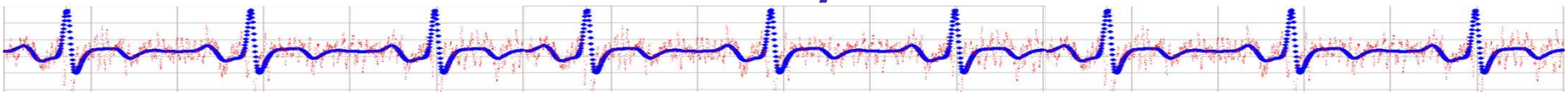
Empirical Research Methods in Information Science

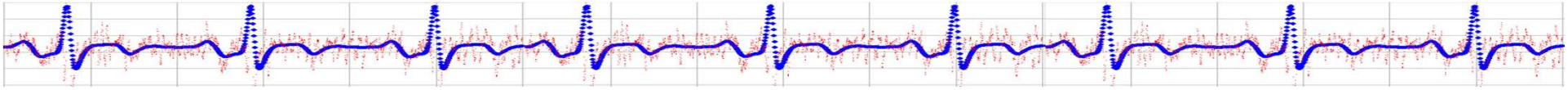
IS 4800 / CS6350



Lecture 9

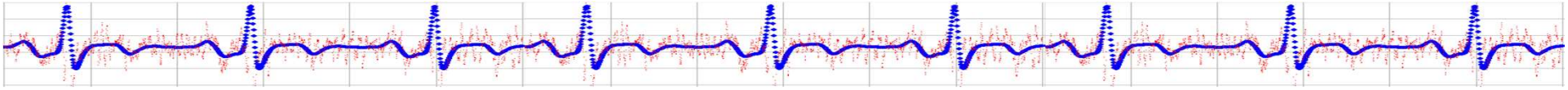
Descriptive statistics, observing behavior, survey design

Outline



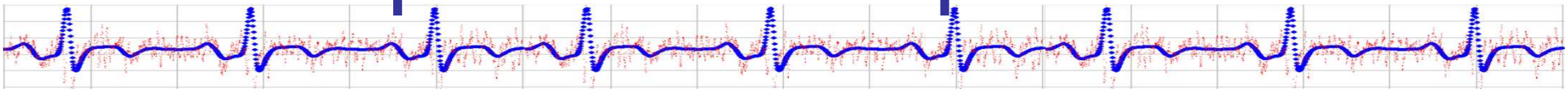
- Reading assessment
- Homework I3 – thoughts?
- Homework I4 – plan
- Descriptive statistics
- Observing behavior
- Survey/instrument design

Homework I3



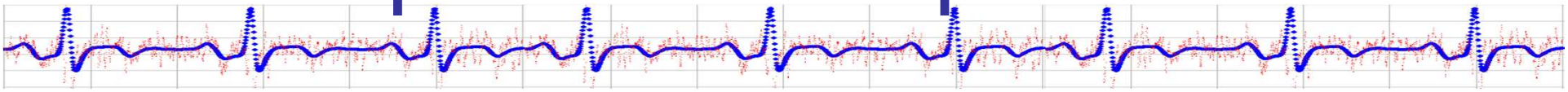
- Usability/performance measures
- What struck you?

Homework I4: Design a composite self-report measure



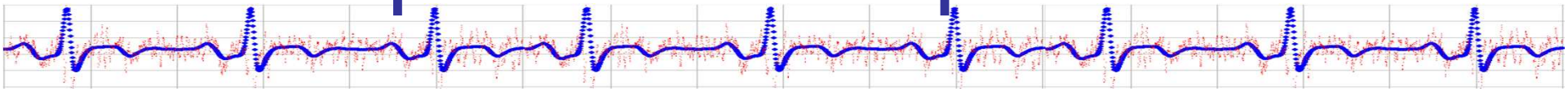
- Design a new composite self-report measure to assess a person's "homework procrastination" (or something else)
- Assume it only has one factor, but use at least five scale items
- Incorporate information from at least one literature reference

Homework I4: Design a composite self-report measure



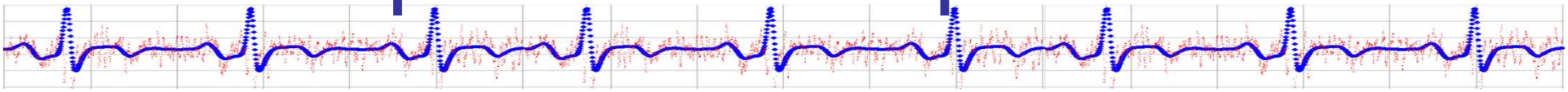
- Assess the face and content validity of your measure and work through a bivariate analysis of your items
- Implement questionnaire on surveymonkey.com or Google forms

Homework I4: Design a composite self-report measure



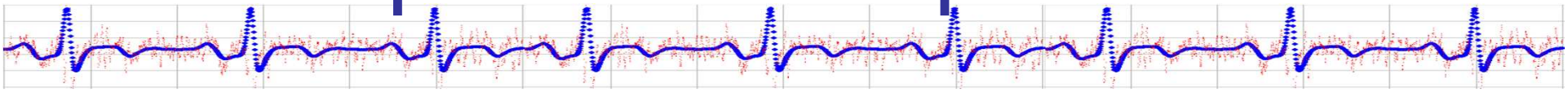
- Decide on one method for assessing validity (besides face & content) for your measure that you can also assess in a self-report questionnaire. This should be an additional question (or an additional previously validated composite measure) on your survey and should provide a numeric measure

Homework I4: Design a composite self-report measure



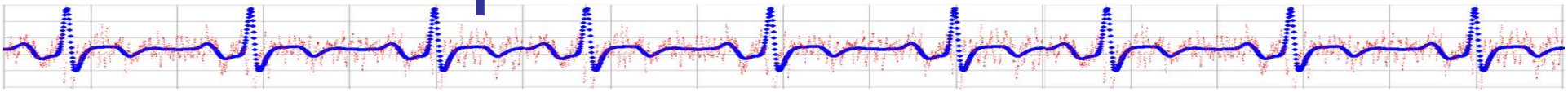
- Post your questionnaire on Piazza
- You are obligated to reply to any questionnaires posted within 48 h!
- Compute the reliability (internal consistency) of your measure using Python
- Compute descriptive statistics for your measure and any other items you may have included on the questionnaire

Homework I4: Design a composite self-report measure



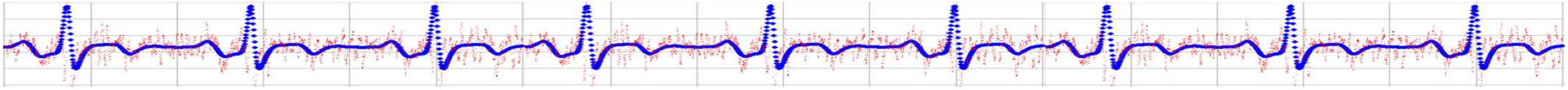
- Assess the validity of your measure (you can do this qualitatively, e.g., using scatterplots)
- Document and submit all of the above
- You may work individually or in teams of two
- Due 2/20

Descriptive statistics



- Statistic = a number used to describe some feature of a group of measurements
- Class rule: For every measure you must have
  - Exactly one statistic describing a measure of center
  - Zero or one statistic describing a measure of spread

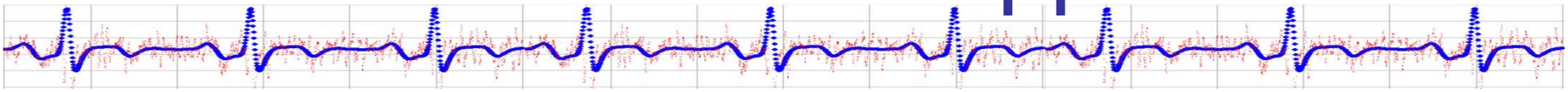
Measures of center



- Mean
- Median
- Mode

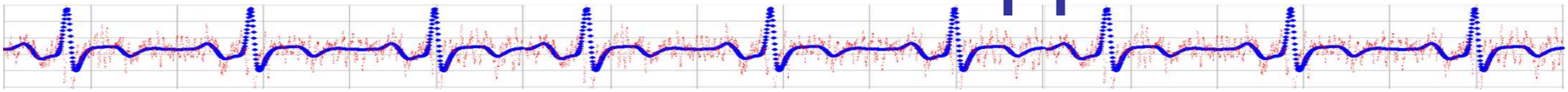
- Whazzit?
- When to use?

Measures of center: Characteristics & applications



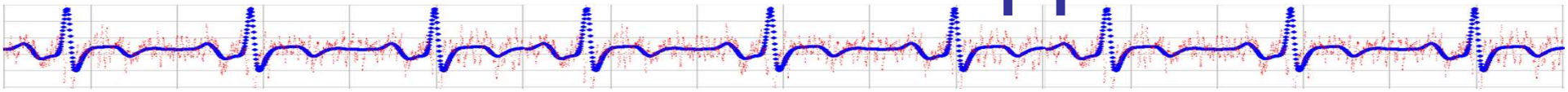
- *Mode*
  - Most frequent score in a distribution
  - Simplest measure of center
  - Scores other than the most frequent not considered
  - Limited application and value

Measures of center: Characteristics & applications



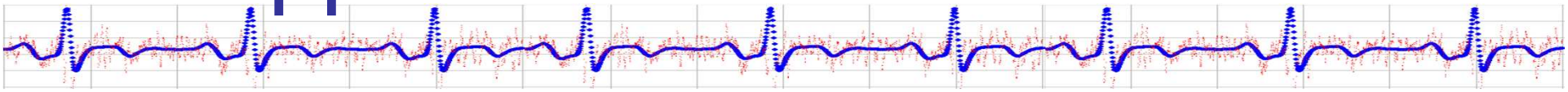
- *Median*
  - Central score in an ordered distribution
  - More information taken into account than with the mode
  - Relatively insensitive to outliers
  - Prefer when data is skewed
  - Used primarily when the mean cannot be used

Measures of center: Characteristics & applications



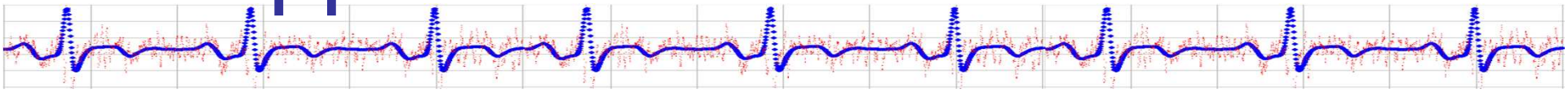
- *Mean*
  - Average of all scores in a distribution
  - Value dependent on each score in a distribution
  - Most widely used and informative measure of center

