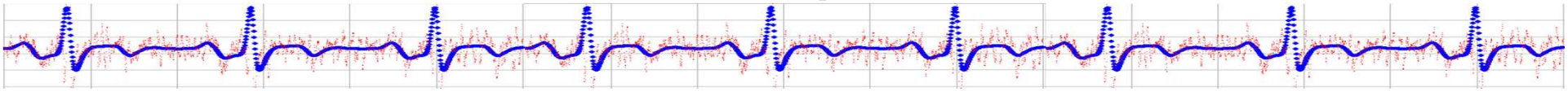
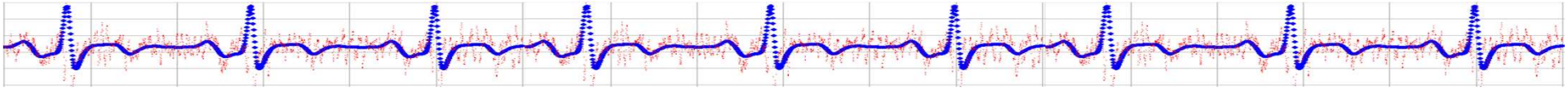
Empirical Research Methods in Information Science

IS 4800 / CS6350



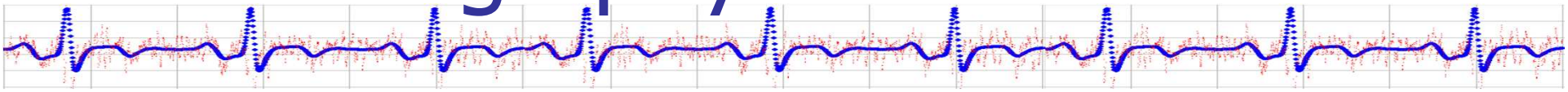
Lecture 13

Outline



- Reading assessment
- Homework I5 – Due date EOD Friday
- Ethnography observations
- Thinking about variables
- Correlation
- Start inferential statistics

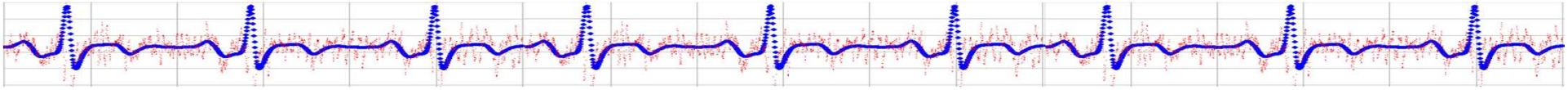
Ethnography



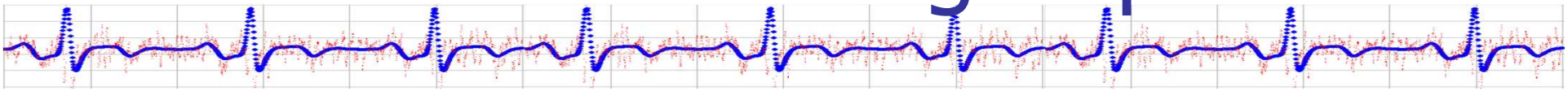
- Richly document experience
- Interviews
  - Consistency
  - Detail (quotes, tone matter)
- Count things
- Measure things
- See repetition … it's your friend

- Variables: measurable
- Variables: clear description
- Boxes: think about the lines!

Student center variables



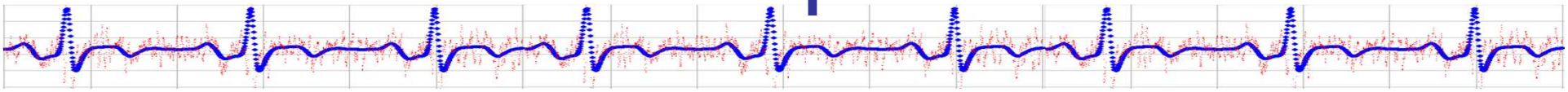
Sidebar: Control groups



- Standard-of-care control (new vs. old)
- Non-intervention control
- "A vs. B" design (shootout)
- "A vs. A+B" design

- Problem: the "intervention" may cause more than just the desired effect
  - Example: giving more attention to intervention Ss in educational intervention
- Some solutions:
  - Attention control
  - Placebo control
  - Wait list control (also addresses ethics)

