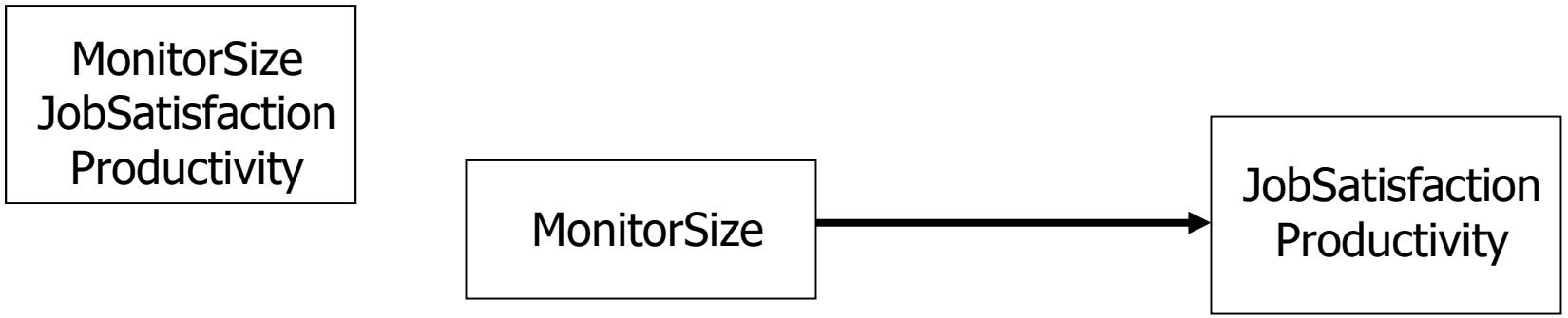


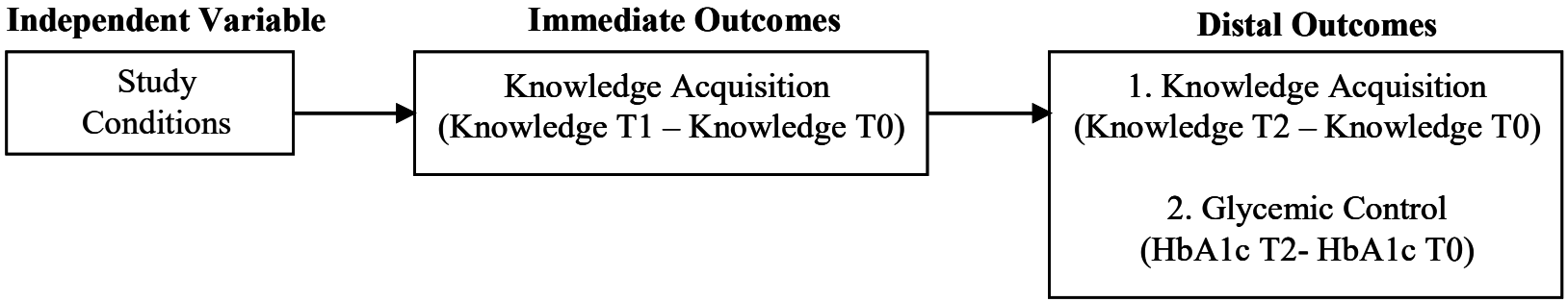
- IVs (treatments/conditions/arms)
  - How to isolate changes:
    - Hold external variables constant
    - Randomize properly
- DVs
- What arrows mean
  - Correlation vs. causality!
- Mediators
- Moderators