Measures of center: Applications



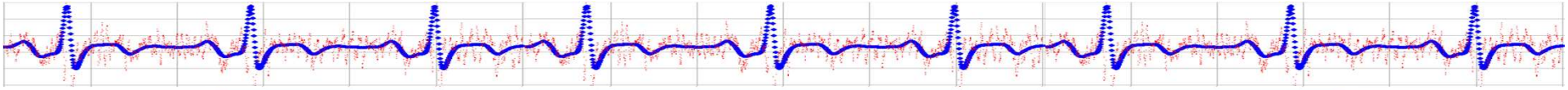
- *Mode*
  - Used if data are measured along a nominal scale
- *Median*
  - Used if data are measured along an ordinal scale
  - Used if interval or ratio data do not meet requirements for using the mean (skewed but unimodal), or if significant outliers

Measures of center: Applications



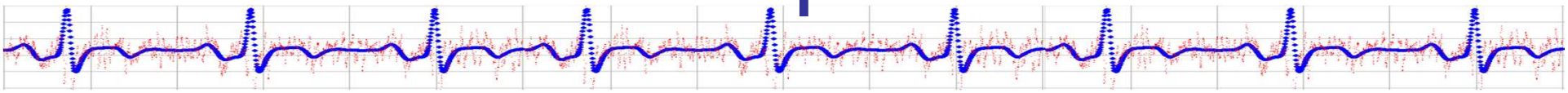
- *Mean*
  - Used if data are measured along an interval or ratio scale
  - Most sensitive measure of center
  - Used if scores are normally distributed

Measurement scales



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"				✔

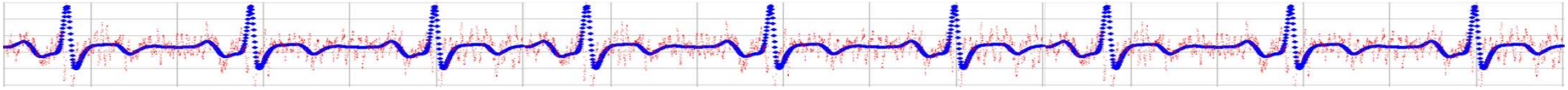
Measures of spread



- Range
- Inter-quartile range
- Standard deviation

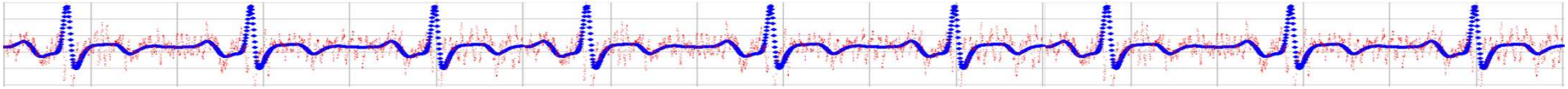
- Whazzit?
- When to use?

Measures of spread: Characteristics



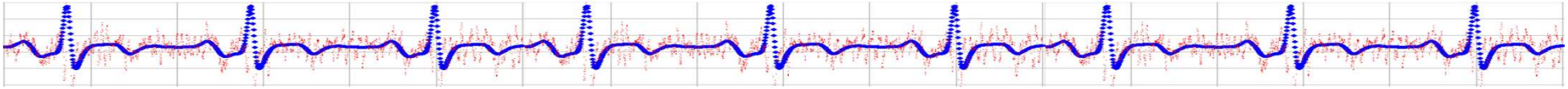
- *Range*
  - Subtract the lowest from the highest score in a distribution of scores
  - Simplest and least informative measure of spread
  - Scores between extremes are not taken into account
  - Very sensitive to extreme scores

Measures of spread: Characteristics



- *Range*
- *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

Measures of spread: Characteristics



- *Variance*
  - Average squared deviation of scores from the mean
  - Divide sum of squared deviations by N or N-1?
- *Standard Deviation*
  - Square root of the variance
  - Most widely used measure of spread

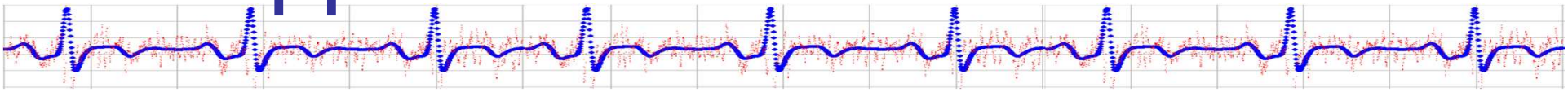
Sample Variance

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

Sample Standard Deviation

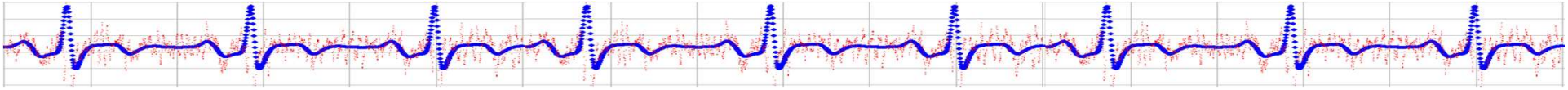
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Measures of spread: Applications



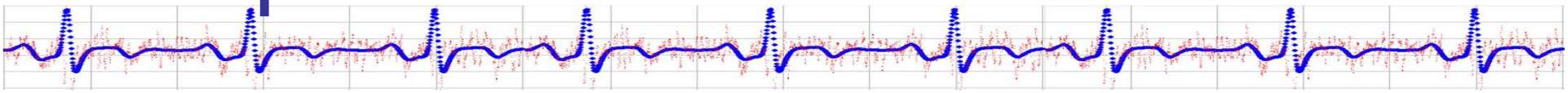
- The range and standard deviation are sensitive to extreme scores ("outliers")
- When your distribution of scores is skewed, the standard deviation does not provide a good index of spread

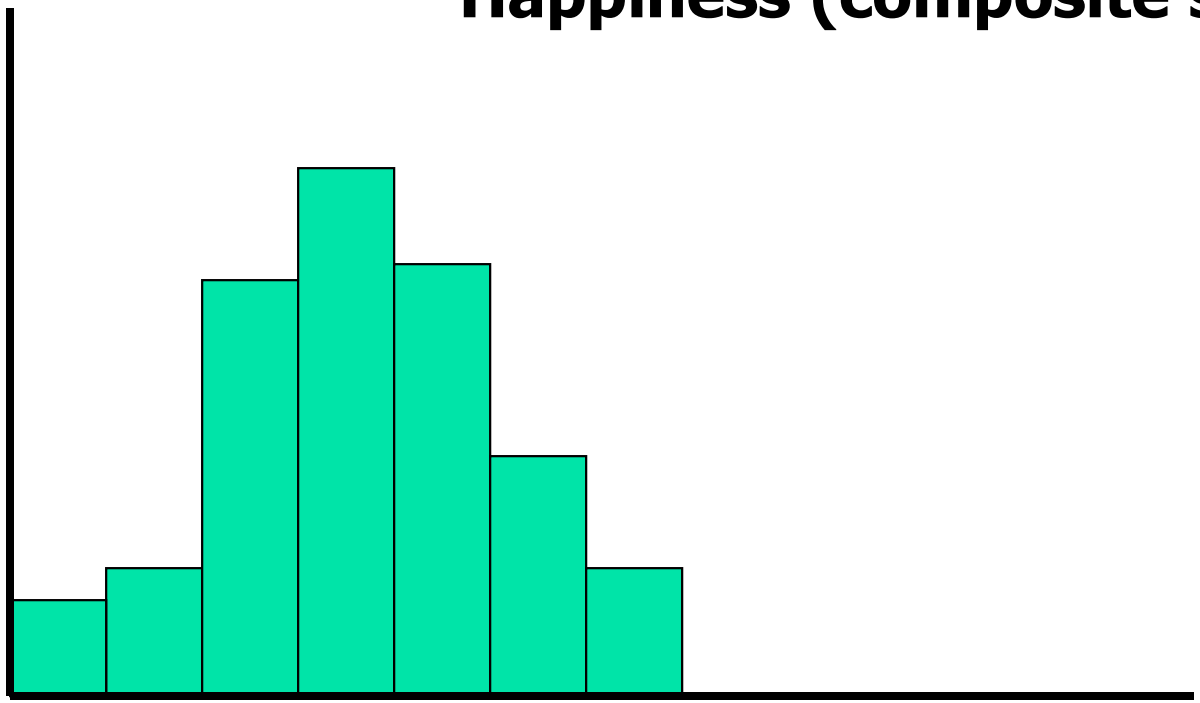
Measures of spread: Decision rule



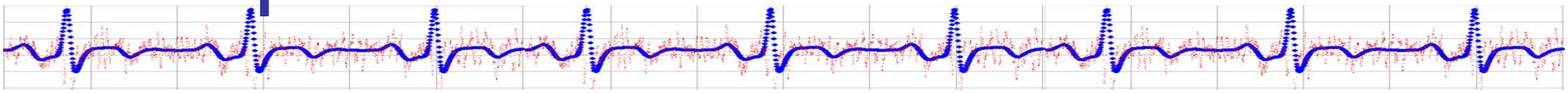
- Nominal, ordinal => no measure of spread
- Interval, ratio & normal & no outliers => SD
- Else: IQR

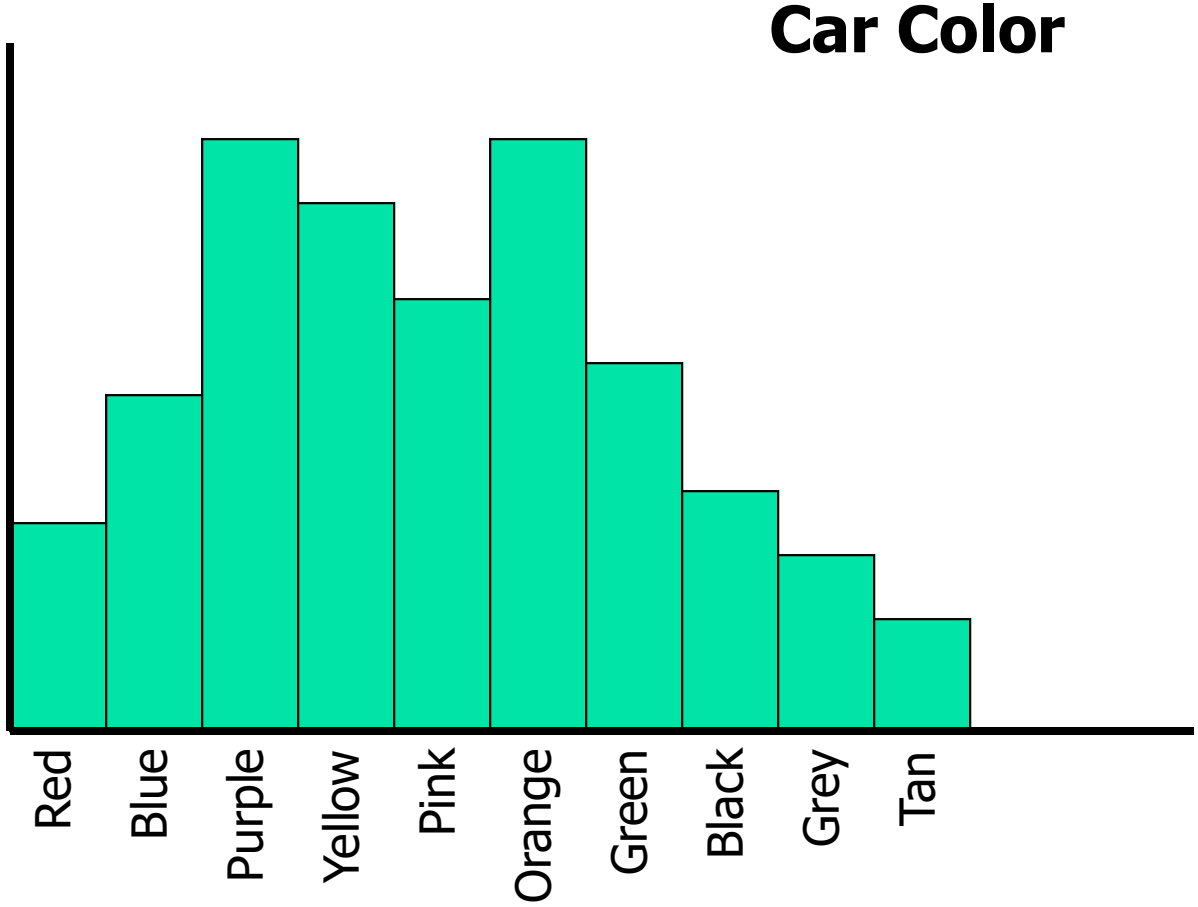
Which measures of center and spread?