Sidebar: Control groups Related concepts

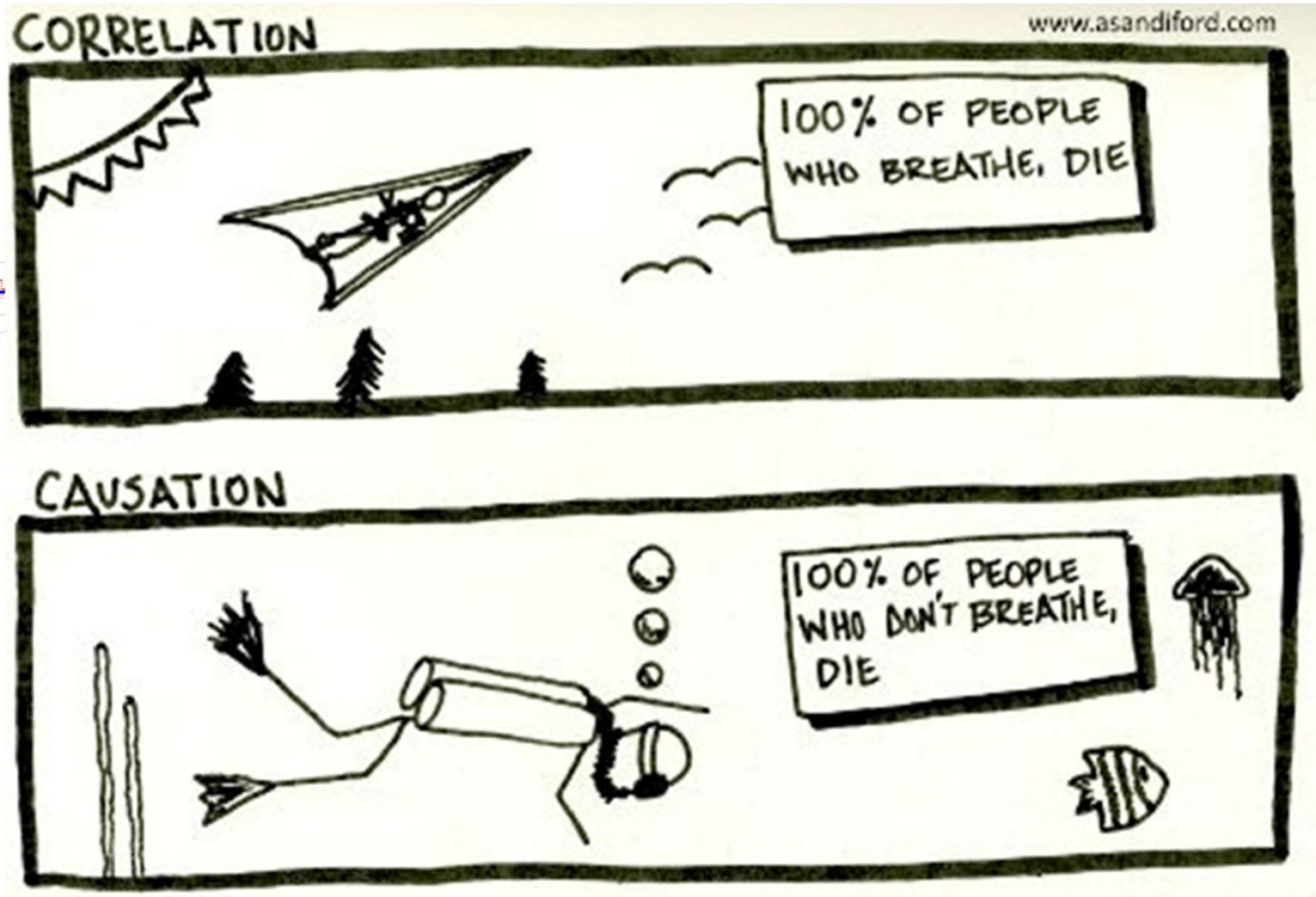


- Blind test – experimenter does not know group
- Double blind test – neither S nor experimenter know

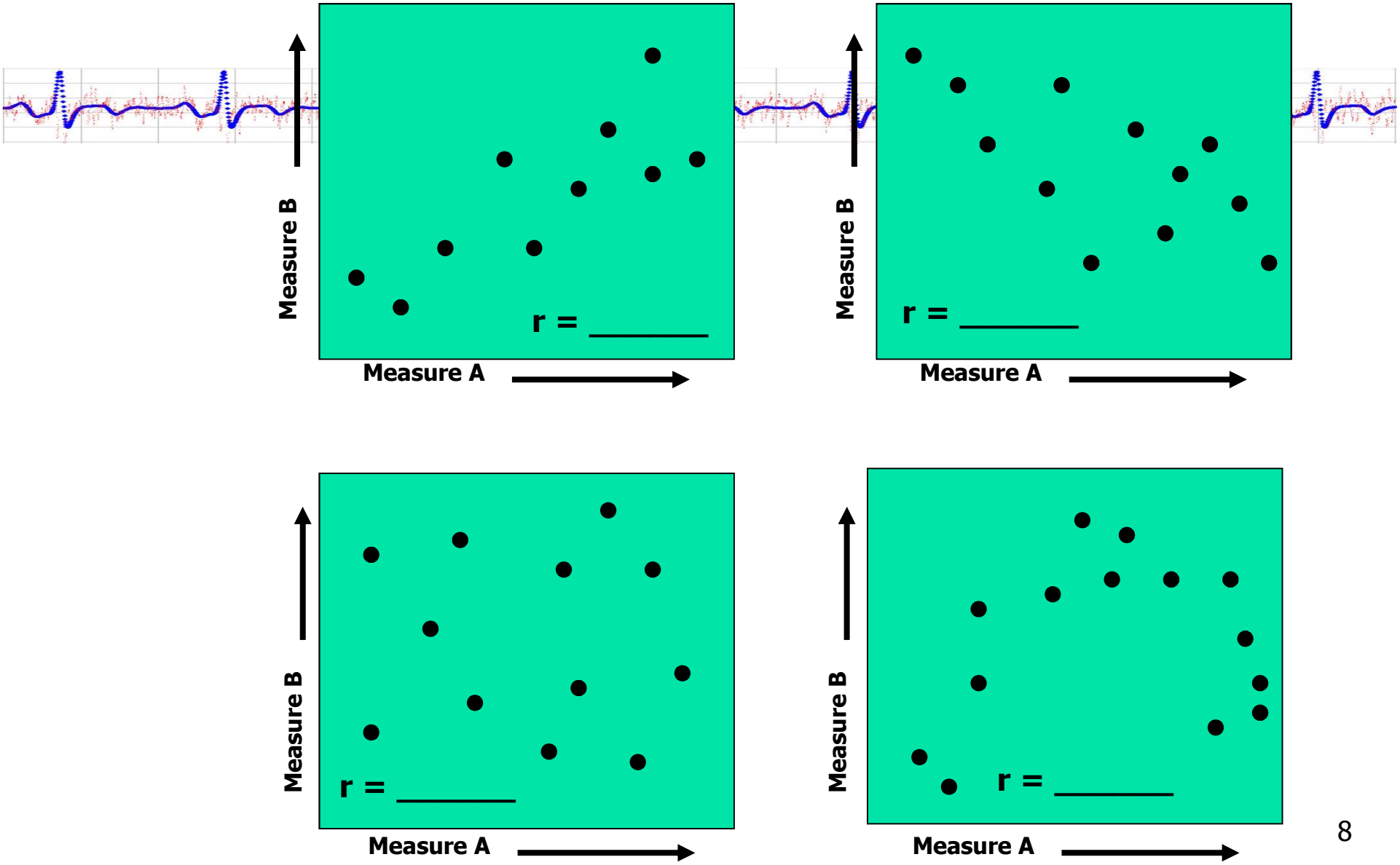
- Manipulation check
  - Test performed just to see if your manipulation is working. Necessary if immediate effect of manipulation is not obvious.
  - "Positive control" test for intervention effect
  - "Negative control" test for lack of intervention effect
  - Example:
    - Student Center Sign: ask students if they saw & read the sign

- Contamination
  - Some control subjects get some of the intervention

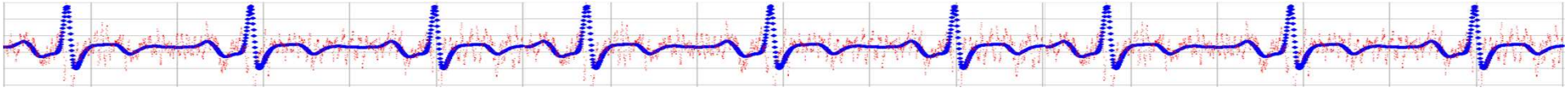
Remember, correlation!