Diagramming

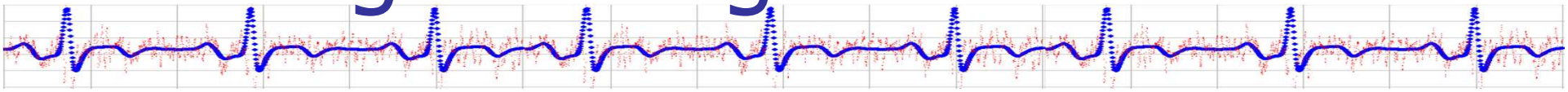


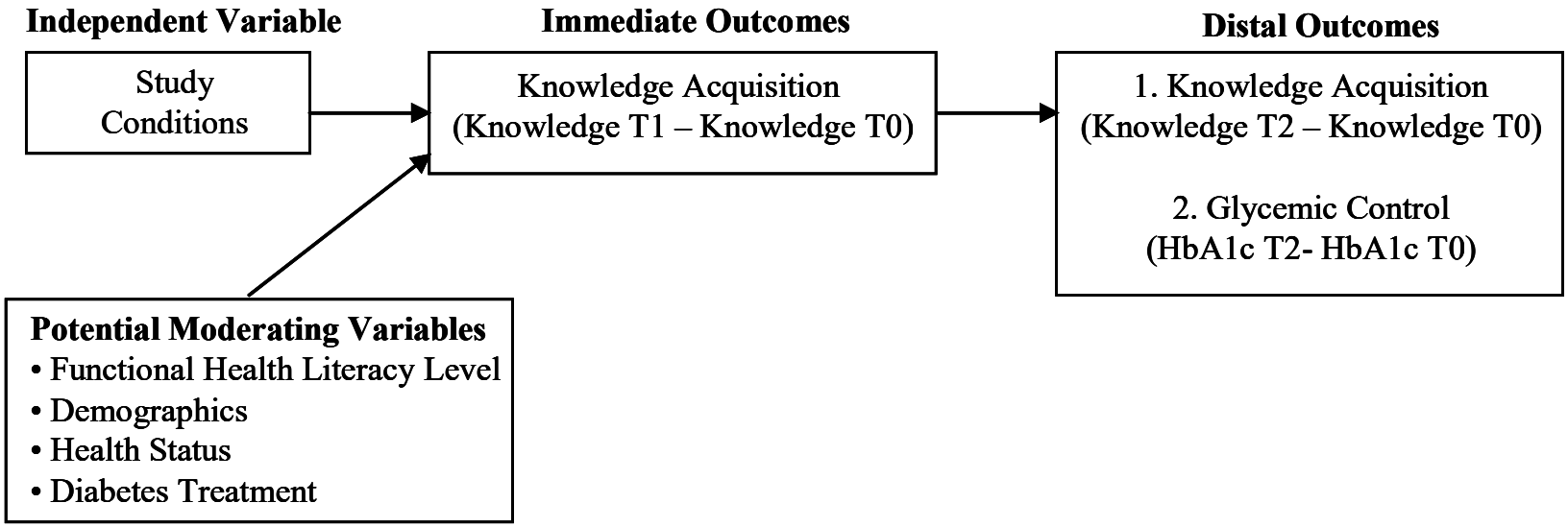




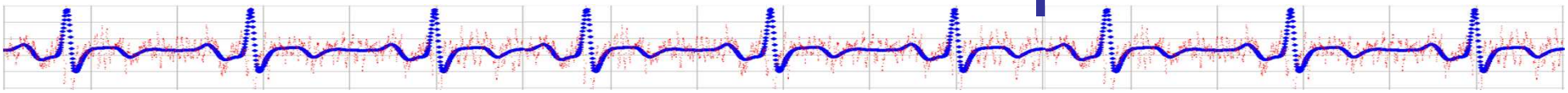


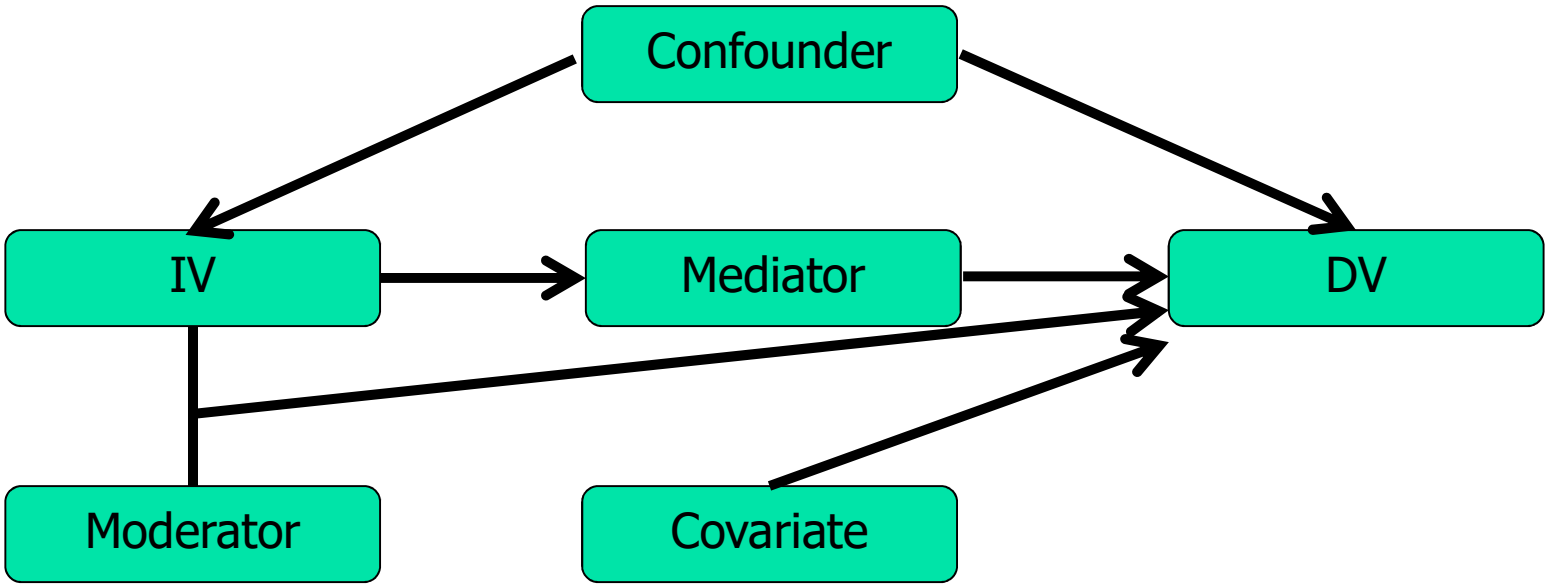
Diagramming



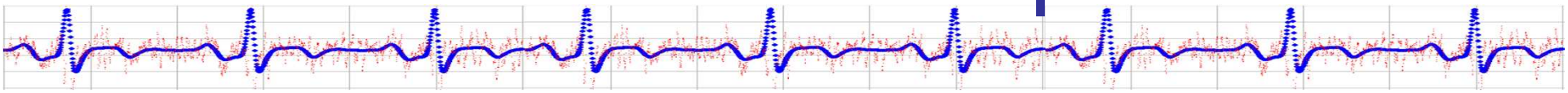


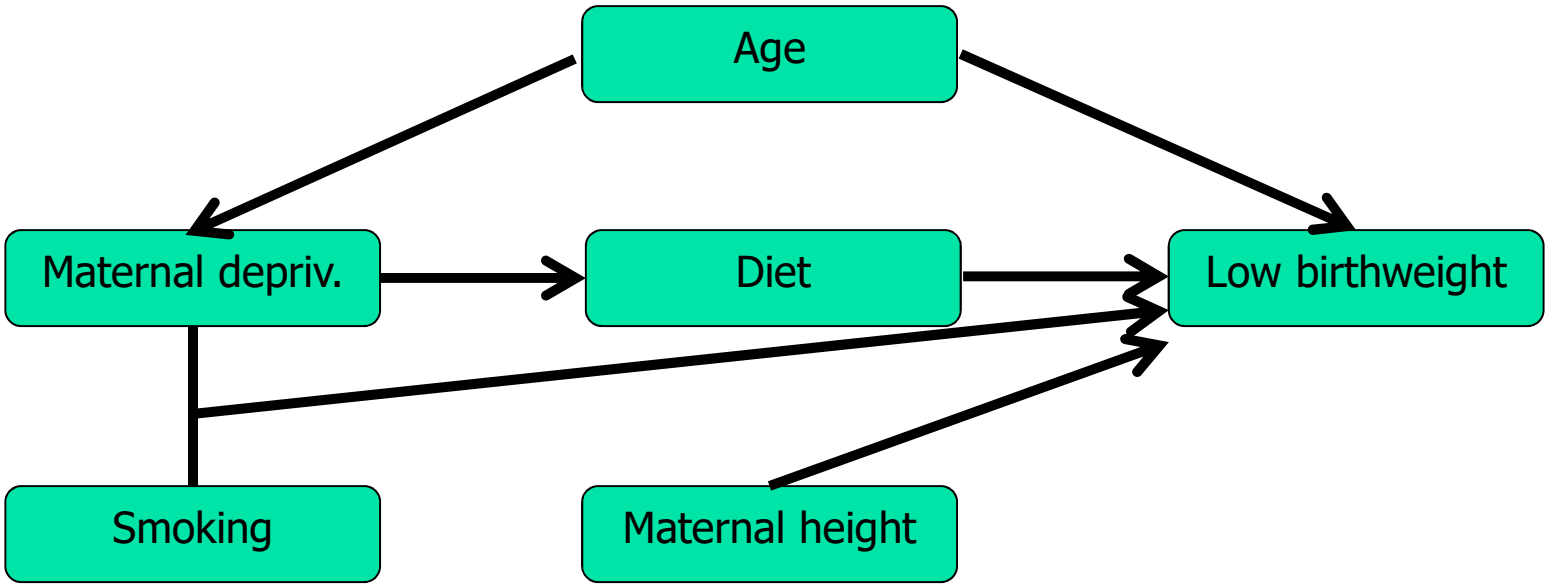
Measure relationships



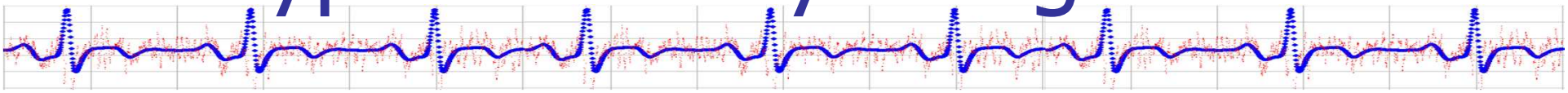


Measure relationships



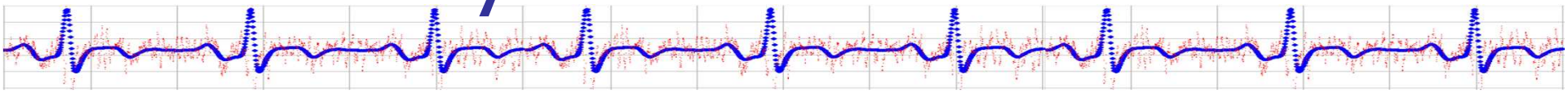


Types of study designs



	Number of Variables	Number of IV Levels	Manipulation
Descriptive	1	NA	NA
Demonstration	≥ 2	1	√
Correlational	≥ 2	NA	NA
Experimental	≥ 2	≥ 2	√

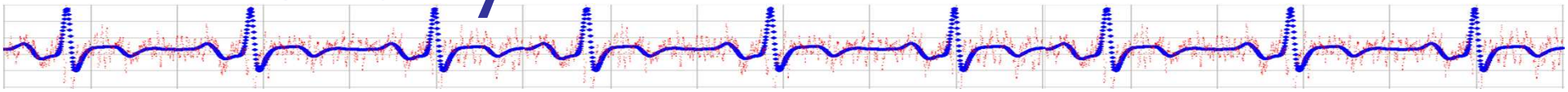
Validity



- Internal validity
  - Considered in design phase
  - Threats: confounding and extra variables

History	Events may occur between multiple observations.
Maturation	Participants may become older or fatigued.
Testing	Taking a pretest can affect results of a later test.
Instrumentation	Changes in instrument calibration or observers may change results.
Statistical regression	Subjects may be selected based on extreme scores.
Biased subject selection	Subjects may be chosen in a biased fashion.
Experimental mortality	Differential loss of subjects from groups in a study may occur.

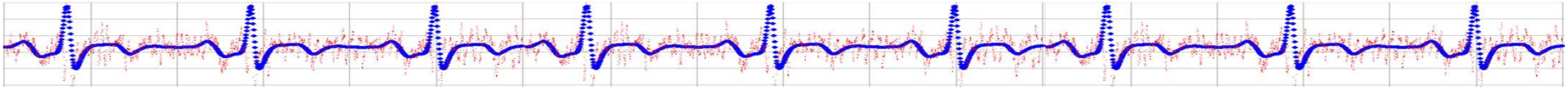
Validity



- External validity
  - Generalization
  - Threats: too much control

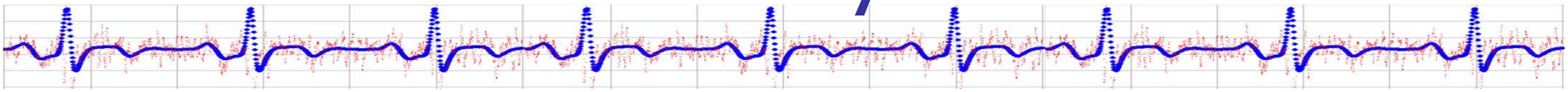
Reactive testing	A pretest may affect reactions to an experimental variable.
Interactions between selection biases and the independent variable	Results may apply only to subjects representing a unique group.
Reactive effects of experimental arrangements	Artificial experimental manipulations or the subject' s knowledge that he or she is a research subject may affect results.
Multiple treatment interference	Exposure to early treatments may affect responses to later treatments.

Measures



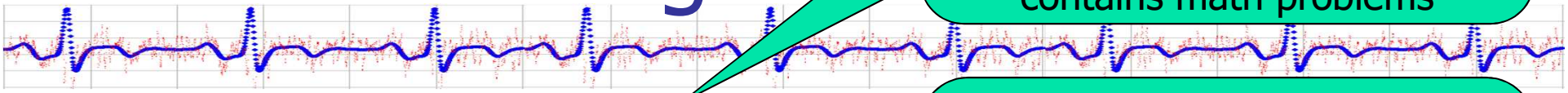
- Reliability – similar results under same conditions
  - Physical measures (precision)
  - Behavioral measures (inter-rater)
  - Questionnaire measures
    - Reliability: Test-retest, parallel forms, split-half
    - Internal consistency: Chronbach's alpha
- Validity – measures what you intend

Measure validity



- Physical: accuracy
- Questionnaire

Establishing vali



- *Face validity:* Subject
  adequacy of content.
  method
- *Content validity:* How
  a test sample beha
  measure?
  - Does each item relate t
  - Do the items collectivel

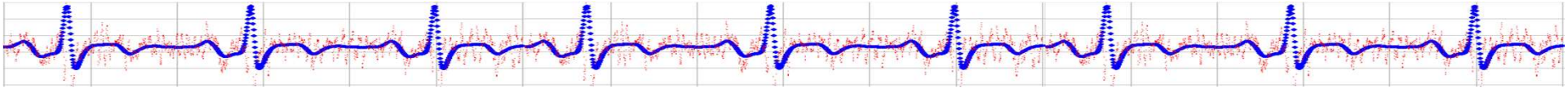
Whether the test "looks valid" to people using/taking it (important sometimes)
E.g., test of math ability contains math problems

Use of recognized domain experts; assess agreement among subject matter expert raters or judges regarding how essential a particular item is

Example: Political attitudes: include items relevant to all the major issues related to such attitudes (e.g., abortion, health care, the economy, and defense)

Example: Final exam covers all material in the course

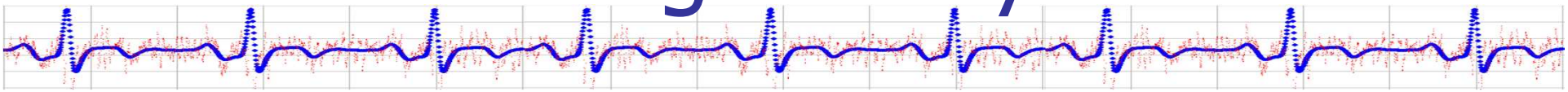
Construct



A variable, not directly observable, that has been developed to explain behavior on the basis of some *theory*

Examples: "intelligence," "self-esteem," "achievement motivation"

Establishing validity



- *Construct validity:* Do the results of a test correlate with what is theoretically known about the construct being evaluated? Does it measure what it claims?
  - Convergent validity (subtype): measures of constructs that *should* be related to each other are
  - Discriminant validity (subtype): measures of constructs that *should not* be related are not

ng validi

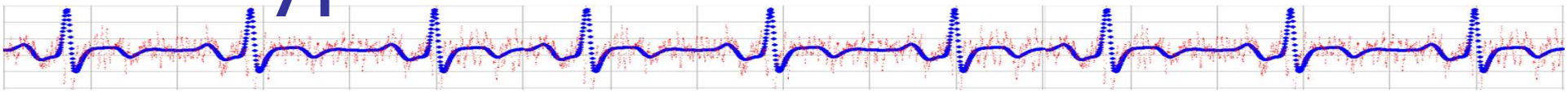
Unlike construct validity, not necessarily related to theory

E.g., Check if survey results correlate with another measure of the same dimension taken at the same time

- *Criterion-related validity*: How adequately does a test score match some criterion score? Takes two forms
  - Concurrent validity: Does test score correlate highly with score from a measure with known validity **taken at the same time**?
  - Predictive validity: Does test predict behavior at a later time known to be associated with the behavior being measured?