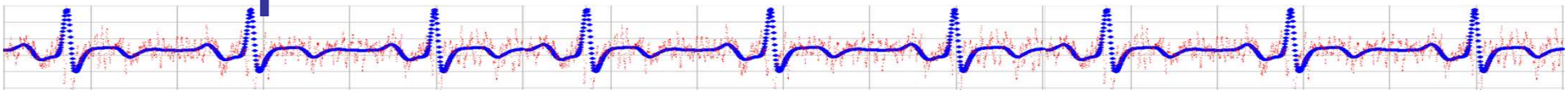


Which measures of center and spread?

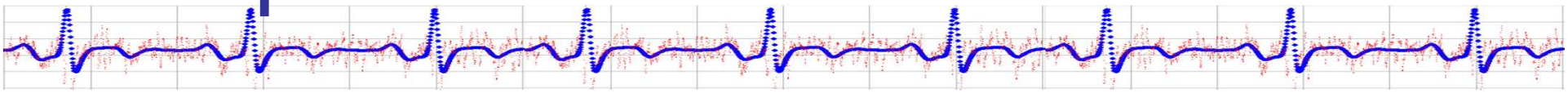


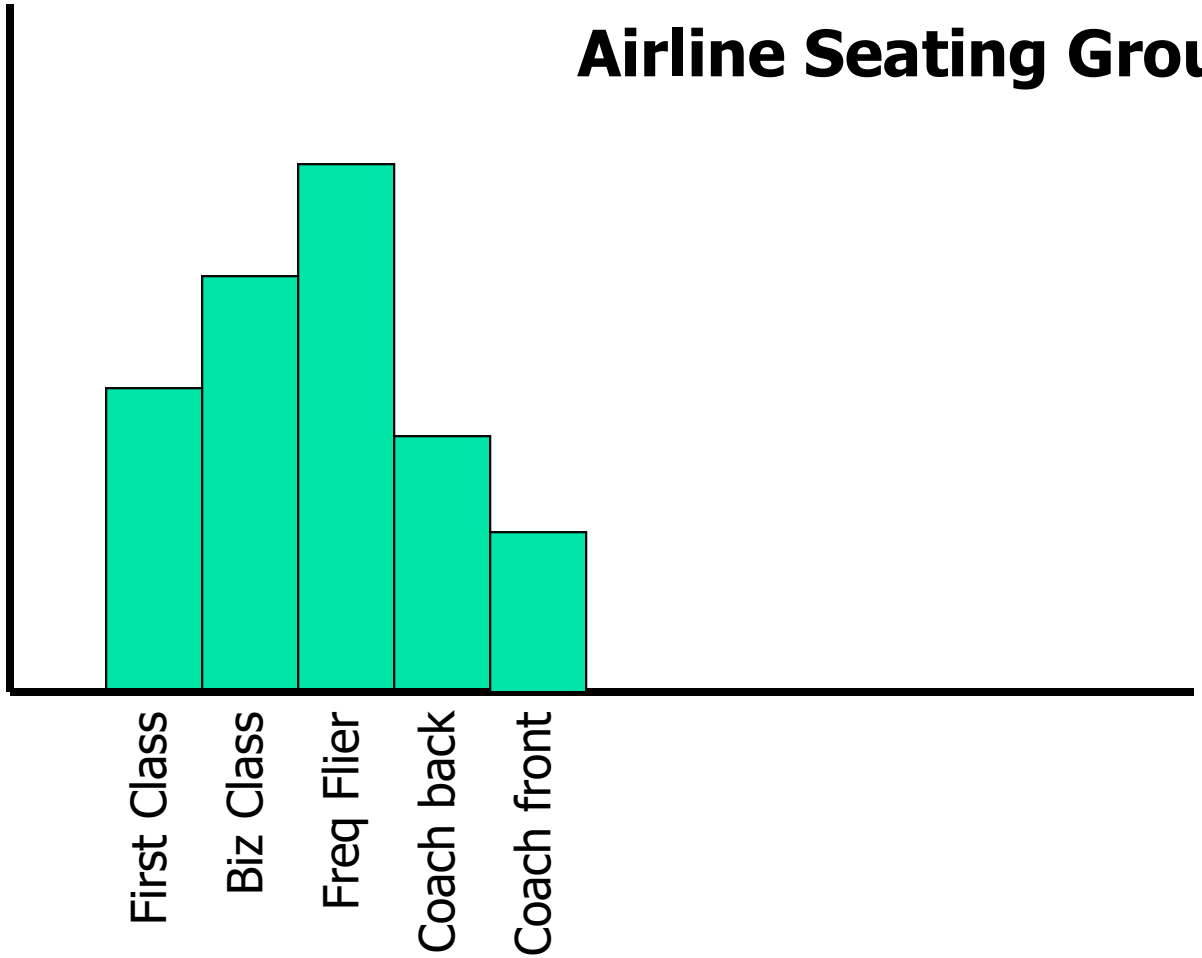


Which measures of center and spread?

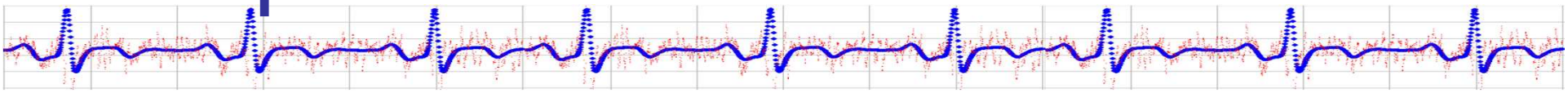


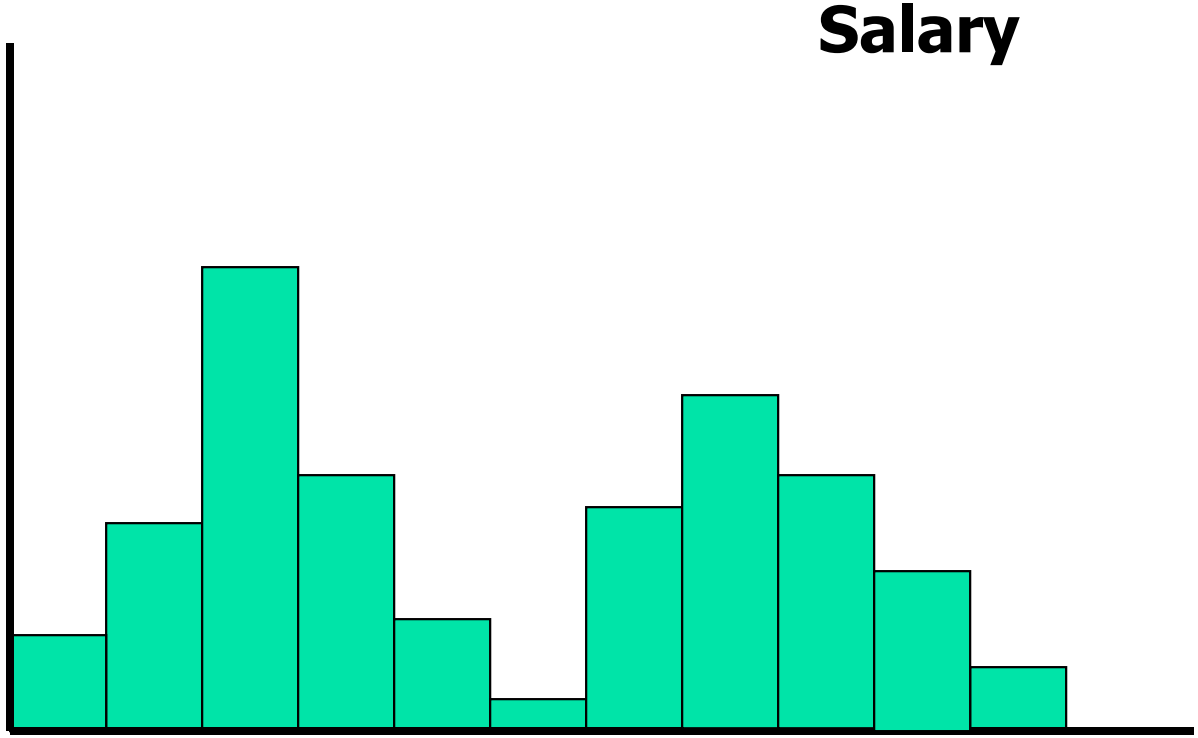
Which measures of center and spread?