Reminder


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

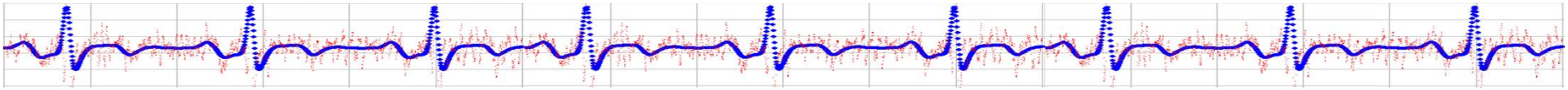
Pearson Correlation Coefficient

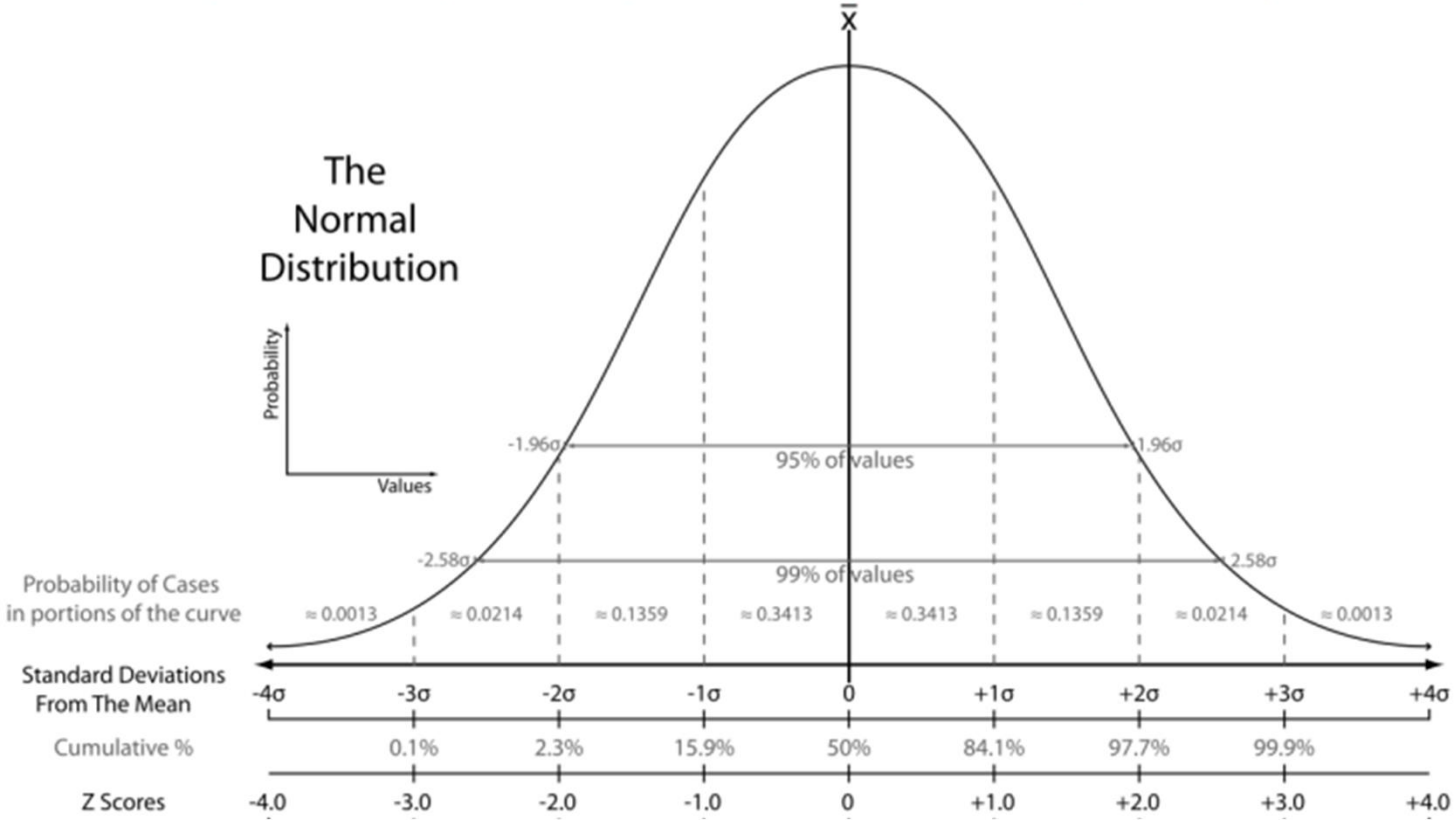


- Assumptions
  1. Two interval (or ratio) measures.
  2. Not an obviously curvilinear relationship.
  3. Both populations normally distributed*.

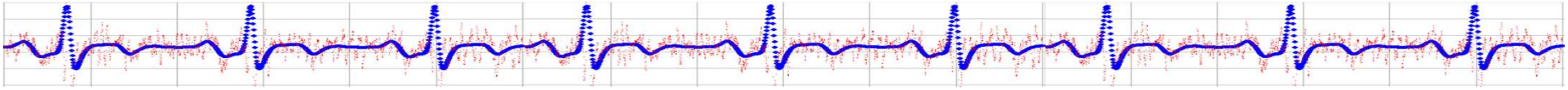
*Unimodal and symmetric frequency distributions. Most important if doing a significance test.

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-

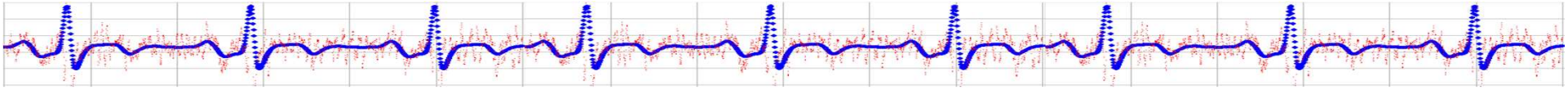
$$r = \frac{\sum [(X - M_X)(Y - M_Y)}{\sqrt{(SS_X)(SS_Y)}}$$