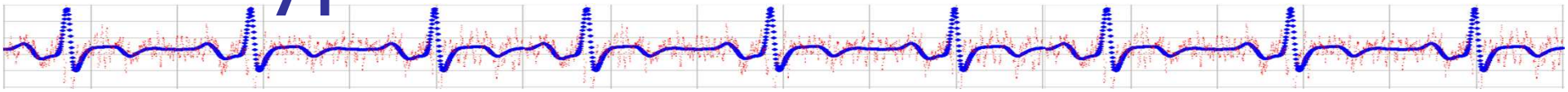
E.g., Survey predicts election results

Types of data



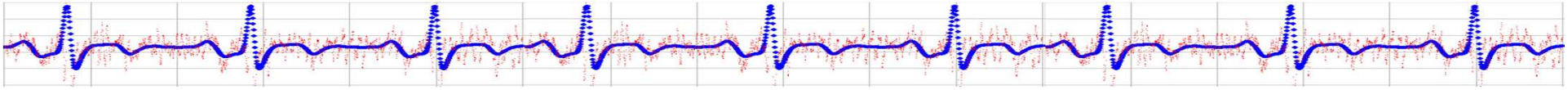
- Nominal
- Ordinal
- Interval
- Ratio

Types of data



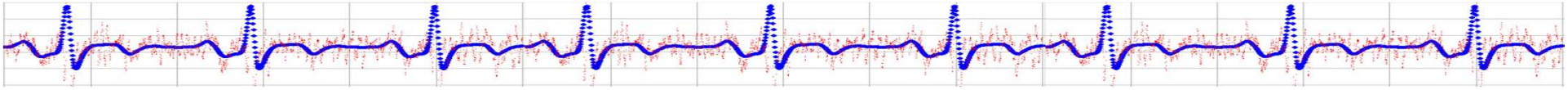
Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Measures of center: Decision rule



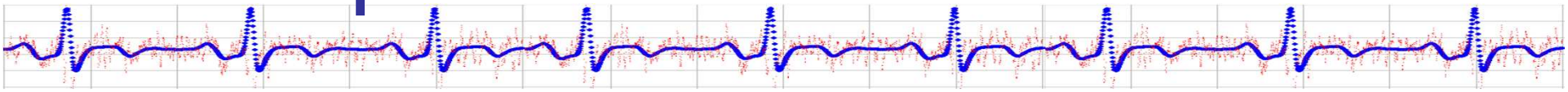
- Nominal
  - Mode
- Ordinal
  - Median
- Interval, Ration & Normal & No Outliers
  - Mean
- Else
  - Median

Measures of spread: Decision rule



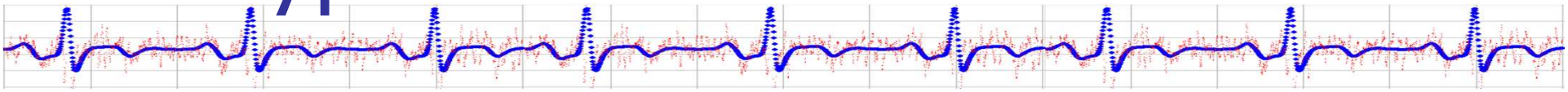
- Nominal, Ordinal
  - No measure of spread
- Interval, Ratio & Normal & no outliers
  - SD
- Else:
  - IQR

Properties of measures



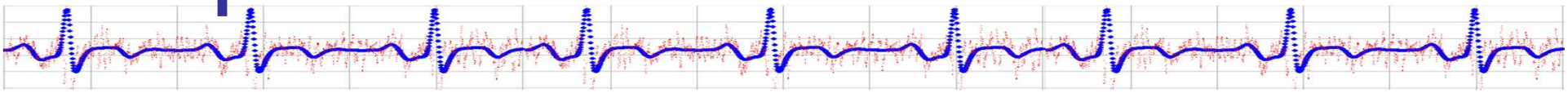
- Ecological validity (extent to which corresponds to real-world)
- Sensitivity
- Range effects (ceiling, floor)

Types of measures



- Behavioral
- Physiological
- Self-report
- Physical
- System
- Implicit

Defining a behavioral protocol



Draft coding manual

Frequency, duration, interval, sampling

Two or more observers code data for pilot subjects

Is Kappa acceptable?

no

Coders review differences & update coding manual

yes

Coding manual provides reliable measure, proceed with study

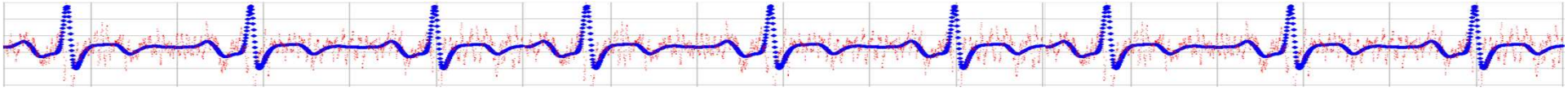
Calculation of Cohen's kappa may be performed according to the following formula:

> 0.7?

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

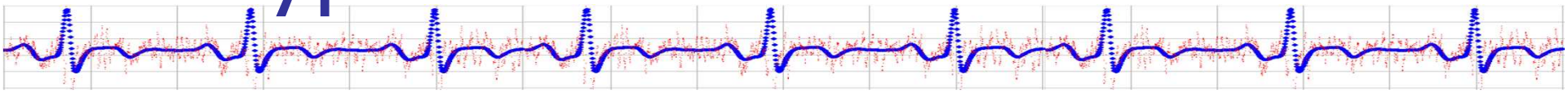
Where Pr(a) represents the actual observed agreement, and Pr(e) represents chance agreement.

Measurement bias



- Reactivity
  - Demand characteristics (performance cues)
  - Role attitude (cooperative, defensive/suspicious, negative)
- Experimenter effects (addressed via blinding)
  - Expectancy bias
  - Treating groups differently
- Manipulation check

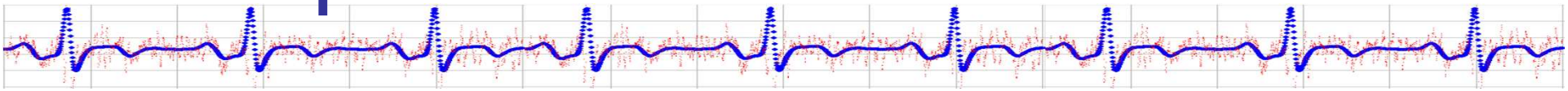
Types of studies



- Univariate (single DV)
- Multivariate (experimental, multiple DV)
- Multivariate (correlational, multiple var)

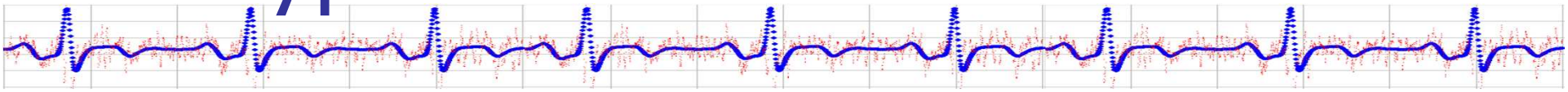
For later: n-factor, n-level from sample midterm: We haven't covered that yet

Experimental considerations



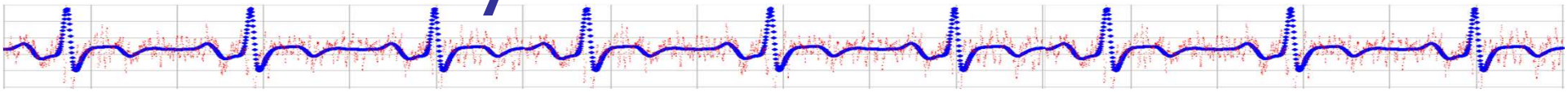
- Control groups
  - Standard of care
  - Non-intervention
  - A vs. B
  - Attention control
  - Placebo control
  - Wait-list contro
- Intention to treat

Types of statistics

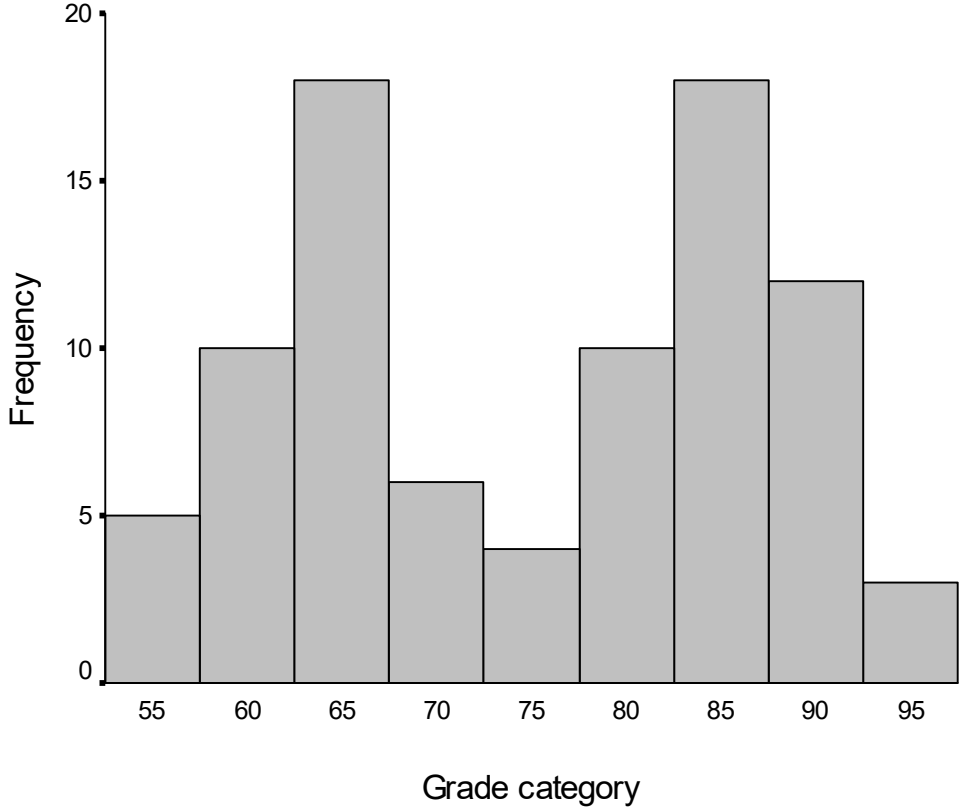


- Pearson correlation
- Chi-square goodness-of-fit
- Chi-square test for independence
- t-test for independent means

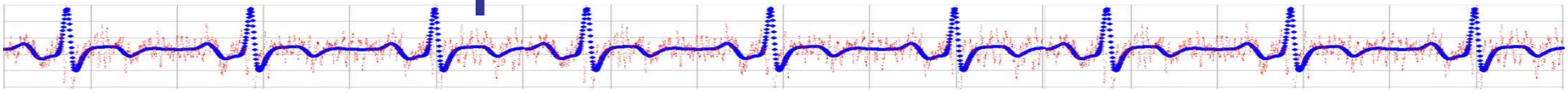
Know your data



- Normal
- Skew
- Uni/bimodal
- Outlier



Descriptive stats



- Every measure
  - 1 stat describing measure of center
  - 0 or 1 stat describing spread
- Be able to compute
  - Mean
  - Median
  - Mode
  - Range
  - StdDev
  - Interquartile-range
  - Histogram

