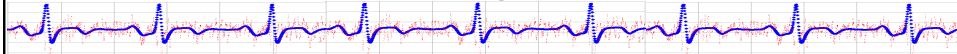


Empirical Research Methods in Information Science

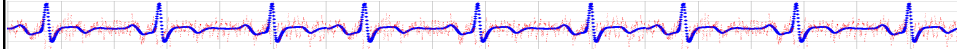
IS 4800 / CS 6350



Lecture 11 Chi-Square Leftovers Correlation

1

A common scenario



IV

DV

Software A
vs.
Software B

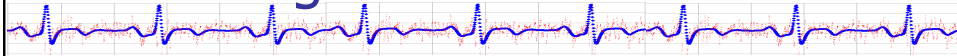
Efficiency
(time to complete)

Sampling Error
Measurement Error
Demand Effects
etc.

- We measure difference A-B
- Is this real? Or just chance?
- We have very good models of randomness (noise).
- We can compute $p(\text{observed difference A-B is due only to chance variation})$
- When should we conclude A-B is real (H1)?
- What can we conclude otherwise (H0)?

2

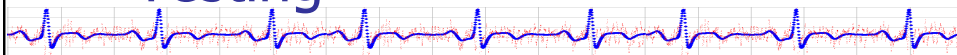
Basic Process of Hypothesis Testing



- 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 $p(\text{observed difference}|H0)$
 - 'p' = probability observed difference is due to random variation
- If $p < \text{threshold}$ then reject H0 => accept H1
 - p typically set to 0.05 for most work
 - p is called the "level of significance"

3

Type of Errors in Hypothesis Testing

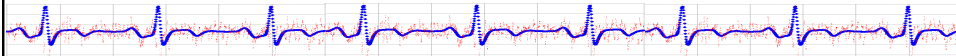


		"The Truth"	
		H1 False	H1 True
Your conclusion	Accept H1	Type I Error	Correct Decision
	Reject H1	Correct Decision	Type II Error

'p' = Probability of Type I Error

4

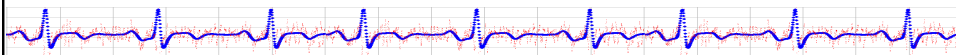
Chi-Square for Goodness of Fit



Is an observed frequency distribution significantly different from an expected distribution?

5

Chi-Square for Goodness of Fit

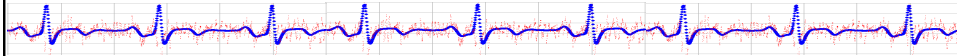


■ Assumes

1. You have a nominal variable
 - Values are exhaustive & mutually-exclusive
2. You have an *Expected Frequency* table for the nominal variable
3. None of the expected frequencies are “too small” (≥ 5)
4. Random sampling

6

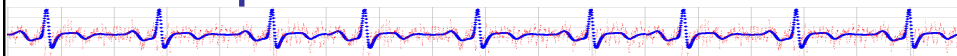
Chi-Square for Goodness of Fit



- Form of null hypothesis H_0 ?
 - Observed frequency = Expected frequency
 - Populations (expected, observed) are actually the same on the nominal measure of interest
- Form of hypothesis H_1 ?
 - Observed frequency \neq Expected frequency
 - Populations (expected, observed) are different

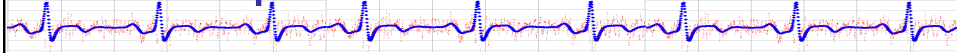
7

Chi-Square Test for Independence



8

Chi-Square Test for Independence

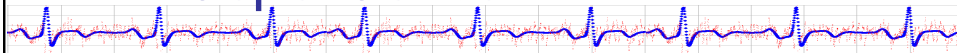


- Are two nominal variables related (H1), or are they independent (H0)?

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

9

Chi-square Test for Independence



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

10

Example from chapter

- Q1: How do you get to work?
- Q2: Are you a Morning person or a Night person?
- What kind of table is this?

	Bus	Carpool	Own Car
Morning	60	30	30
Night	20	20	40

- What kind of study is this?