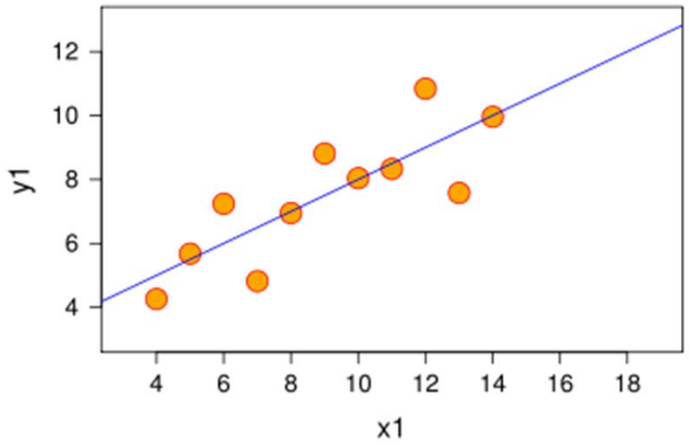
Where

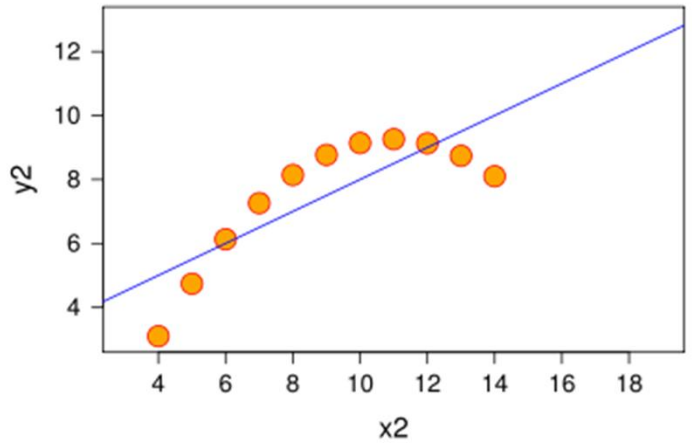
$$SS_X = \sum (X - M_X)^2$$

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

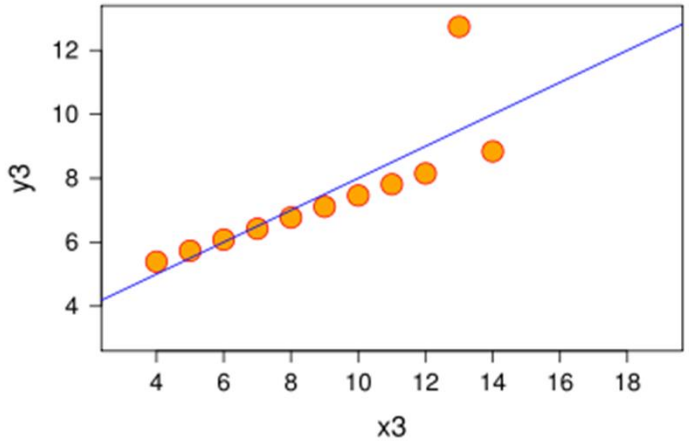
Correlation?

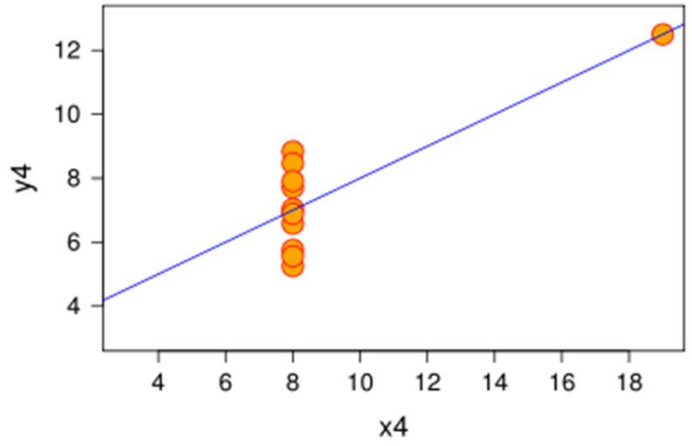




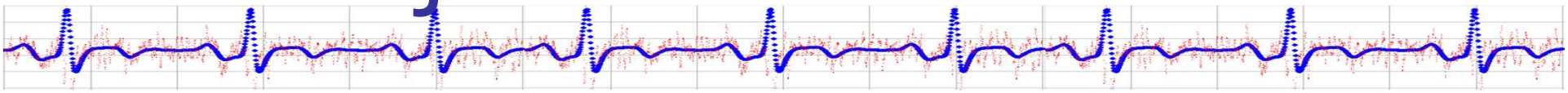


All
r=0.81





Matejka et al. 2017



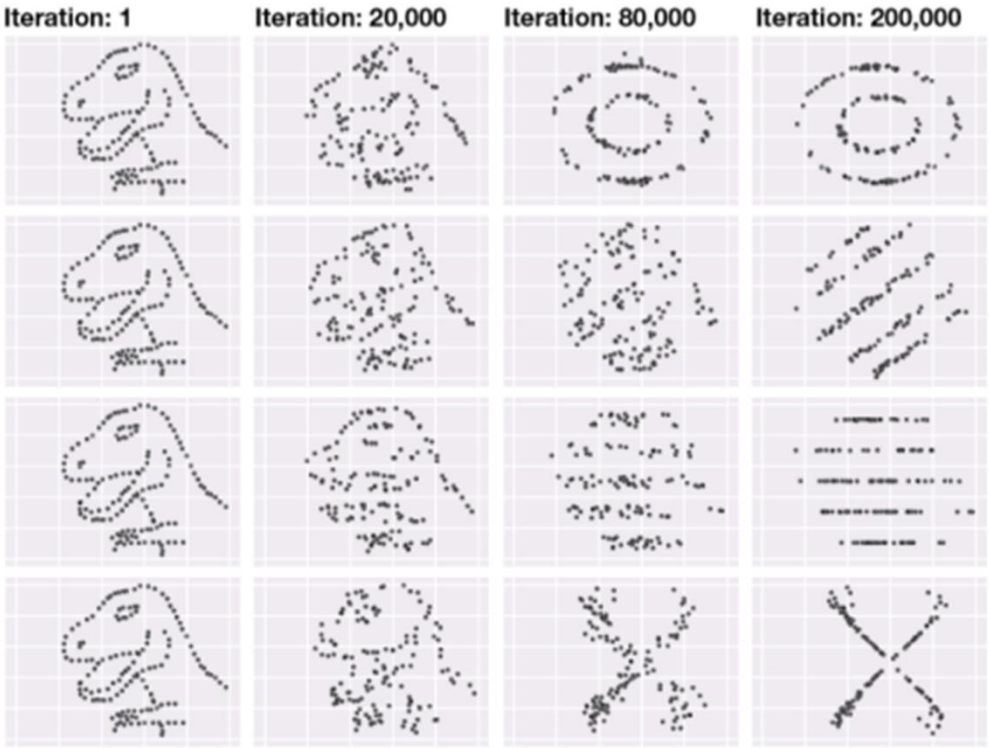
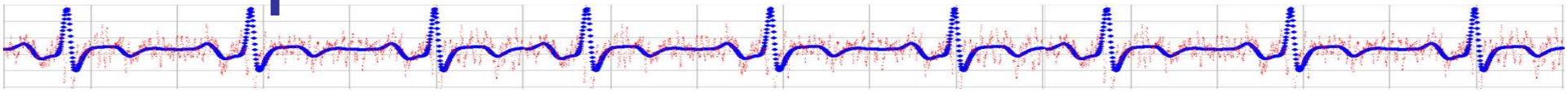