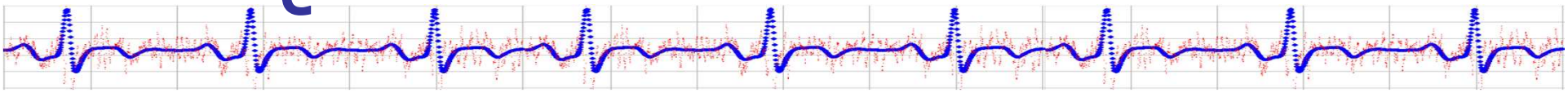
Sample Variance

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

Sample Standard Deviation

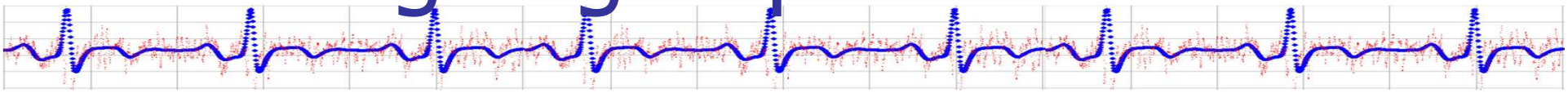
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

IQR



- *Interquartile Range*
  - Compute:
    - Order the data from least to greatest.
    - Find the median.
    - Calculate the median of both the lower and upper half of the data.
    - The IQR is the difference between the upper and lower medians.
  - Less sensitive than the range to extreme scores
  - Used when you want a simple, rough estimate of spread

Designing a questionnaire

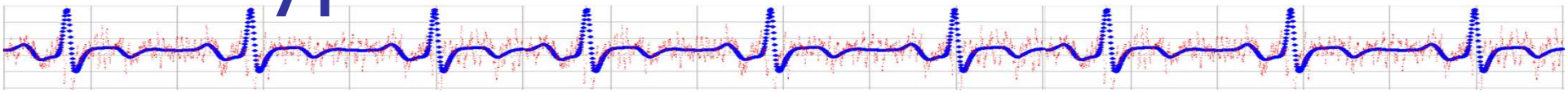


- Items
- Construct / scale
- Factors
- Negated items (helps with response bias)

Rule of thumb:

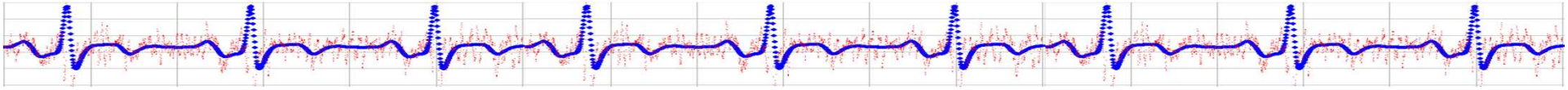
- Use existing measure if exists
- Otherwise, run a pilot study

Types of items



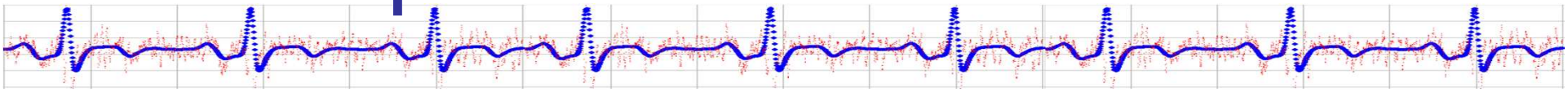
- Rating scale. Consider:
  - Number points
  - Anchors
- Semantic differential scale – bipolar space (e.g., sad 1 2 3 4 5 happy)
- Likert-item (degree of (dis)agreement)
- Visual analog scale

Questionnaire vs. composite measure



- Why composite measure?
  - Complex ideas not captured in single Q
  - Single item not enough variation across people
  - Several items more stable
  - Can impute missing values

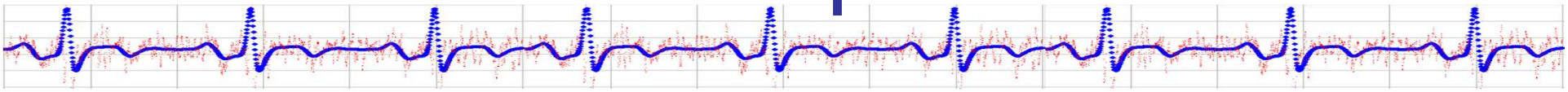
Overall process to develop a composite measure



- Identify factors
- Identify items
- Face and content validity for each item
- Check response variance for each item
  (Check floor/ceiling effects)
- Bi-variate analysis
- Test reliability
- Test validity

Good checklist for assignment!

Chronbach's alpha



- Measure of internal *consistency*
  - Not homogeneity
  - Not unidimensionality (assumes all items measure single dimension … test using factor analysis)
- Function of the number of test items and the average inter-correlation among the items
- Average of all possible split halves

Chronbach's alpha

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$$

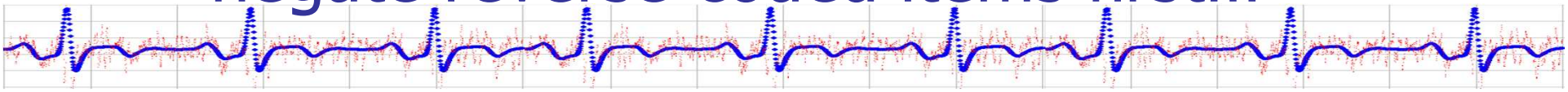
- N = number of items
- c-bar is the average inte
  covariance among the items
- v-bar equals the average variance

("Standardized alpha" based on correlations instead of covariance)

- Increase # items, increase alpha
- If average inter-item correlation is low, alpha will be low. As the average inter-item covariance increases, Cronbach's alpha increases as well (holding the number of items constant).

Cronbach's alpha

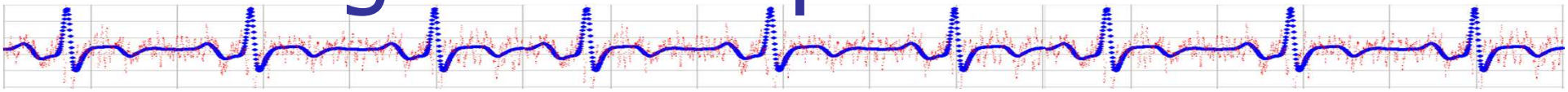
negate reverse-coded items first…



- Scores
  - .00 (no consistency) to 1.0 (perfect)
  - .70 (70% of variance reliable variance)
    Commonly cited as acceptable
    (Something important, might want .90+)

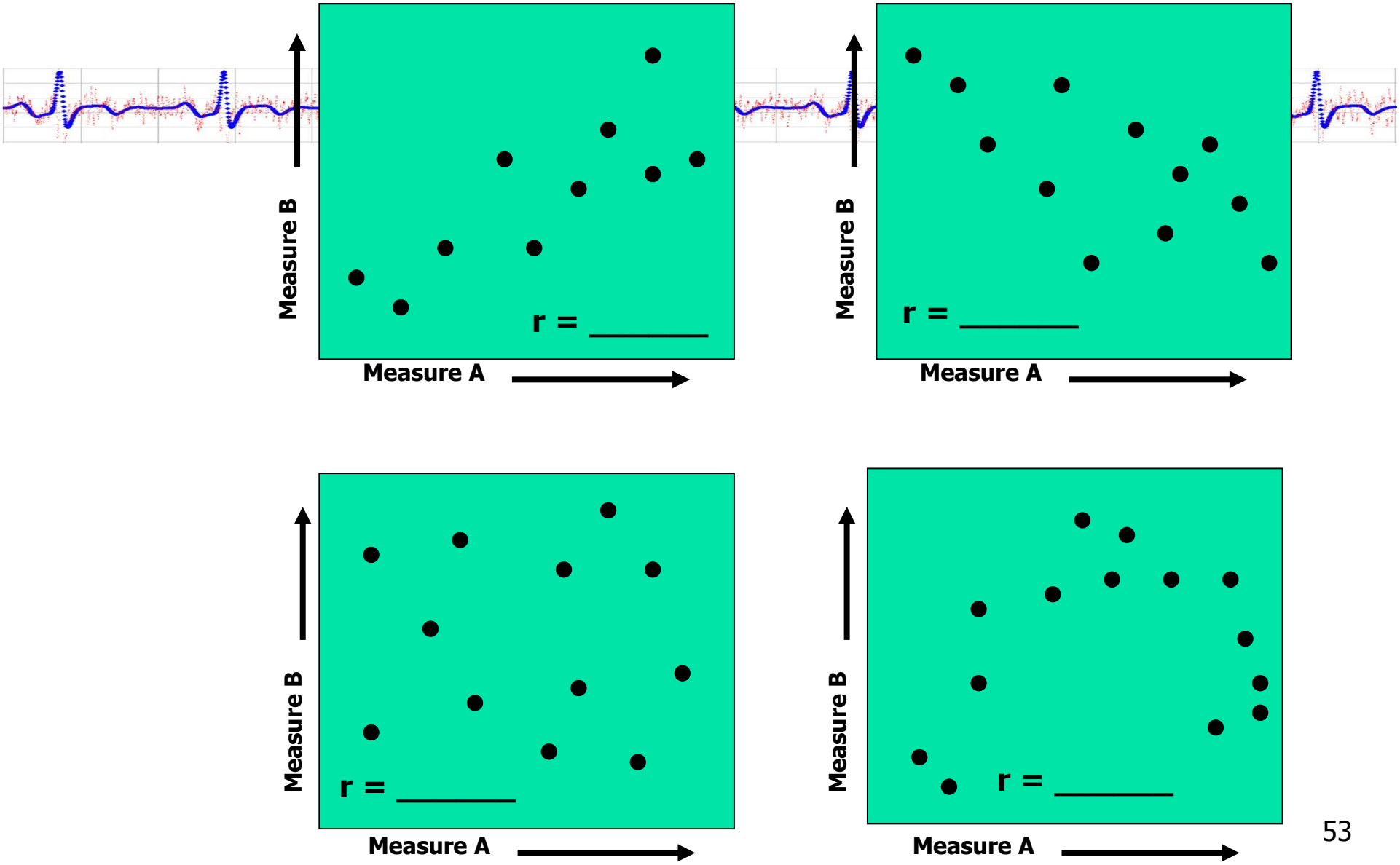
- See Python guide to compute…

Single Likert question?

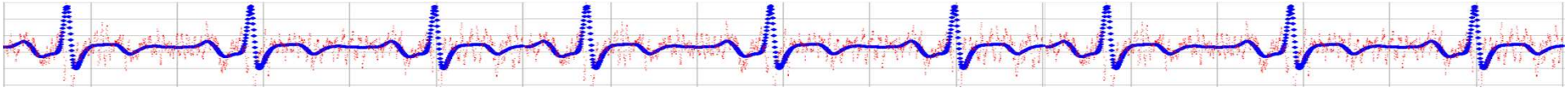


- Interval?