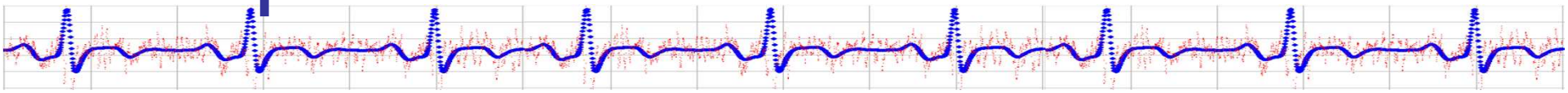


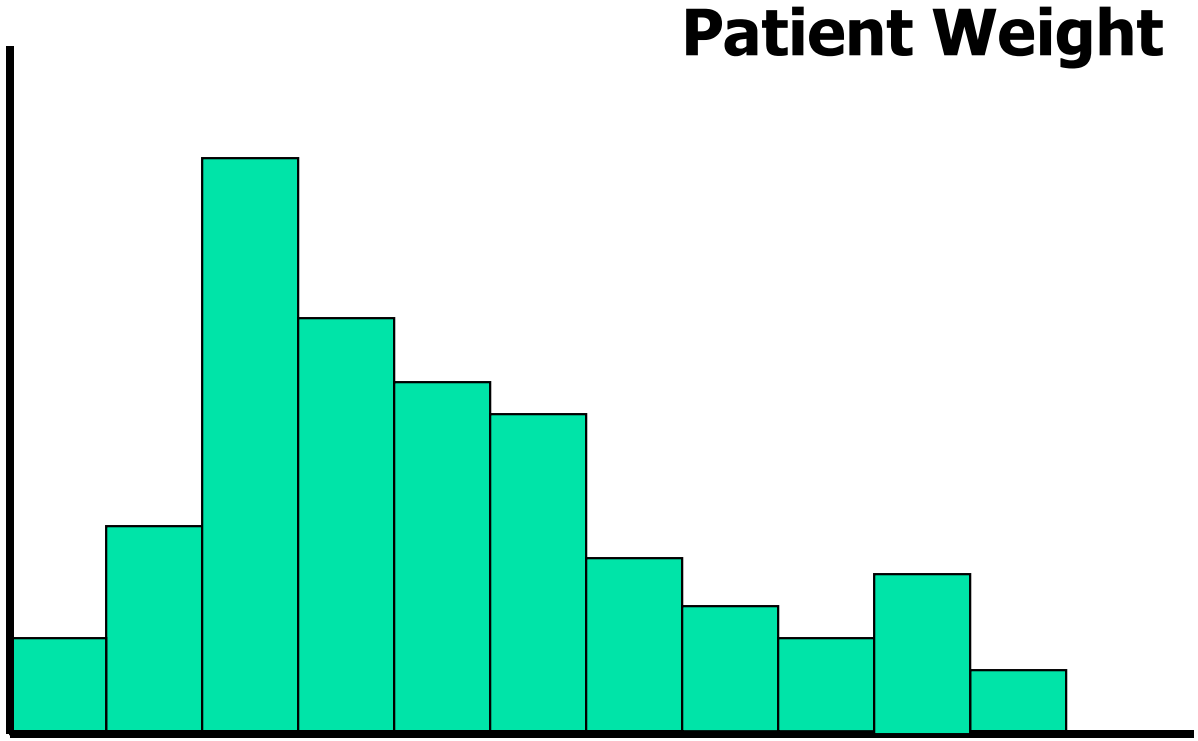
Which measures of center and spread?



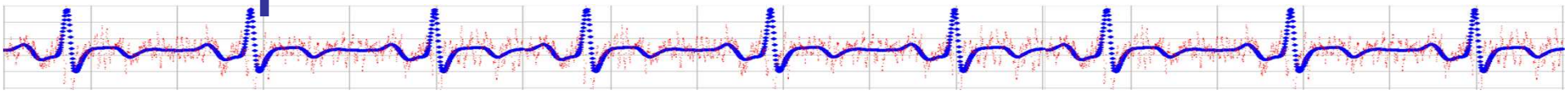


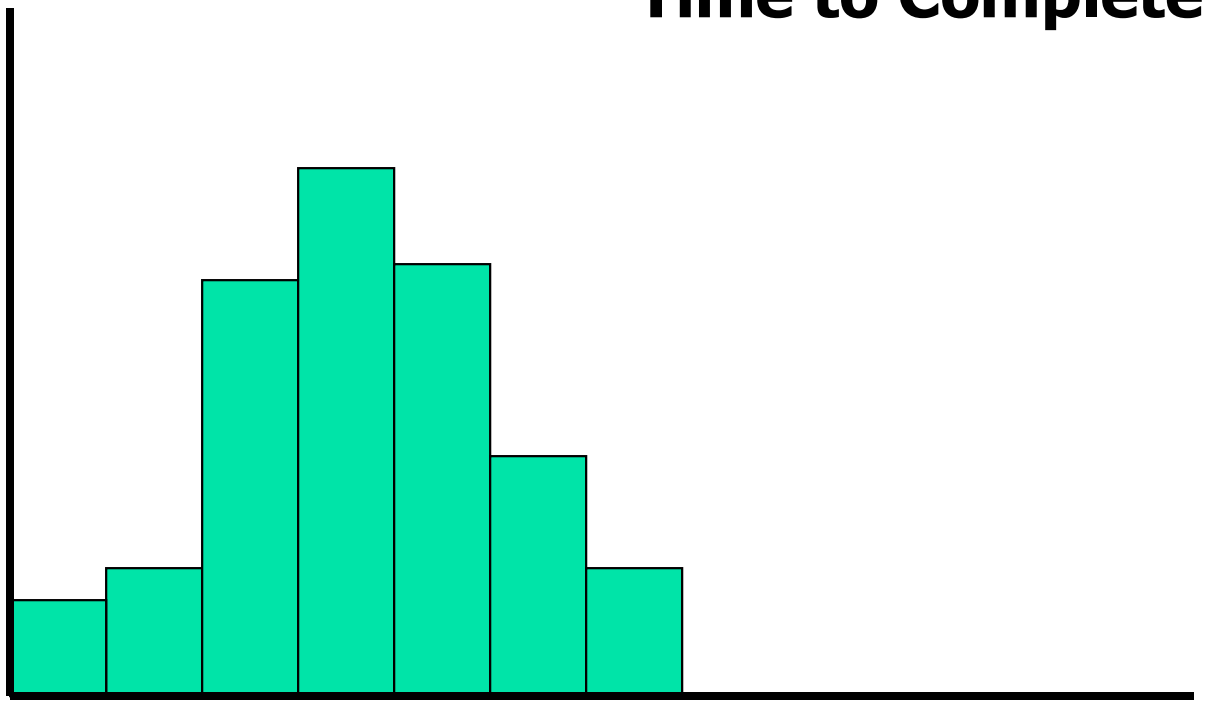
Which measures of center and spread?



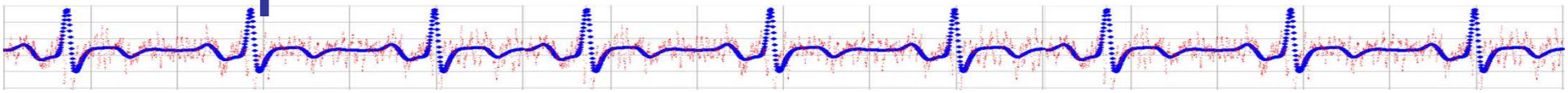


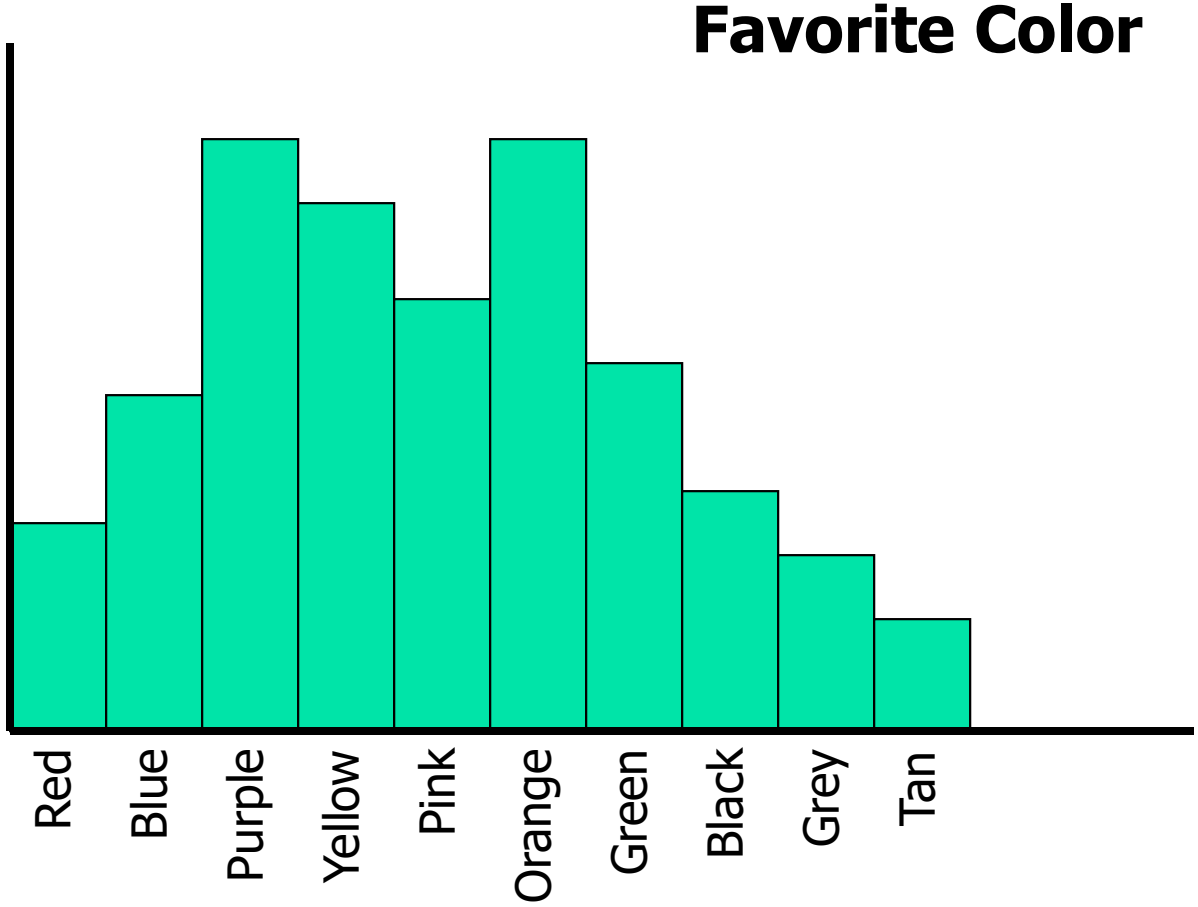
Which measures of center and spread?



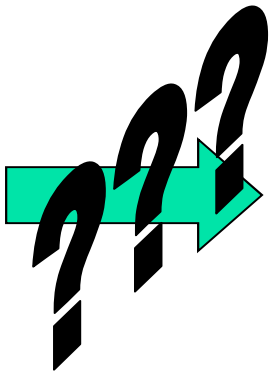


Which measures of center and spread?