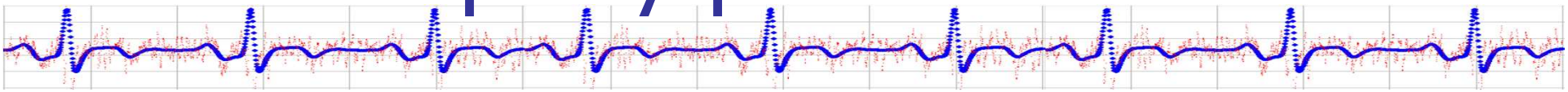
Figure 6. Creating a collection of datasets based on the "dinosaurus" dataset. Each dataset has the same summary statistics to two decimal places: ($\bar{x}$ =54.26, $\bar{y}$ = 47.83, sd_x = 16.76, sd_y = 26.93, Pearson's r = -0.06).

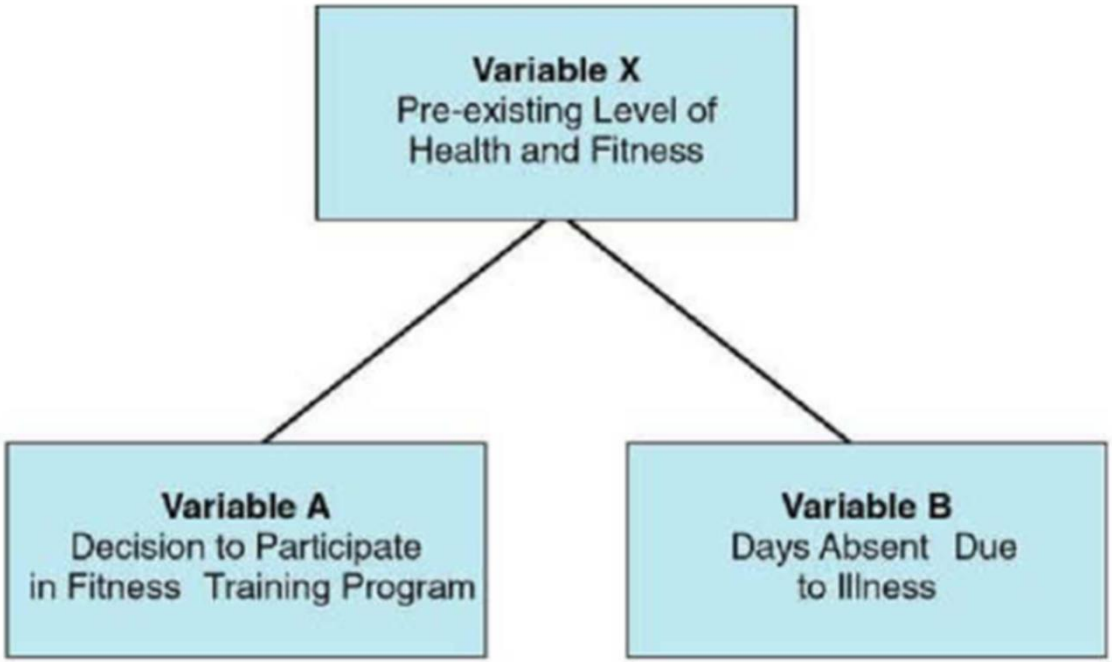
Spurious correlations



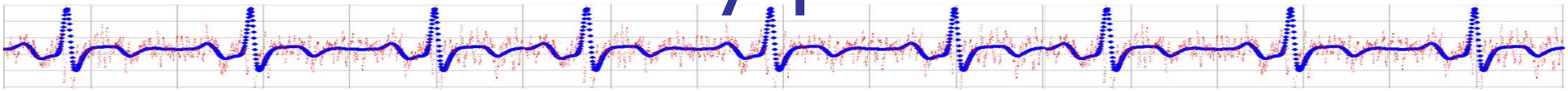
http://www.tylervigen.com/spurious-correlations

Third party problem





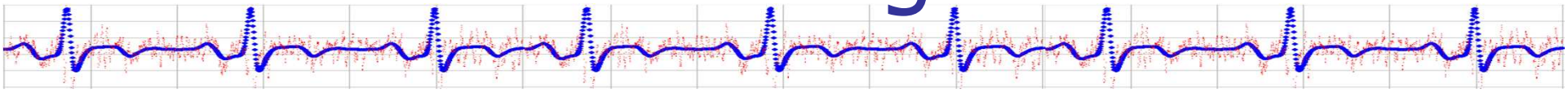
Directionality problem



Seeing violent acts on television causes children to behave aggressively.

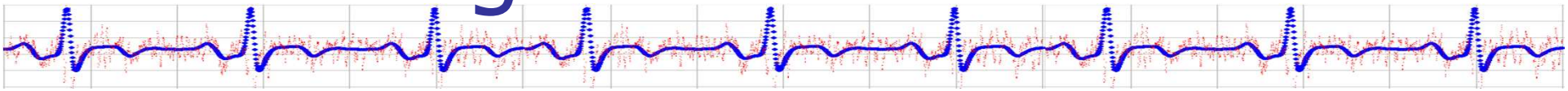
Children choose to watch TV shows with a level of violence that matches their own level of aggression.