Reminder


Measure B
r = _____
Measure A
Measure B
r = _____
Measure A
Measure B
r = _____
Measure A
Measure B
r = _____
Measure A

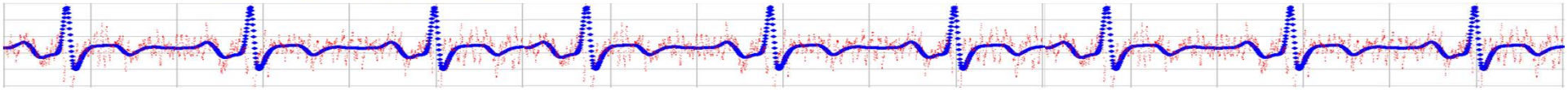
Pearson Correlation Coefficient

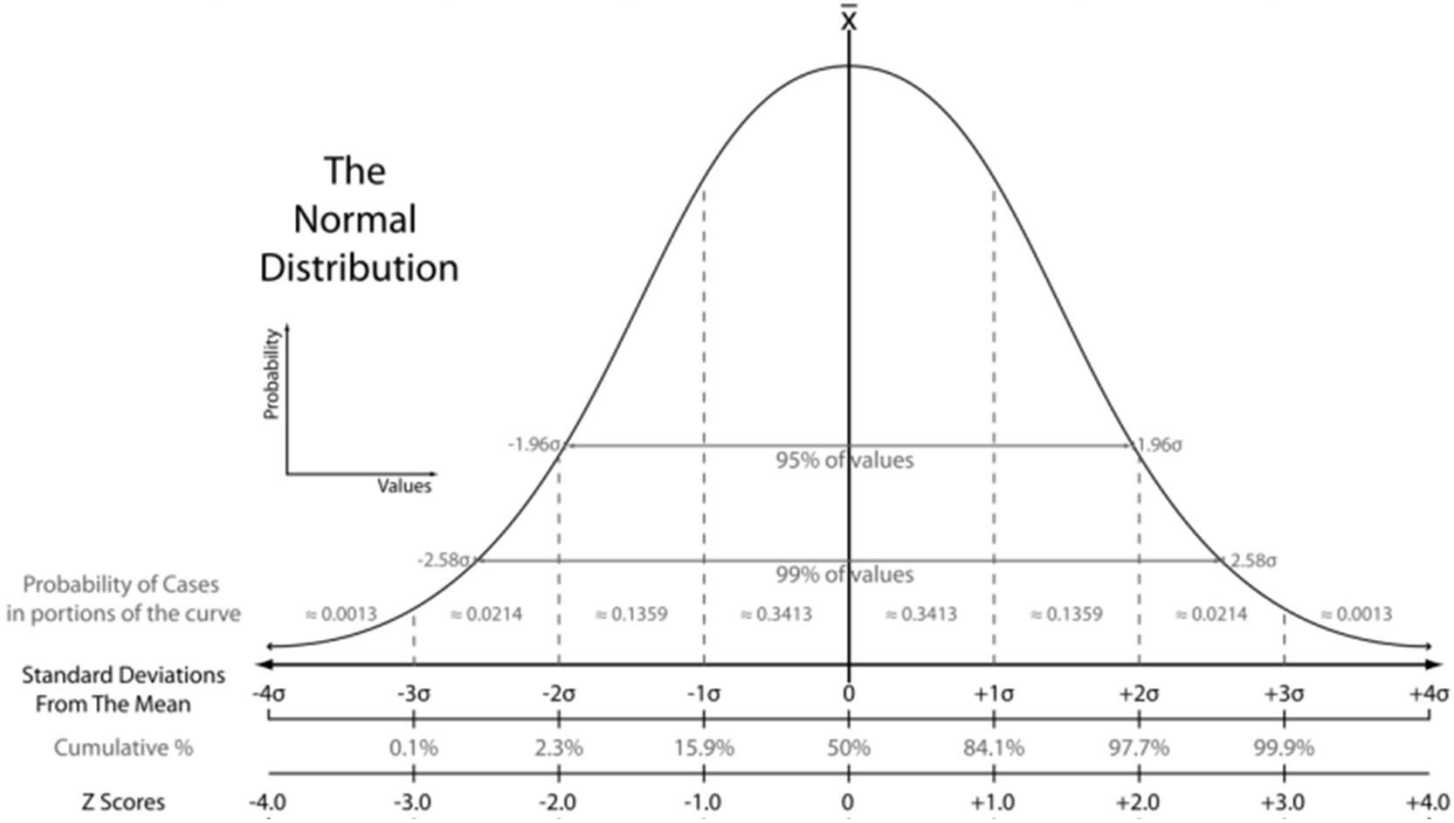


- Assumptions
  1. Two interval (or ratio) measures.
  2. Not an obviously curvilinear relationship.
  3. Both populations normally distributed*.

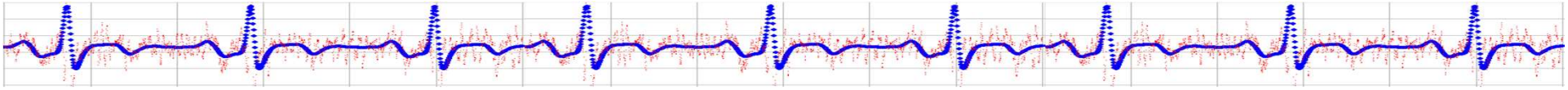
*Unimodal and symmetric frequency distributions. Most important if doing a significance test.

z-score




The Normal Distribution
$\bar{X}$
Probability
Values
-1.96σ
1.96σ
95% of values
-2.58σ
2.58σ
99% of values
Probability of Cases in portions of the curve
≈ 0.0013
≈ 0.0214
≈ 0.1359
≈ 0.3413
≈ 0.3413
≈ 0.1359
≈ 0.0214
≈ 0.0013
Standard Deviations From The Mean
-4σ -3σ -2σ -1σ 0 +1σ +2σ +3σ +4σ
Cumulative %
0.1% 2.3% 15.9% 50% 84.1% 97.7% 99.9%
Z Scores
-4.0 -3.0 -2.0 -1.0 0 +1.0 +2.0 +3.0 +4.0

Formula for Pearson correlation



$$r = \frac{\sum Z_X Z_Y}{N-1}$$

-OR-

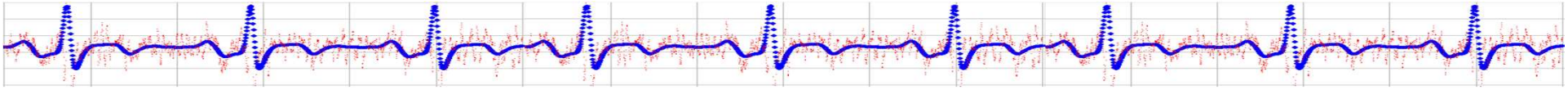
$$r = \frac{\sum [(X - M_X)(Y - M_Y)}{\sqrt{(SS_X)(SS_Y)}}$$

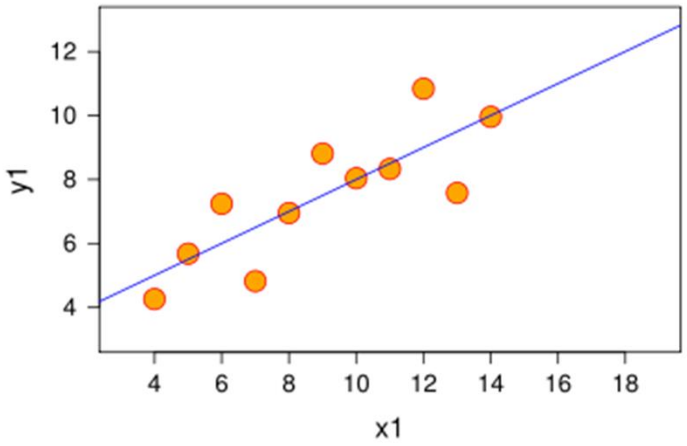
Where

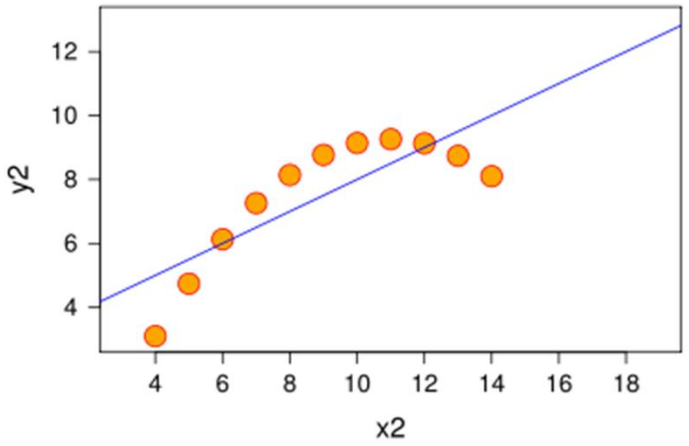
$$SS_X = \sum (X - M_X)^2$$

$$SS_Y = \sum (Y - M_Y)^2$$

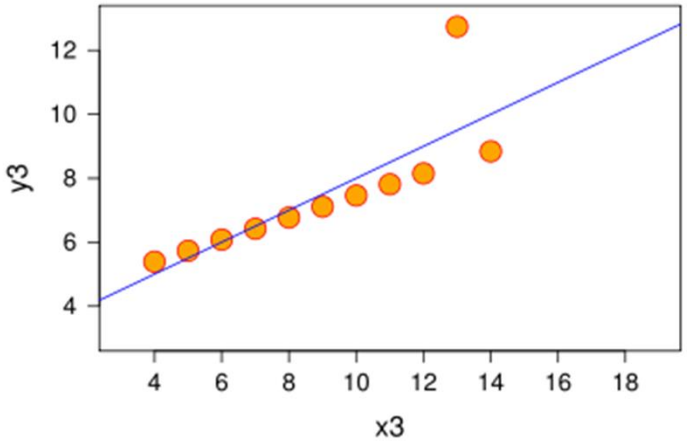
Correlation?