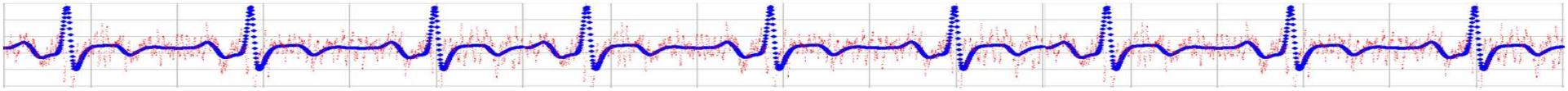


Observational Research



Nonexperimental Research

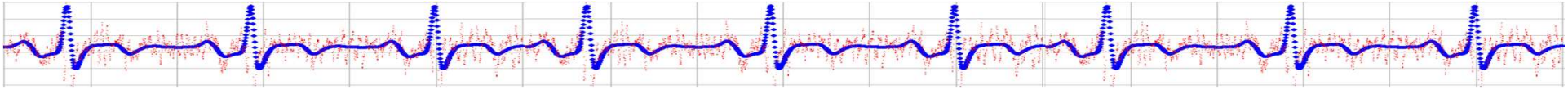
Example: Handheld ECAs

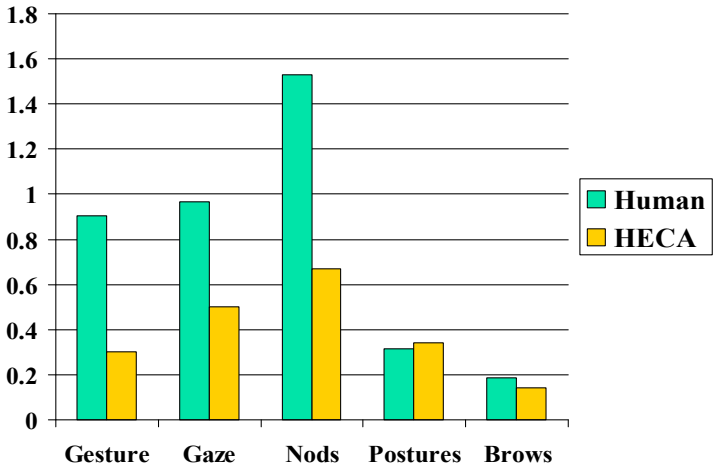


- Research Question:
  - Do people exhibit the same nonverbal conversational behavior when talking to a 2” tall character than when talking to another person face-to-face?
- Exercise:
  - Design the study

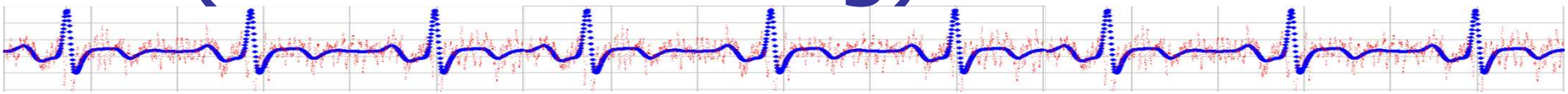
Example: Handheld ECAs

Results



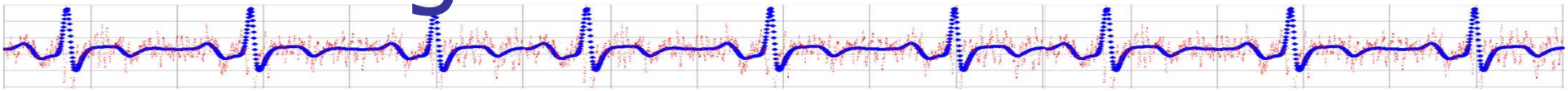

1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Gesture
Gaze
Nods
Postures
Brows
Human
HECA

Observational research (behavior coding)



Watching people and quantifying their behavior

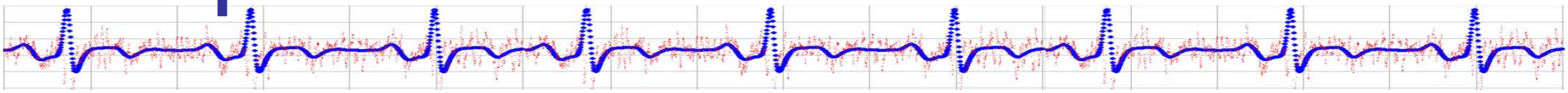
Defining behavioral categories



- Only need enough detail to provide a reliable measure.
  - What is reliability?
  - How to measure it?

- E.g. do people in the student center get more rude 10 minutes before class times?

Defining a behavioral protocol



Draft coding manual

Two or more observers code data for pilot subjects

Is Kappa acceptable?

no

Coders review differences & update coding manual

yes

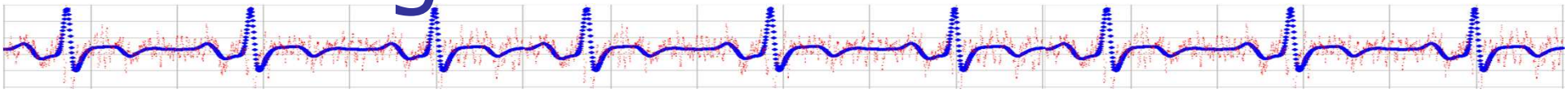
Coding manual provides reliable measure, proceed with study

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

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

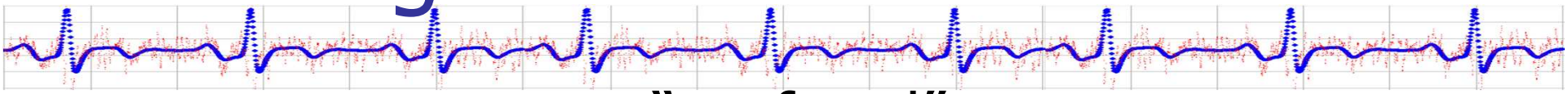
Where Pr(a) represents the actual observed agreement, and Pr(e) represents chance agreement.

Developing behavioral categories



- Categories must be operationally defined
- Behavioral categories must be clearly defined to avoid ambiguity
  - E.g., "flailing arms around" vs. "moved arms from below to above waist and back more than 3 times per minute"
- However, must be practical

Developing behavioral categories



"confused"

vs.

("clicked mouse at least 5 times on inappropriate menu"

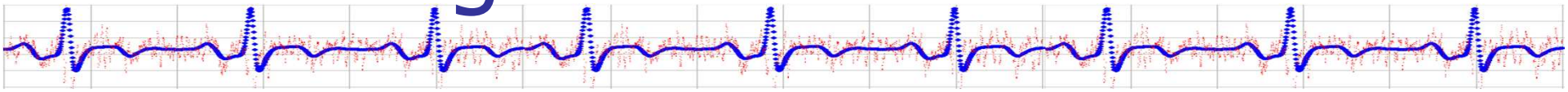
OR

"gazed at interface with mouth open AND no mouse clicks or keyboard presses for 5 minutes")

AND

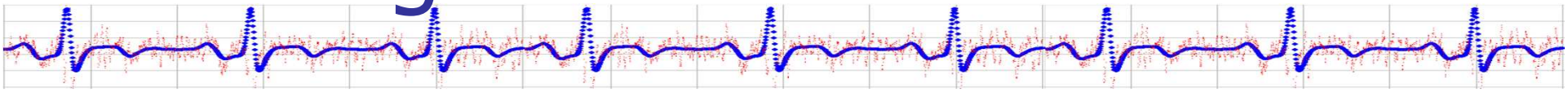
"furrowed brows"

Coding manual



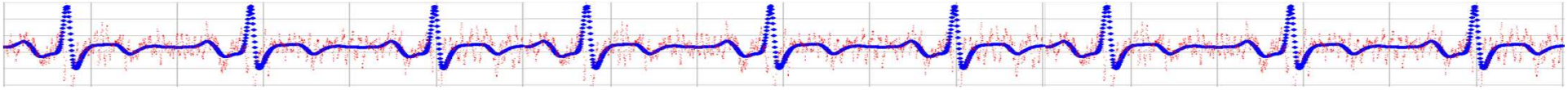
- You should write your behavior identification rules down so that you could give them to someone else to follow reliably
- You should also write down the sampling and coding methods you will use, as well as your recording instrument (e.g., paper form)

Developing behavioral categories



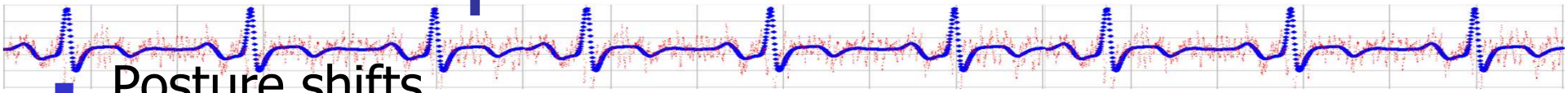
How do you know when your definitions are good enough?

Quantifying behavior: What is the metric?



- *Frequency Method*
  - Record the frequency with which a behavior occurs within a time period
- *Duration Method*
  - Record how long a behavior lasts
- *Intervals Method*
  - Divide the observation period into several discrete time intervals (e.g., ten two-minute intervals), and record whether a behavior occurs within each interval

Example



- Posture shifts
  - Body part
    - Upper body
    - Lower body
    - Both
  - Type
    - Shift
    - Return
  - Energy level
    - 0-100%

StartTime	EndTime	BodyPart	Type	Energy
00:00:03	00:00:04	Upper	Return	50%
...	...	...	...	

- Hand gestures and other communicative behavior does not count – nor their effects.
- Video reviewed and start/stop/type coded.
- From this, we can compute frequency, duration, or intervals

Example: Code Posture Shifts

Posture Shifts

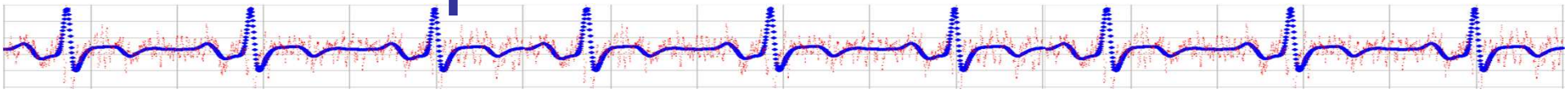
Duration, Frequency, or Interval Measures?

Posture shifts with respect to discourse segment

	Monologues (0.06/s)			Dialogues (0.07/s)		
	ps/s	ps/int	energy	ps/s	ps/int	energy
Inter-dseg	0.340	0.837	0.832	0.332	0.533	0.844
intra-dseg	0.039		0.701	0.053		0.723

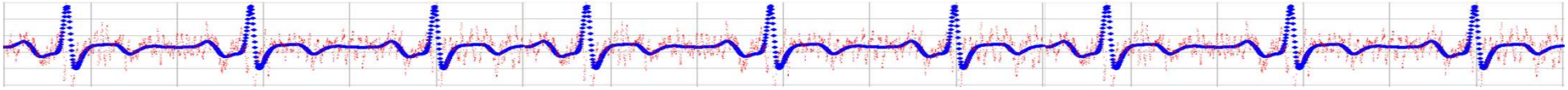
Tools for Coding: ANVIL

Example



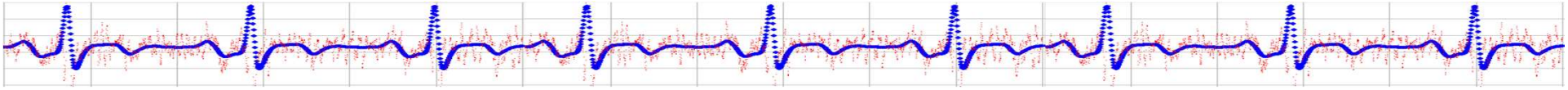
- Is this an example of
  - Frequency method?
  - Duration method?
  - Intervals method?