Predictions Using Correlations



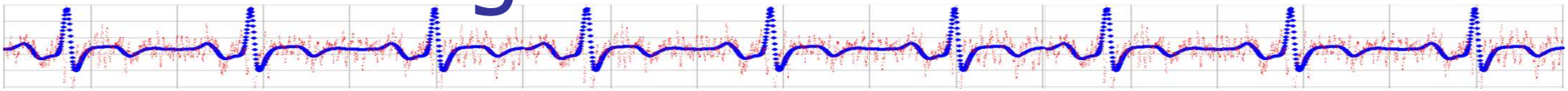
- Predictor vs. Criterion (Dependent) Variable
- Can you assume directionality?
- Depends entirely on your study design.

Procedure for Hypothesis Testing with Correlations



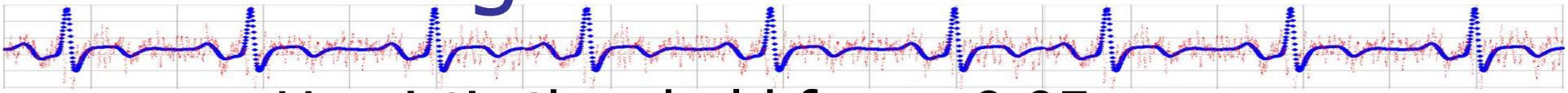
- Populations being compared:
  - ***Test:*** The population from which the observed sample was drawn.
  - ***Comparison:*** A hypothetical population in which the variables are unrelated, i.e., have a correlation of zero.

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

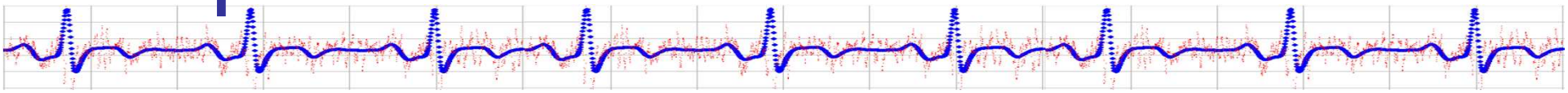


- Heuristic threshold for α=0.05:

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

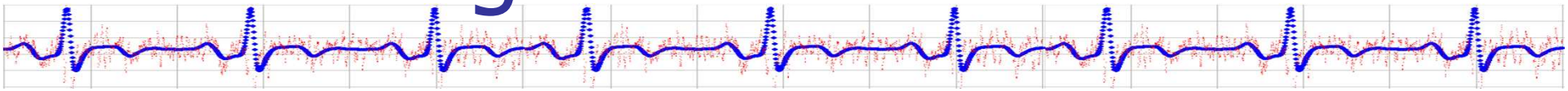
- *Exact form given in Aron, or in R.*

p-value and correlation



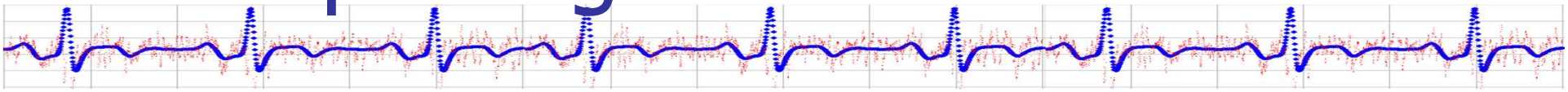
If a correlation coefficient has been determined to be statistically significant this does *not* mean there is a strong association. It simply tests the null hypothesis that there is no relationship. By rejecting the null hypothesis, you accept the alternative hypothesis that states that there is a relationship, 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*

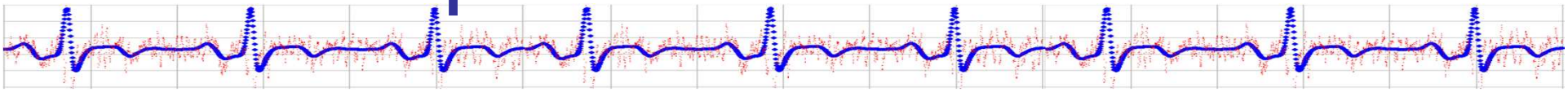
Where,

- 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**

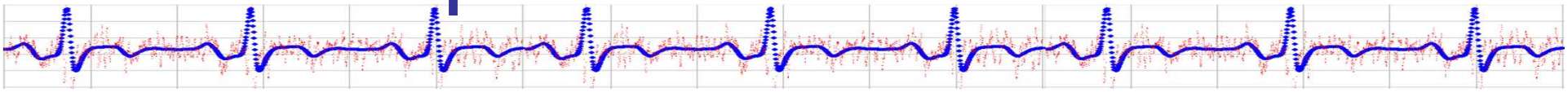
If not significant, than use "n.s." instead of "p<…".