11

Expected frequencies if variables are independent

- $E = (R \times C)/N$ for each cell
 - R = row count
 - C = column count
 - N = total number in all cells

	Bus	Carpool	Own Car
Morning	60	30	30
Night	20	20	40

12

Expected frequencies if variables are independent

- Step 1 – compute row & col totals

	Bus	Carpool	Own Car	
Morning	60	30	30	120
Night	20	20	40	80
	80	50	70	

13

Expected frequencies if variables are independent

- Step 1 – compute row & col totals
- Step 2 – ea cell = $(R \times C)/N$

	Bus	Carpool	Own Car	
Morning	(48) 60	(30) 30	(42) 30	120
Night	(32) 20	(20) 20	(28) 40	80
	80	50	70	

14

Formula

- Same as goodness-of-fit test.

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

- $df = (\text{NumRows}-1) \times (\text{NumColumns}-1)$

15

Text example


- Chi-Sq = 16.07
- $df = ?$

Cutoff
for $\alpha = .05$

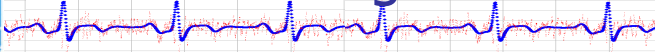
<i>df</i>	<i>cutoff</i>
1	3.84
2	5.99
3	7.82
4	9.49
5	11.07

- Conclusion?

16



Survey Feb 5, 2013 Guns in Congress




- Given the data, what questions could we ask about the relatedness of nominal measures?

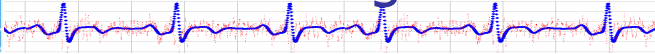
Title Member Party State GunOwner NRAGrade

- Q: Is Gun Ownership related to party?
- Q: Is Gun Ownership related to NRAGrade?

17



Survey Feb 5, 2013 Guns in Congress




- Q: Is Gun Ownership related to party?

	No	Yes
D	115	44
R	18	119

$df = (2-1) \times (2-1)$
 $\chi^2 = 102$

18

THE NATION'S NEWSPAPER

 NO. 1 IN THE USA

Survey Feb 5, 2013 Guns in Congress

■ Q: Is Gun Ownership related to NRA Grade?

	No	Yes
A	14	131
B	5	5
C	7	8
D	4	2
F	103	17

$df = (2-1) \times (5-1)$
 $\chi^2 = 155$

19

Group Exercise

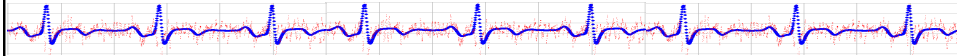
■ For each problem, write

1. What kind of study design is it?
2. Two populations being compared
3. Research hypothesis
4. Null hypothesis
5. Test criteria
6. Expected frequencies
7. Observed frequencies
8. Test results
 - publication format and
 - English

Cutoff for $\alpha = .05$	
<i>df</i>	<i>cutoff</i>
1	3.84
2	5.99
3	7.82
4	9.49
5	11.07

20

χ^2 Test for Independence in R

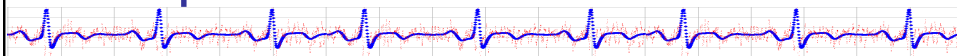


```
#use contingency table  
chisq.test(table(x,y))
```

```
#e.g., data$favcolor, data$ownmac  
chisq.test(table(data$favcolor,  
                 data$ownmac))
```

21

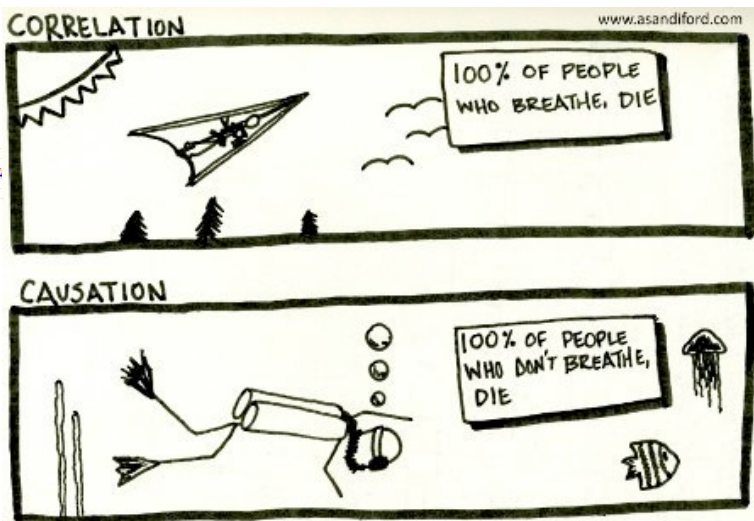
Can you use χ^2 tests to evaluate the outcomes of experiments?