Coping with complexity in observational research



- *Recording*
  - Use a recording device to make a record of behavior for later review
- *Time Sampling*
  - Scan subjects for a specific period (e.g., 30 seconds), and then record your observations during the next period
- *Individual Sampling*
  - Select a subject and observe behavior for a given period (e.g., 30 seconds), and then shift to another subject and repeat observations

Coping with complexity in observational research

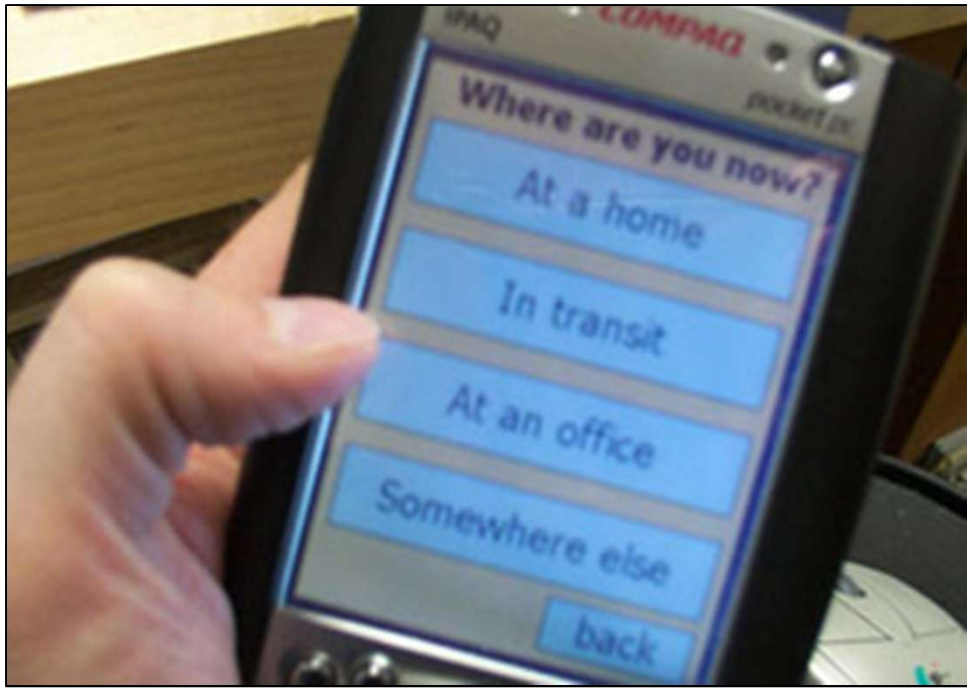


- *Event Sampling*
  - Select one behavior for observation and record all instances of that behavior
  - It is best if one behavior can be specified as more important than others

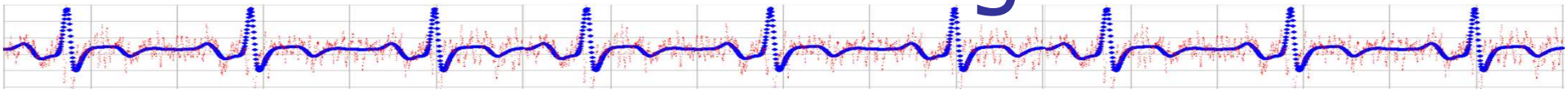
Coping with complexity in observational research

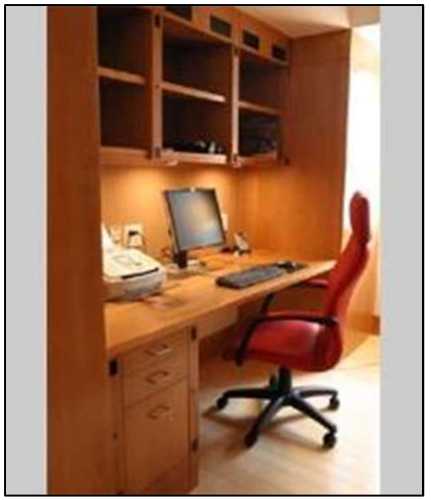
- Ecological momentary assessment
- Intelligent/Context Aware EMA

- What kind of sampling is this?

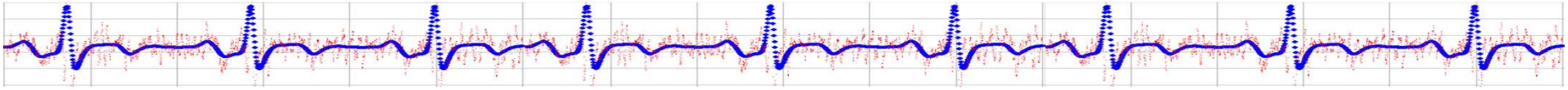


Smart Rooms – e.g. PlaceLab





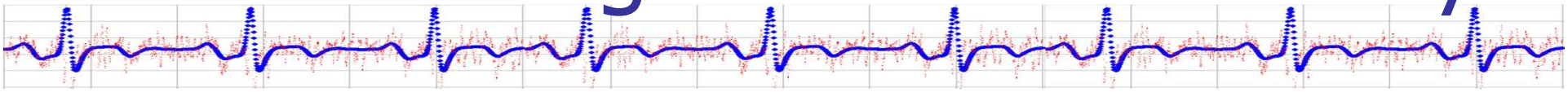
Issues in observational research



- IRB issues with video/audio recording?

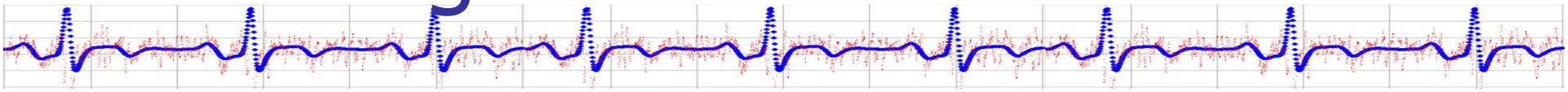
- Behavior vs. function/intent

Evaluating interrater reliability



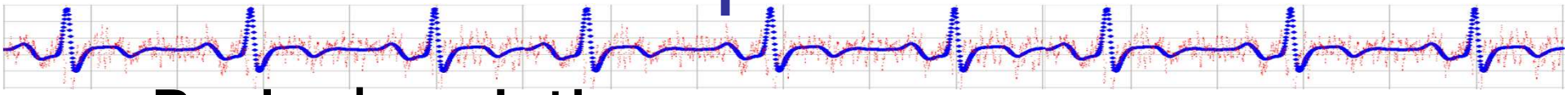
- You must establish reliability of observations from multiple observers (*interrater reliability*)
- Most common/acceptable method for evaluating interrater reliability for a nominal measure, 2 raters
  - *Cohen's Kappa*
    - Allows you to determine if agreement observed is due to chance
    - Kappa of 0.70 or more indicates acceptable interrater reliability

Histogram



- hist(data)
- hist(wage, col = "yellow")

Basic Descriptive Stats



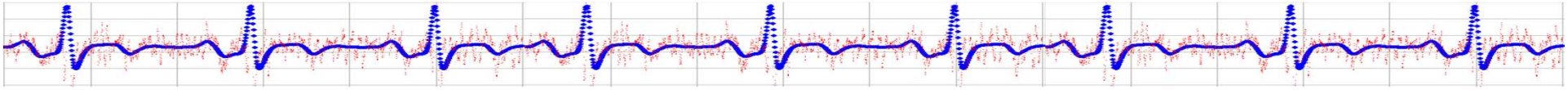
- Basic descriptives

```
summary(eruptions)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.600   2.163   4.000   3.488   4.454   5.100
```

- Histogram

```
hist(eruptions)
```

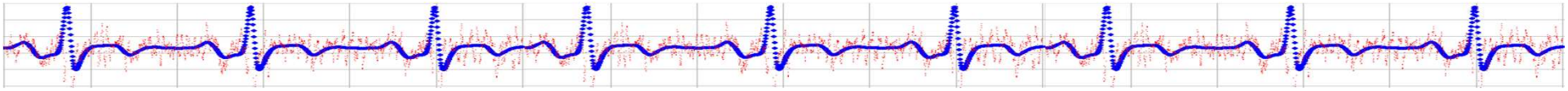
More Basics



- mean(data)
- median(data)

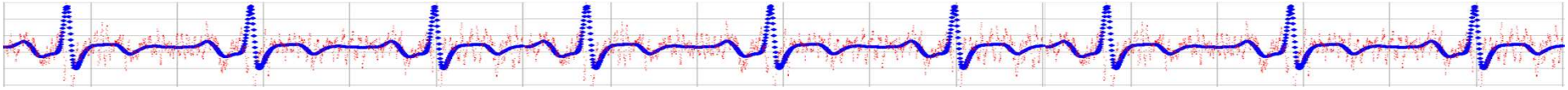
- var(data)
- sd(data)
- IQR(data) #inter-quartile range

Frequency Tables



- table(data) #frequency counts
  - Returns table of frequency counts for each unique value in argument
  - Positions in table are "named" with the value counted
- E.g. table(c(2,2,2,3,3,4))
  - returns table 3,2,1
  - with names 2,3,4
- names(table(data))
  - Returns vector of unique names, sorted

Mode

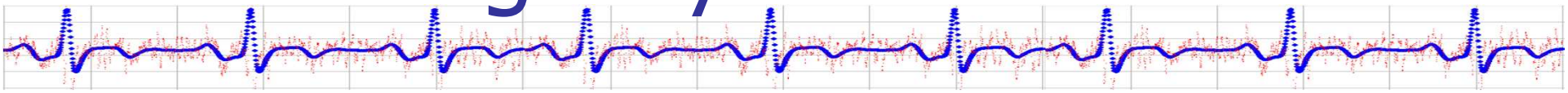


- Given values in vector ‘x’

```
temp <- table(x)
names(temp)[temp == max(temp)]
```

#this returns the names of the values that have the highest frequency in x.

Contingency Tables

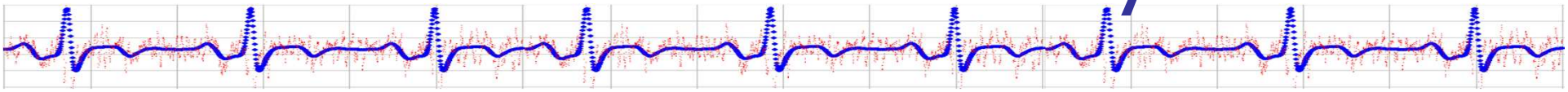


- Co-occurrence of values for 2 variables.
- table(val1,val2) #freq counts

e.g.