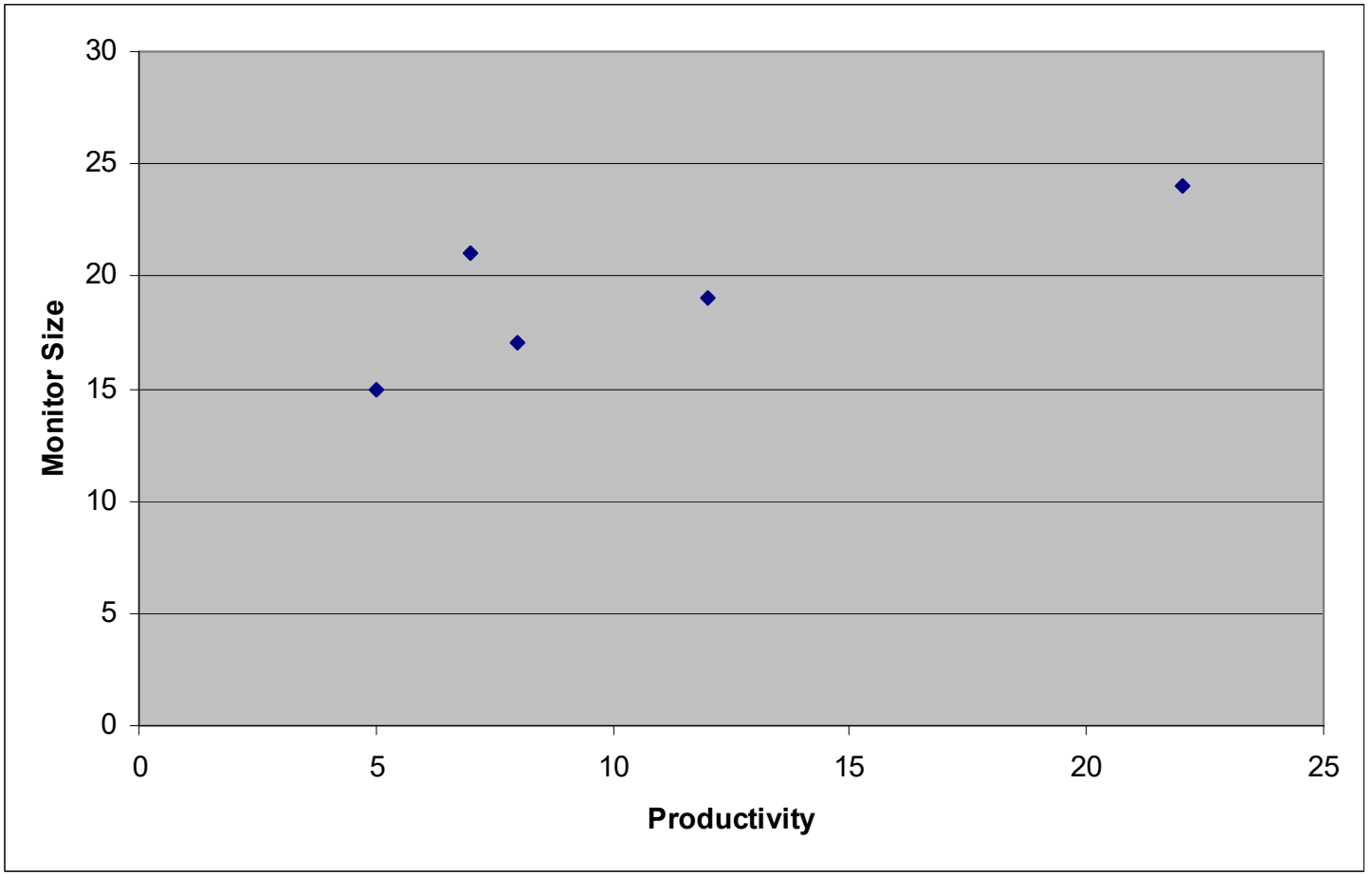
Example



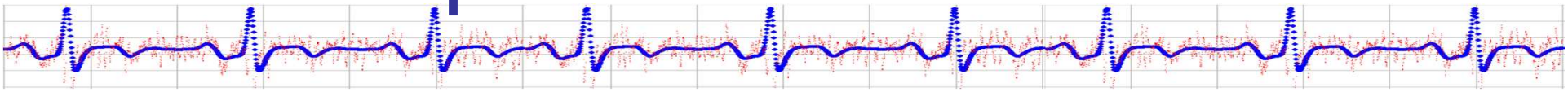
Employee	Sue	Sam	Sid	Sal	Sierra
Productivity	8	5	7	12	22
Monitor Size	17	15	21	19	24

Example Continued





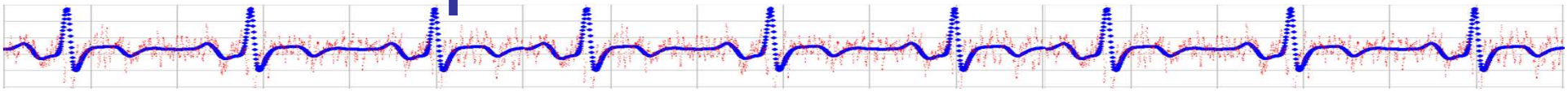
Example



Employee	Sue	Sam	Sid	Sal	Sierra
Productivity	8	5	7	12	22
Monitor Size	17	15	21	19	24

P = 10.8 (6.8), M = 19.2 (3.5)

Example



Employee	Sue	Sam	Sid	Sal	Sierra
Productivity	8	5	7	12	22
Monitor Size	17	15	21	19	24

P = 10.8 (6.8), M = 19.2 (3.5)

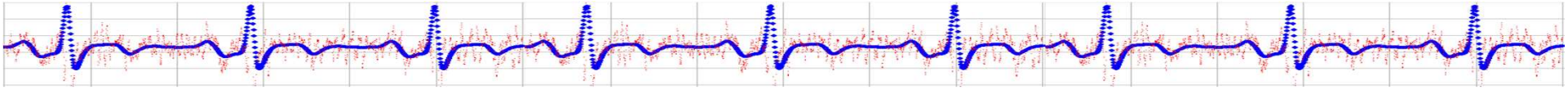
Z scores					
	-0.414	-0.858	-0.562	0.178	1.657
	-0.63	-1.202	0.515	-0.057	1.374

Z score products					
	0.260881	1.031666	-0.28968	-0.01016	2.276781

r = sum / (N-1) = +0.82

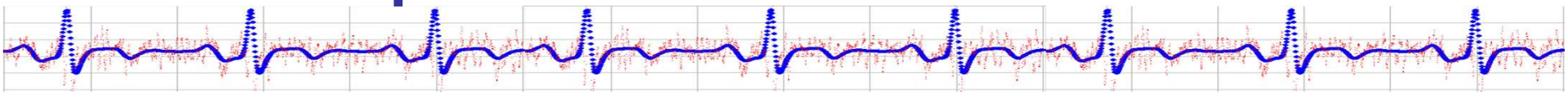
$$r > \frac{2}{\sqrt{N}} \quad =.89$$

Pearson Correlation coefficient

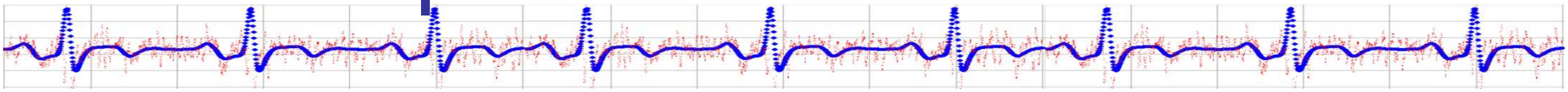


- Which of the following is it appropriate for?
  - Descriptive study designs
  - Demonstration study designs
  - Correlational study designs
  - Experimental study designs

Group Exercise



Group Exercise

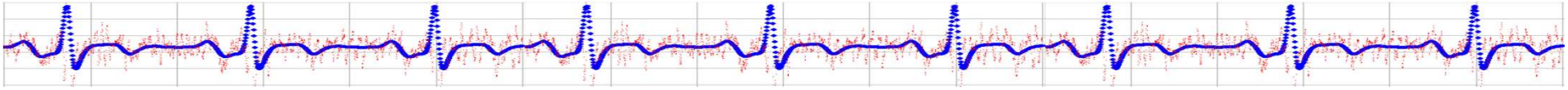


- For each problem, write
  1. Two populations being compared
  2. Research hypothesis
  3. Null hypothesis
  4. Test criteria
  5. Scatter plot
  6. r (if appropriate)
  7. Hypothesis test results
     - publication format and
     - English