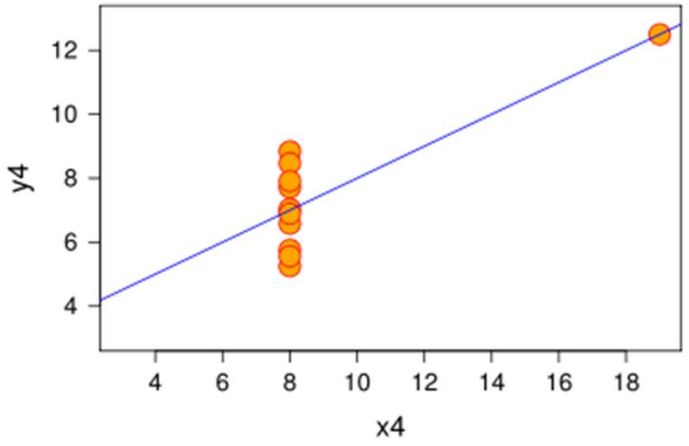




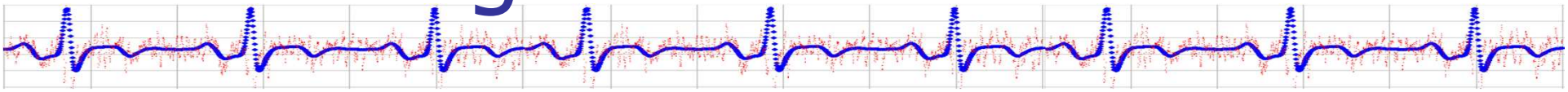


All
r=0.81



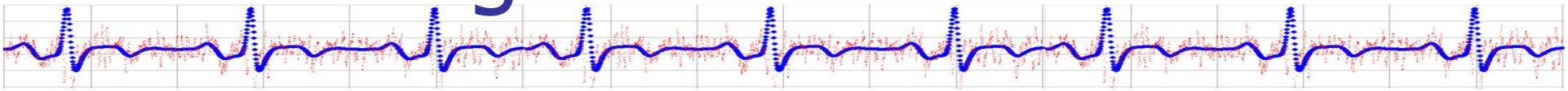


Procedure for Hypothesis Testing with Correlations



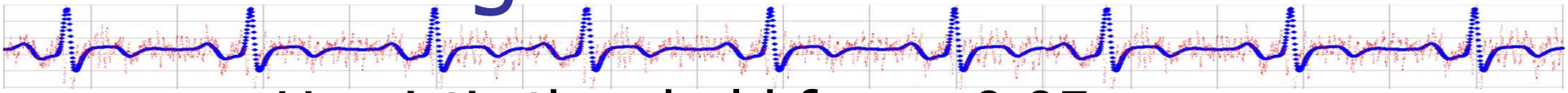
- Populations being compared:
  - ***Test:*** The population from which the observed sample was drawn.
  - ***Comparison:*** A hypothetical population in which the variables are unrelated, i.e., have a correlation of zero.

Procedure for Hypothesis Testing with Correlations



- Form of hypothesis H1?
  - The correlation in the observed population is different from a population in which the correlation is zero.
  - Unlikely we would have obtained a correlation this big if the variables actually were unrelated.
- Form of null hypothesis H0?
  - The correlation in the observed population is the same as a population in which the correlation is zero.

Procedure for Hypothesis Testing with Correlations

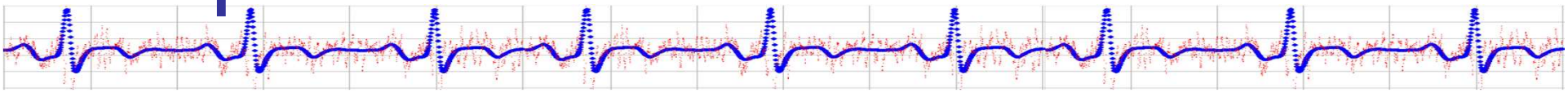


- Heuristic threshold for α=0.05:

$$r > \frac{2}{\sqrt{N}}$$

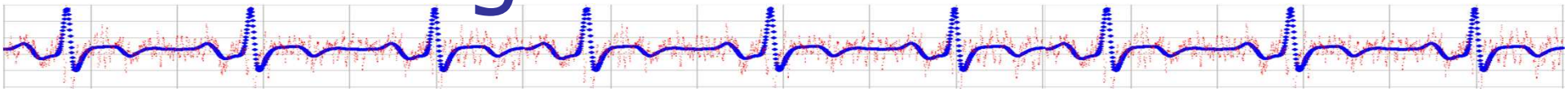
- *Exact form given in Aron, or in R.*

p-value and correlation



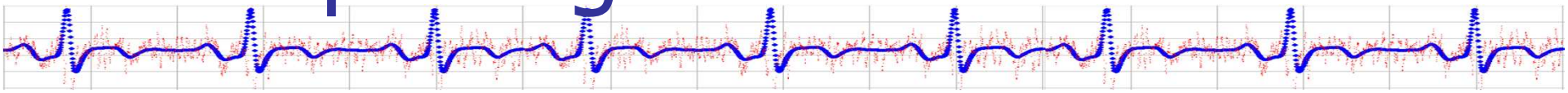
If a correlation coefficient has been determined to be statistically significant this does *not* mean there is a strong association. It simply tests the null hypothesis that there is no relationship. By rejecting the null hypothesis, you accept the alternative hypothesis that states that there is a relationship, but with no information about the strength of the relationship or its importance!

Procedure for Hypothesis Testing with Correlations



- R: (looking for Python turnkey equivalent)
  - cor.test(v1,v2)
  - See if significance<threshold
    - Yes => reject H0
    - No => inconclusive
- Manually:
  - Compute r
  - Is $r > \frac{2}{\sqrt{N}}$
    - If yes => reject H0
    - If no => inconclusive

Reporting results



r=*val*, *p<sigthresh*

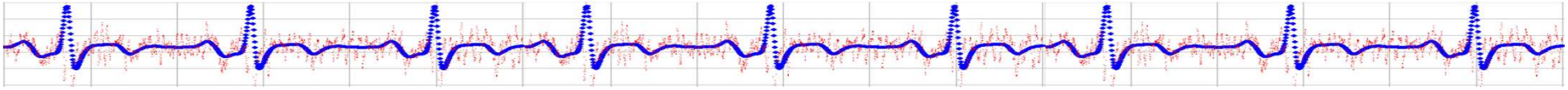
Where,

- sigthresh = pre-defined significance threshold
  - Note: if p<<sigthresh, can report that as well, e.g., "p<.01", "p=.001"

For example: **r=0.82, p<.05**

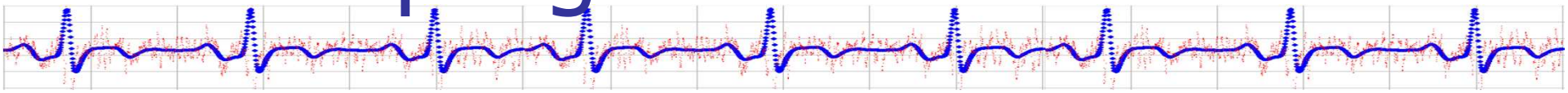
If not significant, than use "n.s." instead of "p<…".

Other measures of association



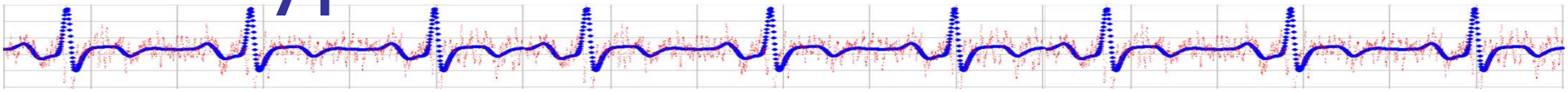
- Point-biserial
  - One numeric & one binary (nominal) measure
  - Just dummy code the nominal (0 and 1) and use Pearson correlation.
- Spearman Rank Order (rho)
  - Two ordinal measures (or for transformed numeric measures if non-linear)
  - Replace each value with its rank order
  - Compute Pearson correlation with ranks
  - Measures degree of monotonicity

Sampling



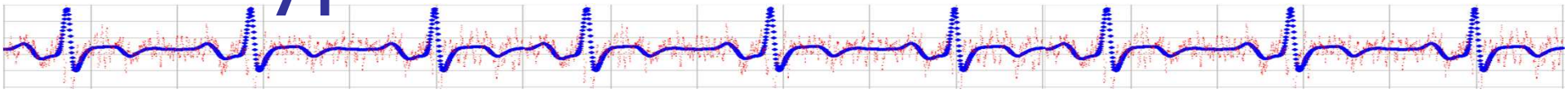
- Convenience
- Simple
- Systematic
- Stratified
- Proportionate
- Cluster

Type I error



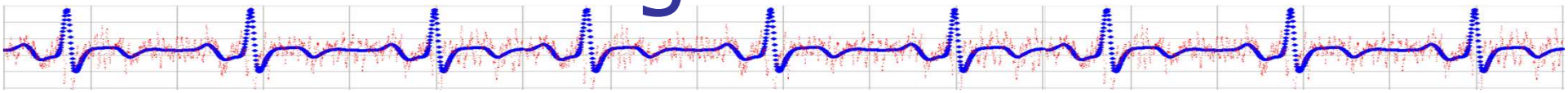
- Rejection of a true null hypothesis (also known as a "false positive" finding)
- Often represented by the Greek letter alpha (α)

Type II error



- Failure to reject a false null hypothesis (also known as a "false negative" finding)
- Often represented by the Greek letter beta **(**β **)**

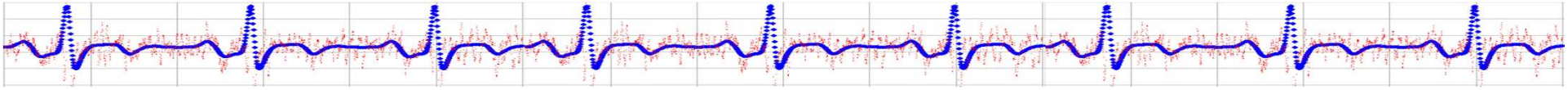
Level of significance



Probability of rejecting a null hypothesis by the test when it is really true, which is denoted as α. That is, P(Type I error) = α.

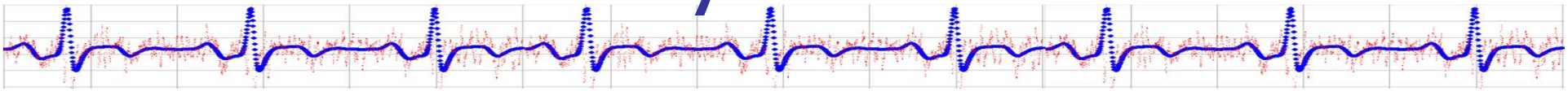
The level of significance 0.05 is related to the 95% confidence level

Power



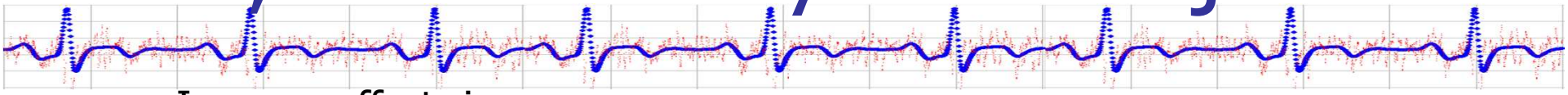
- Probability that a test will reject the null hypothesis when it is, in fact, false.
- 1 – β (type II error rate)
- High power is desirable. Like β, power can be difficult to estimate accurately, but increasing the sample size always increases power

Power Analysis



- Should determine number of subjects you need ahead of time by doing a 'power analysis'
- Standard procedure (part of your study plan):
  - Determine statistic you will use
  - Fix alpha and beta (1-power) (and number of tails if appropriate)
  - Estimate expected effect size from prior studies
  - Then: Determine number of subjects you need
- Note: Power
  - Increases with effect size
  - Increases with sample size
  - Decreases with decreasing alpha

But, I can't study 786 subjects!