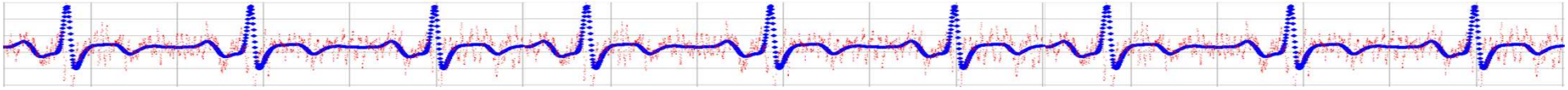
	2	3	4	5	6	7	8	9	10	12
F	0	1	0	1	1	1	0	1	0	1
M	1	1	1	0	0	0	1	0	1	0

Example R data setup for interrater reliability



Time	Judge1	Judge2
1	together	together
2	apart	apart
3	together	together
4	apart	together
5	apart	apart
6	together	together
7	together	together

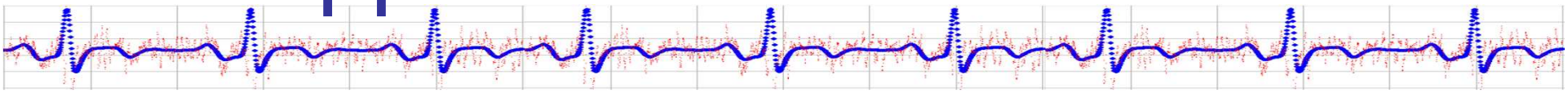
Contingency Table
aka Cross-Tabulation



- table(vector)
  - Matrix:
    - first row = values
    - second row = freq counts

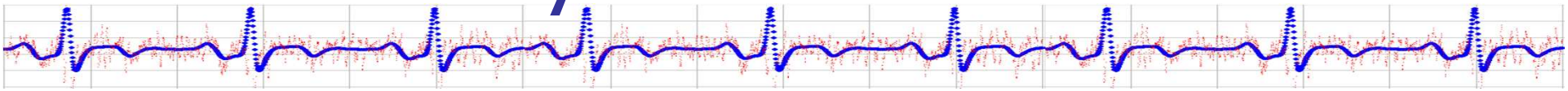
- Co-occurrence of values for 2 variables.
  - table(val1,val2)

Kappa



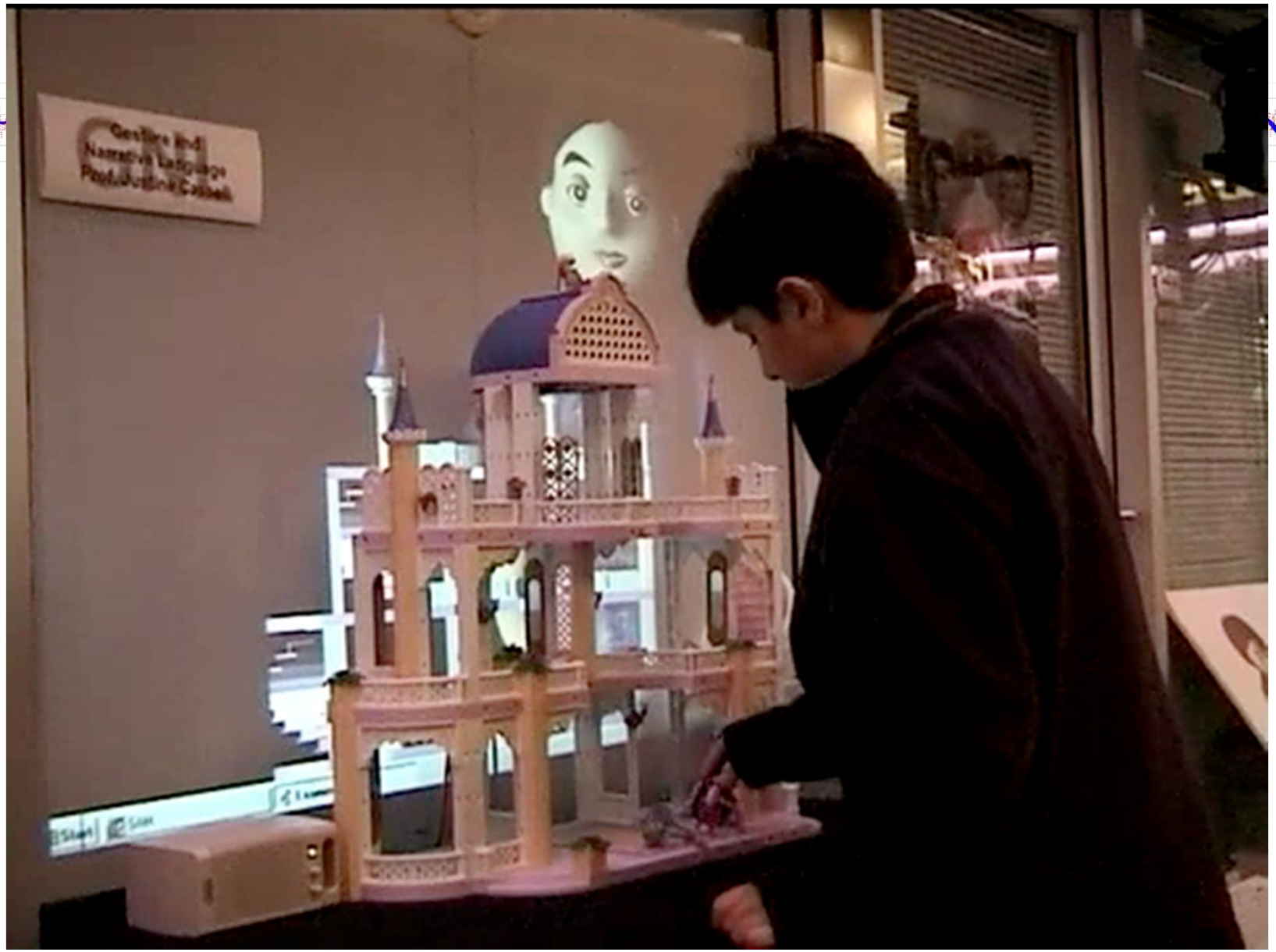
```
> install.packages("psych") #one time

> require(psych)  #every session

> wkappa(table(data$Judge1,data$Judge2))

$kappa [1] 0.6

$weighted.kappa [1] 0.2
#accounts for distance of each discrepancy
#= how bad different disagreements are
#ignore for now
```

Other statistics for inter-rater reliability



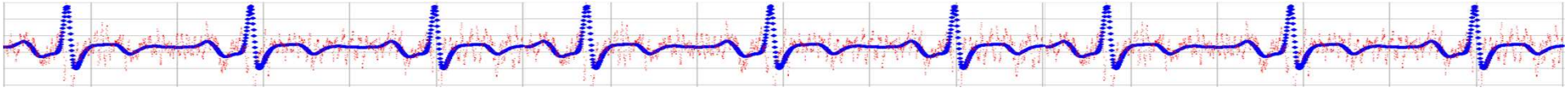
- Fleiss' kappa
  - Nominal, >2 raters
- Kendall's τ, or Spearman's rho
  - Ordinal, 2 raters (not testing absolute match)
- Pearson correlation coefficient
  - Interval or ratio, 2 raters (not testing absolute match – only whether linearly related)
- Intraclass correlation coefficient
  - Interval or ratio, 2+ raters
- Hallgren, KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol. 2012;8(1):23-34.

Behavior Coding Exercise



Behavior Coding Exercise

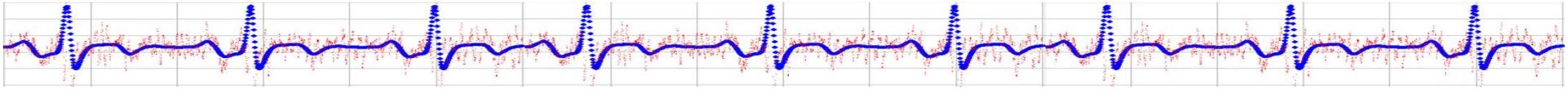
Groups of 2-4, one should have a laptop with R



- You are developing a robotic couples counselor.
- You want to determine how couples react to it.
- Given nominal variable to code
- Discuss
  - Meaning
  - Refine values
  - Behavioral correlates
  - Draft coding manual
  - Focus on nonverbal behavior (poor audio)

Behavior Coding Exercise

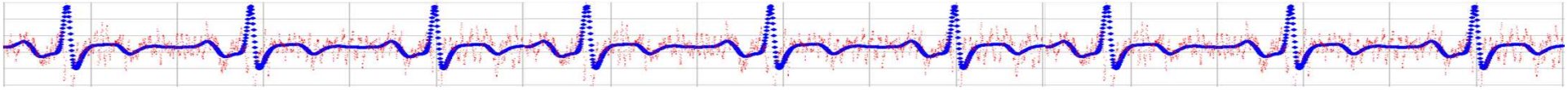
Groups of 2-4, one should have a laptop with R



- Shown 2-3 minute samples from 3 couples
- Interval sampling, 10s intervals
- Each judge codes behavior
  - Suggest shorthand, eg “E” for “Engaged”
  - Annotate Couple ID, landmarks (e.g. start of speaking turn), sample ID

Behavior coding exercise

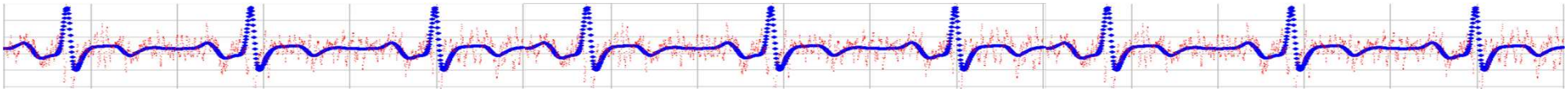
Groups of 2



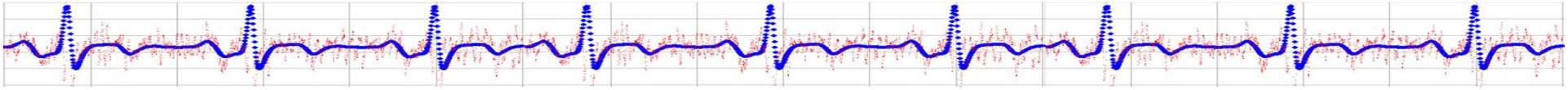
- Put into one spreadsheet with one row per observation, one column per judge
- Compute interrater reliability
  - If <0.7 discuss discrepancies and improvements
  - Update coding manual

- Videos shown 2nd time for discussion
- Videos shown 3rd time for 2nd pass coding
- Repeat kappa calcs

Code!