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

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

Pearson Correlation in R

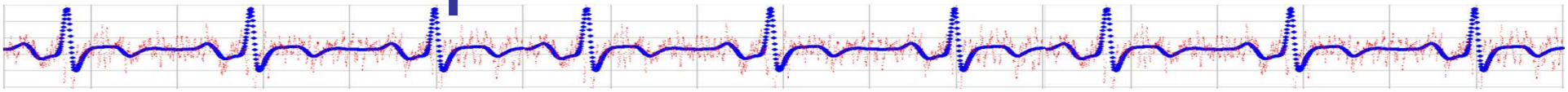


```
#For vectors v1, v2

#Just 'r'
cor(v1,v2)

#Hypothesis test (including r)
cor.test(v1,v2)
```

Example Correlation Matrix

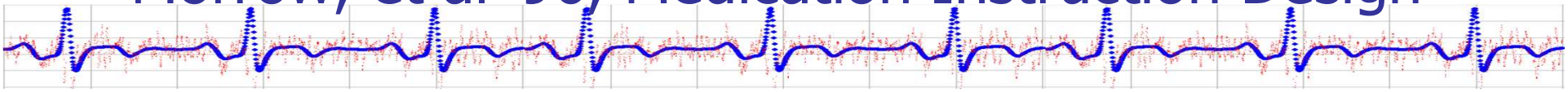


	Mean	CS	SE	EOU	TP	IC	OV
CS - Customer Support	4.7	1.00	***0.39***	***0.60***	***0.53***	***0.48***	***0.51***
SE – Security	5.0		1.00	***0.30***	***0.34***	***0.36***	***0.32***
EOU - Ease of Use	5.4			1.00	***0.49***	***0.53***	***0.62***
TP - Transactions and Payment	5.0				1.00	***0.58***	***0.49***
IC - Information Content and Innovation	5.0					1.00	***0.64***
OV - Overall Satisfaction	5.4						1.00

Table 6: *Correlation Matrix and Means (all $p < 0.05$)*

Example Correlation Matrix

Morrow, et al '96, Medication Instruction Design

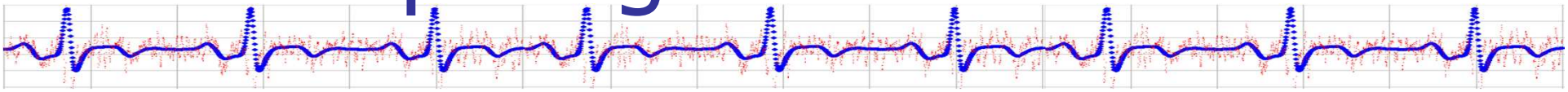


Simple Correlations among Instruction Deviation Scores, Age, Vocabulary, Number of Medications Taken, and Health Beliefs for Older and Younger Participants

	External	Need for Information	Chance	Internal	Vocab	# Meds	Instruction Deviation Scores
			Older Group (N = 42)				
Age	.04	−.35*	.07	.15	.04	.08	.13
Ext		.24	−.16	.35*	−.24	−.09	.32*
Info			−.18	−.04	−.16	.14	.10
Ch				−.08	.02	.09	−.07
Int					.26	−.32*	−.22
Voc						−.26	−.59***
Meds							.12
			Younger Group (N = 42)				
Age	−.06	.09	.23	.10	.02	.15	.13
Ext		.07	−.02	.01	.01	.01	−.15
Info			−.43**	.30*	.16	.01	.04
Ch				−.32*	−.04	.12	.16
Int					−.21	−.08	.18
Voc						.21	−.44**
Meds							.06

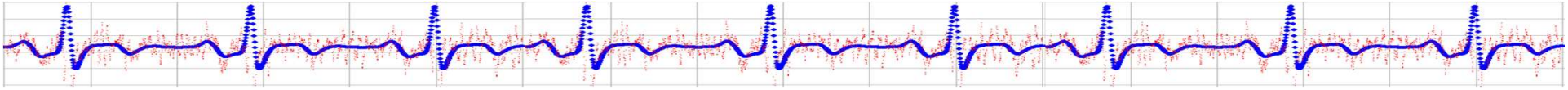
* $p < .05$, ** $p < .01$, *** $p < .001$.

Comparing r's



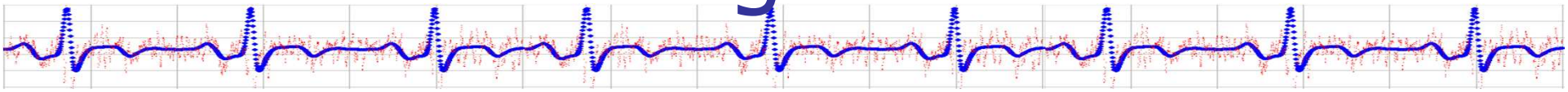
- If you want to make statements about how large one correlation is relative to another.
  - e.g. one is twice as large as another
- Don't compare r's directly…
- Compare r^2 ("proportionate reduction in error")

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

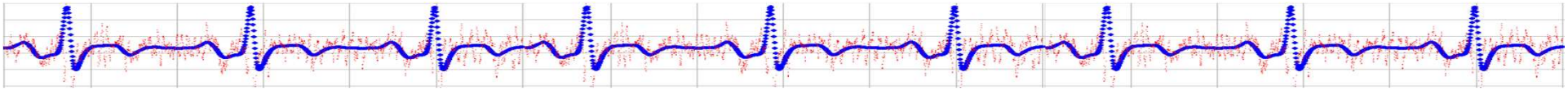
Two meanings of 'correlation'



- Correlation statistic vs.
- Correlational research model

Example: Leashes & Attachment

Example: Leashes & Attachment



- You want to see if toddlers who grow up leashed have better attachment scores.
- You recruit 30 parents of toddlers, and randomly give half of them leashes and sign contracts agreeing to leash their toddler every time they leave the house.
- After one year you administer the strange situation protocol to classify the toddler attachment as secure, avoidant, or resistant.

- What kind of study is this?
- What statistic would you use to evaluate results?
- What is df?
- Assuming X^2(df) = 32.4, what would you conclude?
- Assuming X^2(df)=0.2, what would you conclude?