- Increase effect size
  - Increase difference in population means (change manipulation)
  - Decrease population variance (better measures, control more extraneous vars)
  - Redesign study to collect many trials of measures per subject
- Relax criteria for Type I error
  - Increase α threshold
  - Change from Two-tailed => one-tailed test
  - *Decreases credibility of your findings*
- Decrease power
  - *Decreases likelihood of getting a significant result*
- Use a different statistic
  - *If possible, maybe consult a statistician*

- Practically
  - Usually, redesign experiment so that we have increased effect size or better measures for decreased variance
  - OR, call it a "pilot study"

Type I and Type II errors

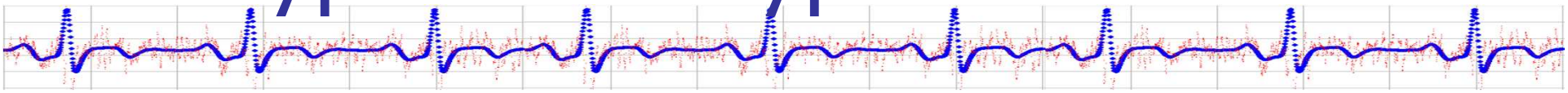
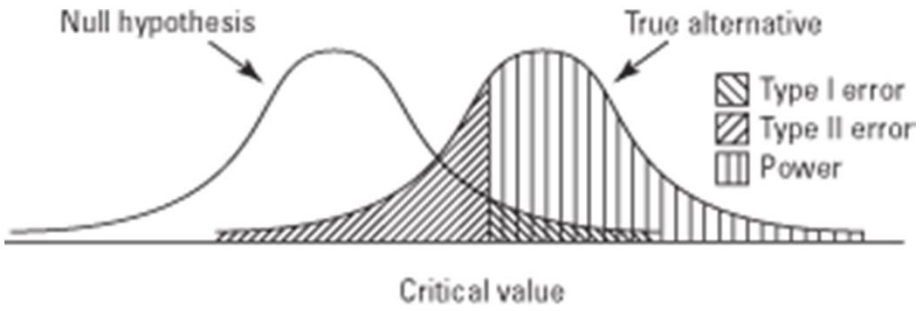
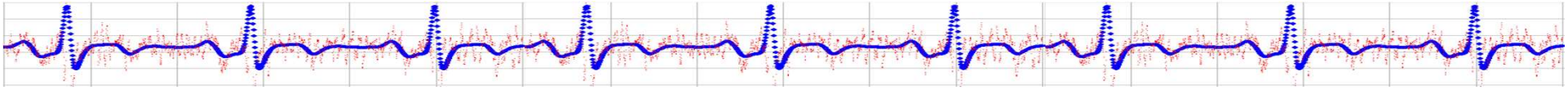


Table 1. Types of Statistical Errors

	H_0 is actually:	
	True	**False**
Reject H_0	Type I error	Correct
Accept H_0	Correct	Type II error



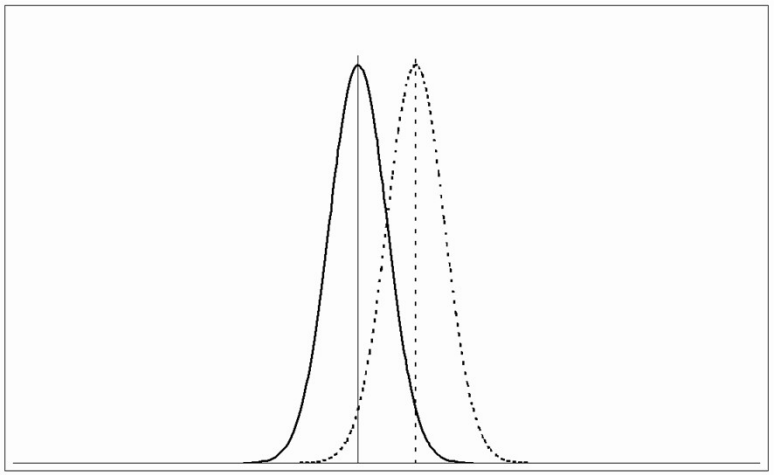
Effect size

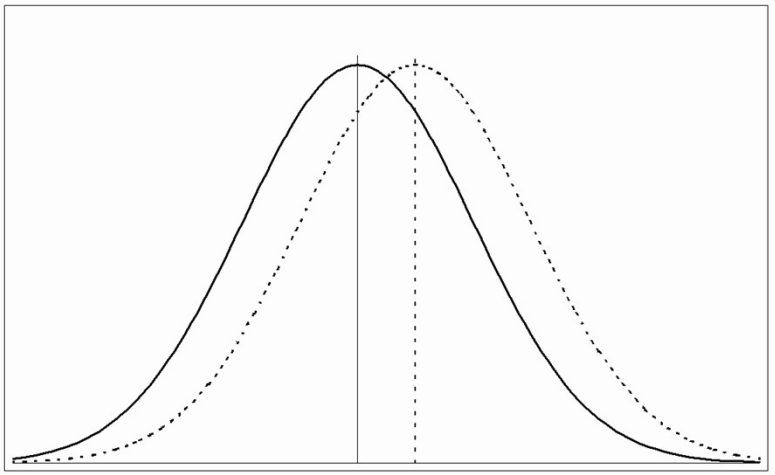


- Measures strength of the relationship between two variables on a numeric scale
- Effect size is usually measured in one of three ways: (1) standardized mean difference, (2) odd ratio, (3) correlation coefficient

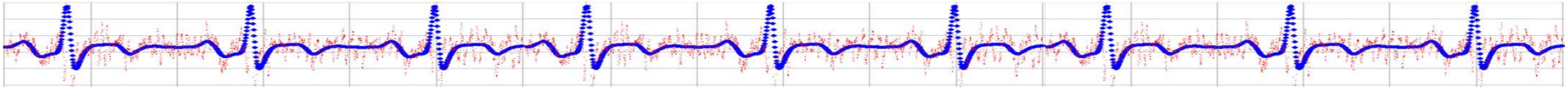
Effect size mean

$$\text{Effect Size} = \frac{\text{[Mean of experimental group]} - \text{[Mean of control group]}}{\text{Standard Deviation}}$$





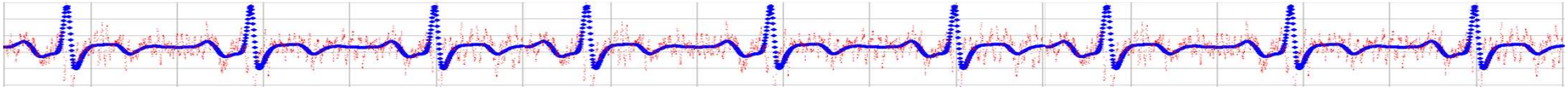
Effect size: correlation



Estimate the amount of the variance within an experiment that is "explained" or "accounted for" by the experiment's model

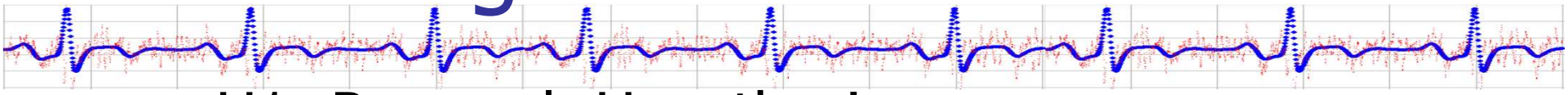
Can use Pearson’s correlation

Pearson's correlation



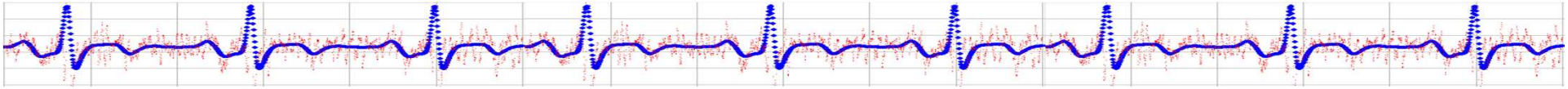
- r for effect size
- -1 to 1
- Guidelines:
  - Small 0.1
  - Medium 0.3
  - Large: 0.5

Basic Process of Hypothesis Testing



- H1: Research Hypothesis:
  - Population 1 is different than Population 2
- H0: Null Hypothesis:
  - No difference between Pop 1 and Pop 2
  - *The difference is "null"*
- Compute *p(observed difference|H0)*
  - 'p' = probability observed difference is due to random variation
- If p<*threshold* then reject H0 => accept H1
  - p typically set to 0.05 for most work
  - p is called the "level of significance"

Formula for Chi-square statistic

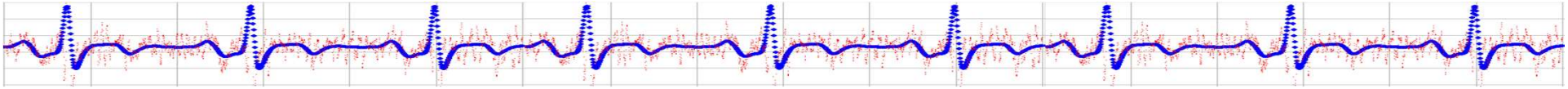


$$X^2 = \sum \frac{(O-E)^2}{E}$$

- O = Observed frequency for a given category
- E = Expected frequency for a given category

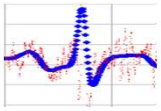
- Note: "statistic" is a function you apply to a set of data (in a statistical analysis)