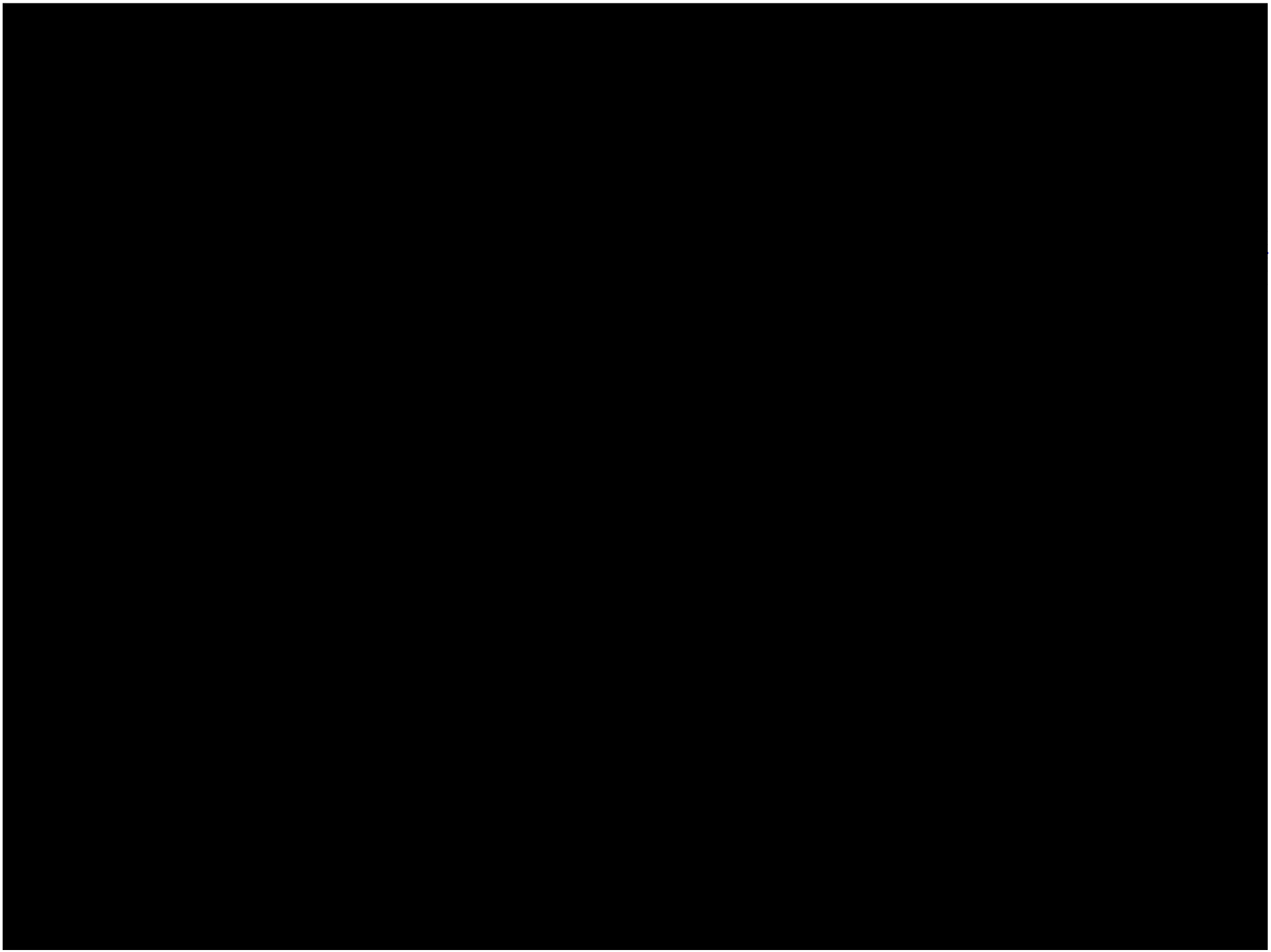


Behavior Coding Exercise: Form Pairs!

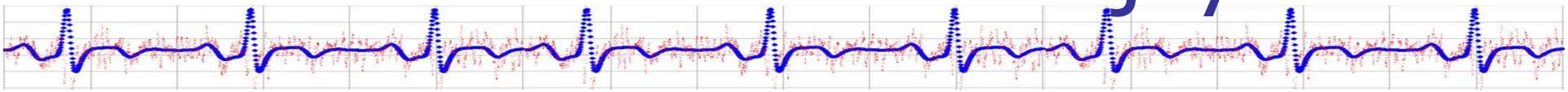


- Does playing with Sam cause children to play together more?
- Draft code manual for:
  - Playing together
  - Playing alone
  - Not playing
- Time sampling
  - 5 second intervals
- Compute interrater reliability

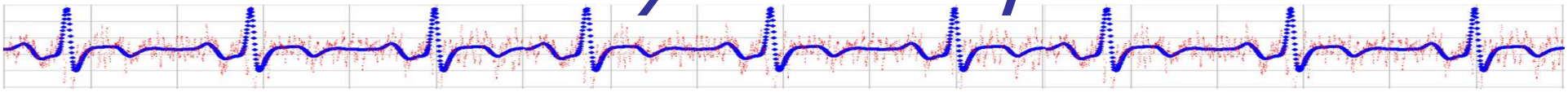
Desktop: dyad w/o sam



Exercise: Evaluate Enjoyment

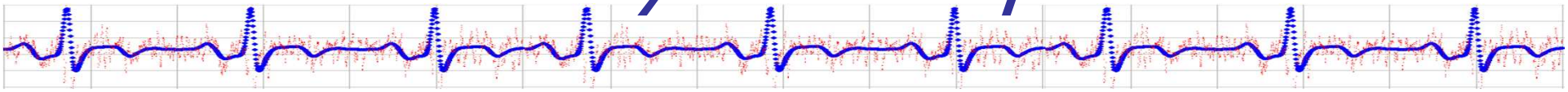


Approaches to data collection: *Necessarily non-experimental?*



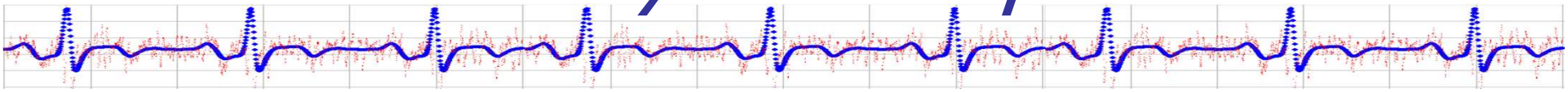
- *Naturalistic Observation*
  - Unobtrusive observations of subjects' naturally occurring behavior are made
- *Ethnography*
  - The researcher becomes immersed in the behavioral or social system being studied. May be conducted as a participant or non-participant observation study
- *Sociometry*
  - You identify and measure interpersonal relationships within a group

Approaches to data collection *Necessarily non-experimental?*



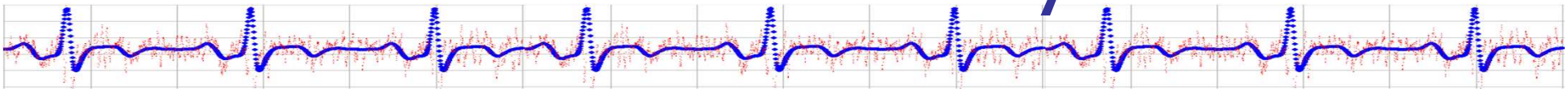
- *Case History*
  - You observe and report on a single case
- *Content Analysis*
  - You analyze spoken or written records for the occurrence of specific categories of events (e.g., a word or phrase)

Approaches to Data Collection *Necessarily non-experimental?*

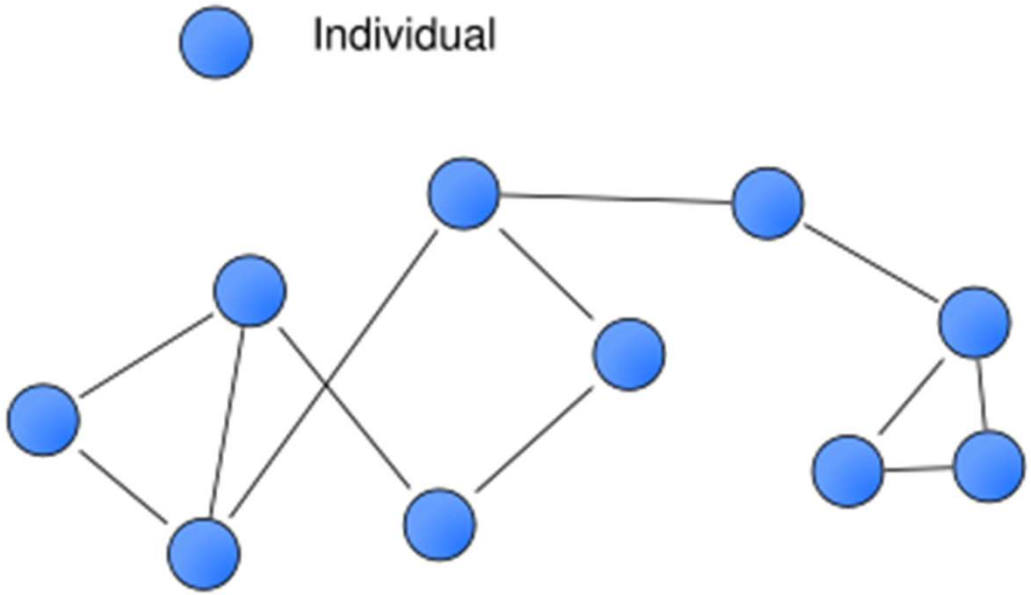


- *Archival Research*
  - You use existing records (e.g., police records) as your source of data
- *Meta-Analysis*
  - Compute overall statistics based on a number of previously-published studies

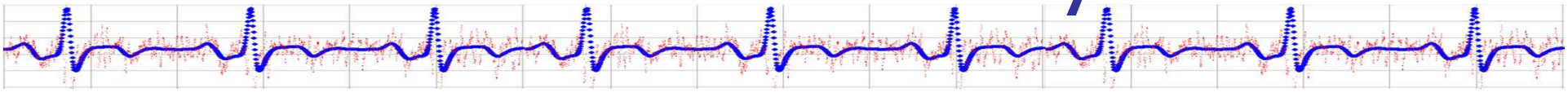
Sociometry on steroids: Social network analysis



- Analyze structures
- Measures
  - Betweeness – how often a node is an intermediary
  - Centrality – number of ties to others
  - Many more…

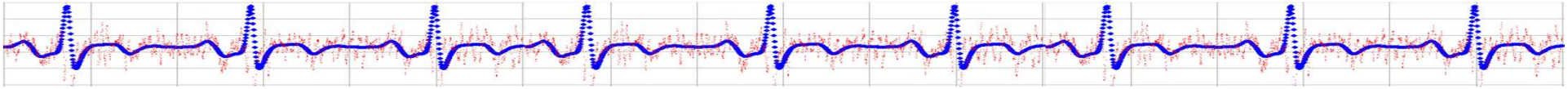


Sequential Analysis
aka time-series analysis



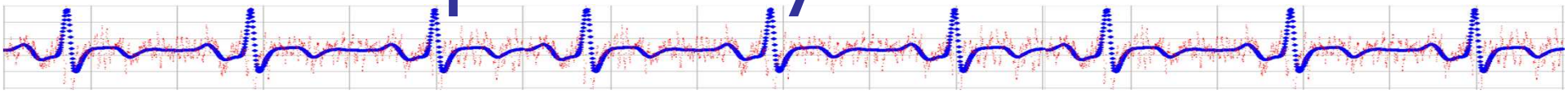
- B&A say recording sequences of behavior may yield more information than individual events
  - E.g., interruption followed by grimace followed by rolling eyes

Content analysis: Defining characteristics



- Used to analyze a written or spoken record for occurrence of specific behaviors or events
- Archival sources often used as sources for data
- Response categories must be clearly defined
- A method for quantifying behavior must be defined
- Tools exist:
  - http://www.lexicoder.com
  - Topic modeling: https://radimrehurek.com/gensim/

Example study



The CEO of Global Enterprises, Inc. is worried about the low morale in the company, as evidenced by the amount of flame email she receives. She considers sending every office on a "ropes" course, but to do this would cost the company $10M. She asks you to do a study to tell how well her scheme might actually work in reducing her flame mail.