Examples?

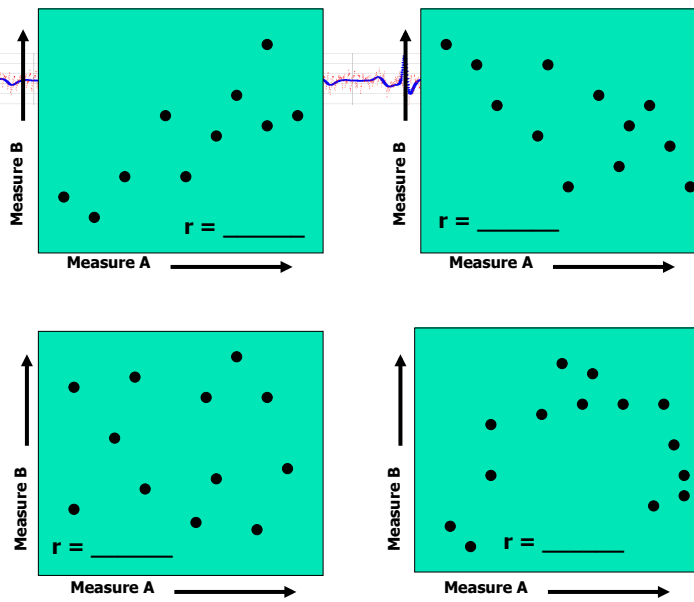
22

Correlation!



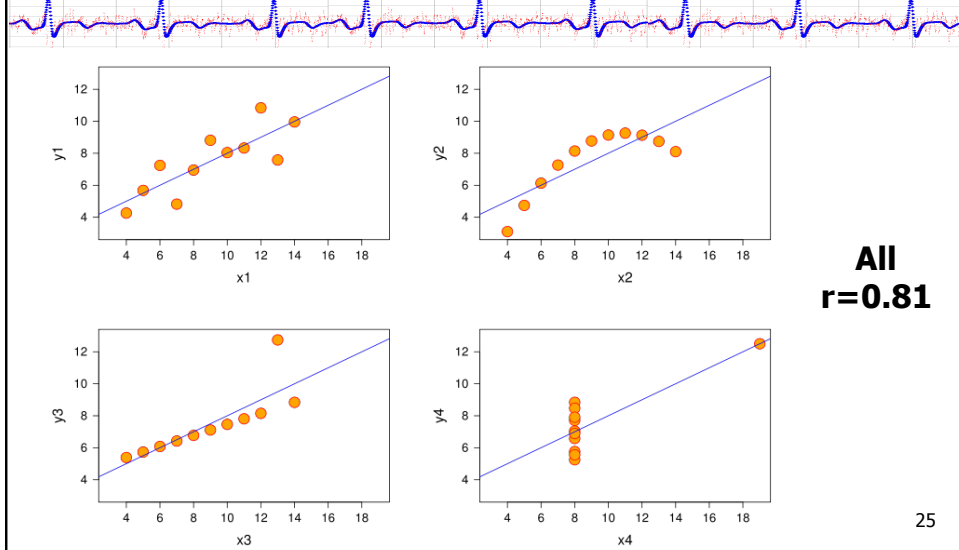
23

Quiz



24

Correlation?



Predictions Using Correlations

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

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.

27

Formula for Pearson correlation

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

-OR-

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

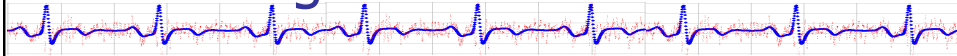
Where

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

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

28

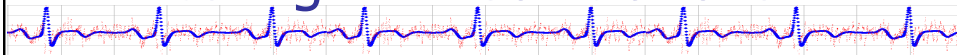
Procedure for Hypothesis Testing with Correlations



- 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.