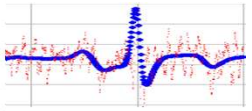
Computing Chi-square Goodness of Fit

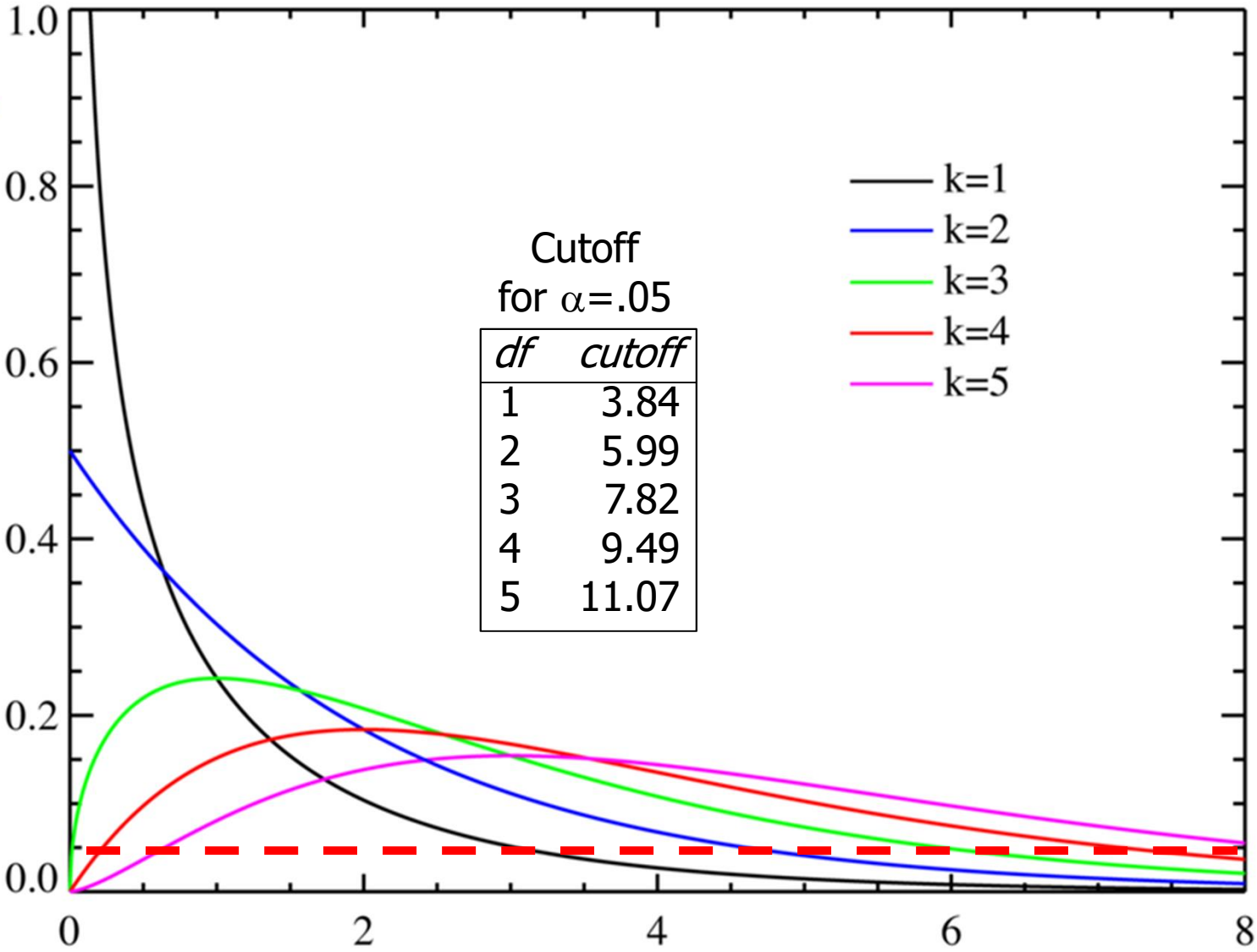


- Manually:
  - Determine df (= num categories – 1)
  - Compute Chi-square using formula
  - Lookup to see if statistic>table entry for significance,df
    - If yes => reject H0
    - If no => inconclusive

Chi-square probability distribution



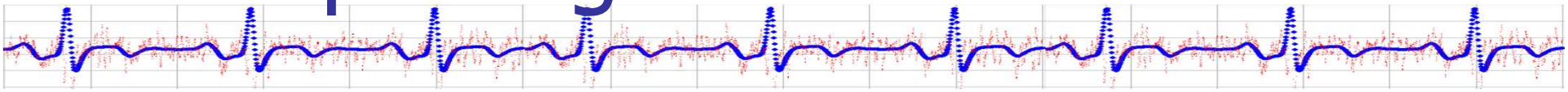




Cutoff for α=.05

df	*cutoff*
1	3.84
2	5.99
3	7.82
4	9.49
5	11.07

Reporting result



$$X^2(df) = chisq, p < sigthresh$$

Where,

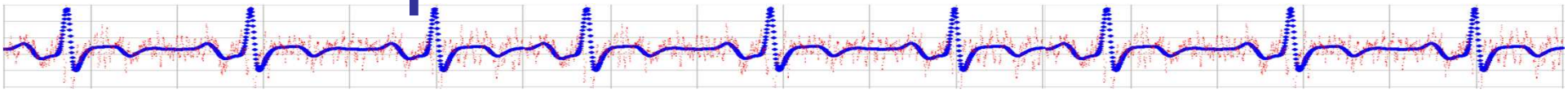
- df = degrees of freedom
- sigthresh = pre-defined significance threshold
  - Note: if p<<sigthresh, can report that as well, e.g., "p<.01", "p=.001"

For example: $X^2(2) = 11.89, p < 0.05$

If not significant, than use "n.s." instead of "p<…".

Usually also report expected and actual frequencies, or at a minimum, the total number of cases considered (aka "n").

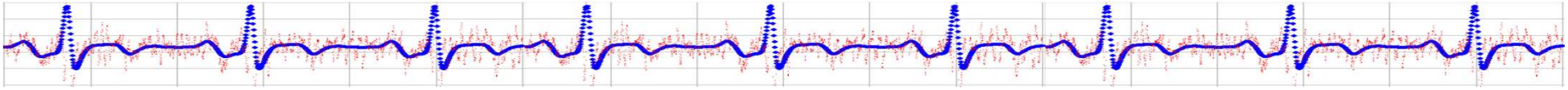
Chi-Square Test for Independence



- Are two variables related (H1), or are they independent (H0)?

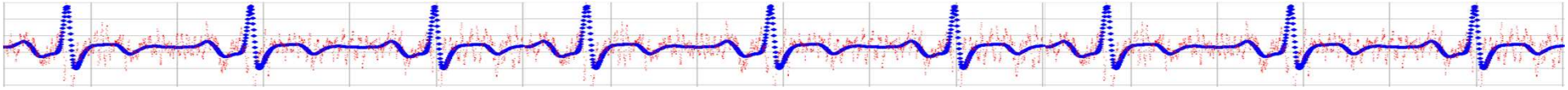
- Assumptons
  - Both variables must be nominal.
  - Cannot be related in a 'special' way (i.e., repeated measures)
  - Random sampling assumed

Review



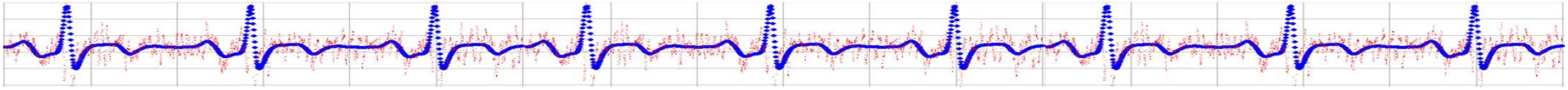
- Kinds of analyses for descriptive studies?
  - Descriptive
  - Chi-square goodness of fit [nominal]

Review



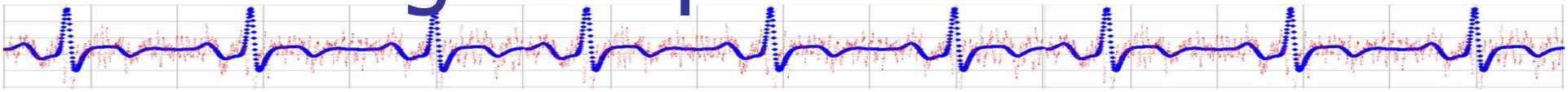
- Kinds of analyses for correlational studies?
  - Descriptive
  - Chi-square goodness of fit [nominal]
  - Chi-square test for independence [nominal/nominal]
  - Correlation [numeric/numeric]
    - Point-biserial [numeric/nominal]
    - Spearman rho [ordinal/ordinal], [ordinal,numeric], or [numeric,numeric and not linear or not normal]

Review



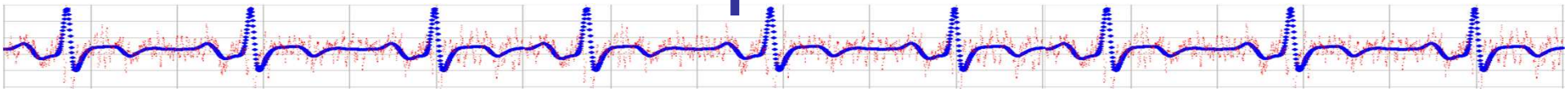
- Kinds of analyses for experimental studies?
  - Descriptive
  - Chi-square goodness of fit [nominal]
  - Correlation (large number of IV values - *parametric*) [numeric/numeric] (atypical)
  - Chi-square test for independence [nominal/nominal]
  - t-test [nominal/numeric]

The grand plan



- X^2 tests
  - For nominal measures
  - Can apply to a single measure
- Correlation tests
  - For two numeric measures
- t-test for independent means
  - For categorical IV, numeric DV

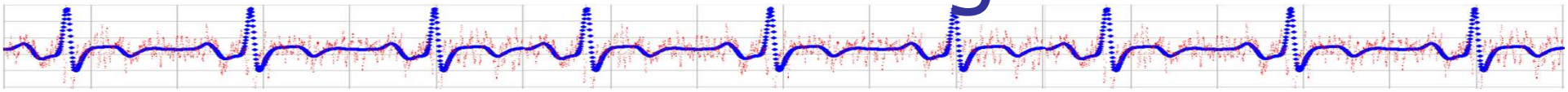
Midterm Prep



- Every test has a long question of the following form:

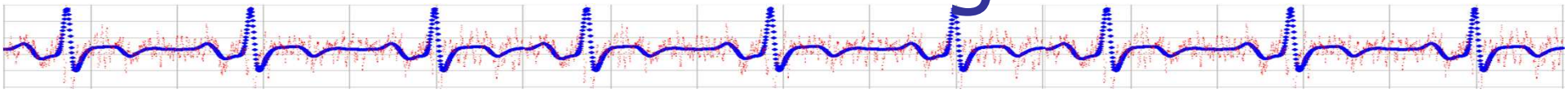
Study Proposal (25%). Sketch a study proposal to prove which search engine is best (Google or Bing) for individuals who have never used a computer before. Your primary outcome measure is learnability (from Nielsen).

What a sketch might have



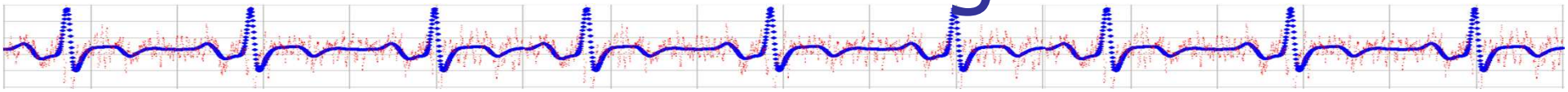
- Hypotheses
  - Null hypothesis
  - Research hypotheses
- Study design
  - 2 group, between subjects
  - Diagram
- Measures
  - Type
  - Reliable? Valid? (How do you know)?

What a sketch might have



- Participants
  - Sampling/recruitment
  - Eligibility
    - Inclusion/exclusion
    - Power analysis
  - Ethics
- Protocol
  - Standardized tasks
  - Randomization

What a sketch might have



- Analysis plan
  - Descriptive statistics
  - Inferential statistics
    - Test type (Pearson r, Chi-Square, t-test, …)
    - Test criteria (alpha, two-tailed, DoF, …)
- Summary: what might be learned?