29

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.

30

Procedure for Hypothesis Testing with Correlations

- Heuristic threshold for $\alpha=0.05$:

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

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

31

Procedure for Hypothesis Testing with Correlations

- R:
 - `cor.test(v1,v2)`
 - See if `significance < threshold`
 - Yes => reject H0
 - No => inconclusive

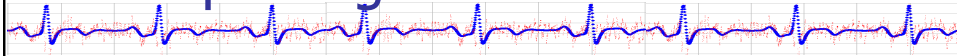
- Manually:

- Compute r
- Is
 - If yes => reject H0
 - If no => inconclusive

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

32

Reporting results



$r = val, p < sigthresh$

Where,

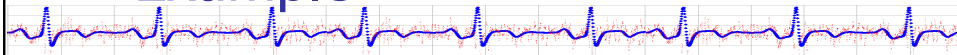
- $sigthresh$ = pre-defined significance threshold
 - Note: if $p < sigthresh$, can report that as well, e.g., "p<.01", "p=.001"

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

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

33

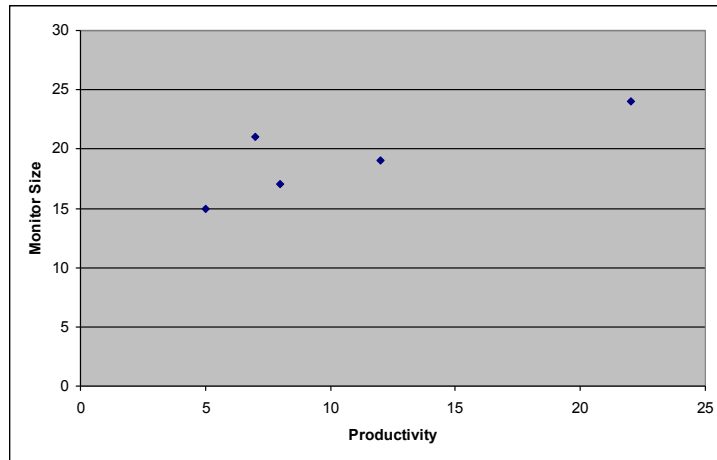
Example



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

34

Example Continued



35

Example

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

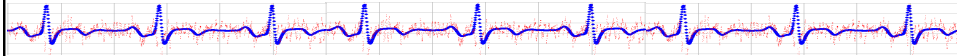
$$P = 10.8 \text{ (6.8)}, \quad M = 19.2 \text{ (3.5)}$$

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
------------------	----------	----------	----------	----------	----------

$$r = \text{sum} / (N-1) = +0.82 \quad r > \frac{2}{\sqrt{N}} = .89$$

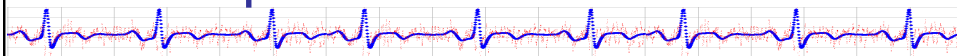
Pearson Correlation coefficient



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

37

Group Exercise



38

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

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

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

39

Pearson Correlation in R

```
#For vectors v1, v2
```

```
#Just 'r'
```

```
cor(v1, v2)
```

```
#Hypothesis test (including r)
```

```
cor.test(v1, v2)
```

40

Example Correlation Matrix

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

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

41

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")

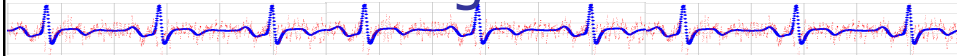
43

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 (ρ)
 - 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

44

Two meanings of 'correlation'



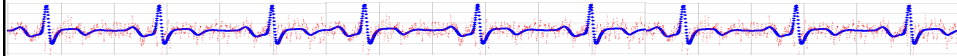
- Correlation statistic vs.
- Correlational research model

45

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 $\chi^2(df) = 32.4$, what would you conclude?
- Assuming $\chi^2(df)=0.2$, what would you conclude?

47