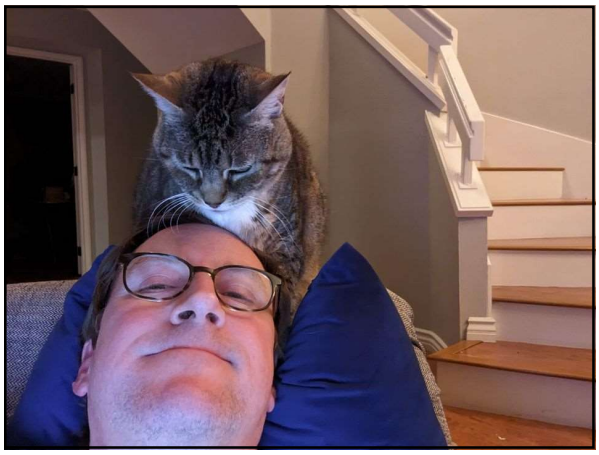


SOC497/L: SOCIOLOGY RESEARCH METHODS

Control Variables:

Causation & Elaboration

Ellis Godard

Pre-lecture dance parties

A. YES! They rock 35%

B. Sure, keep doing them 29%

C. I don't care. Please just move on. 13%

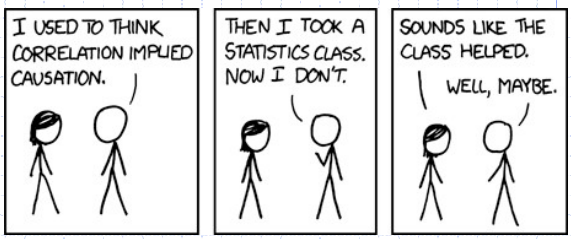
D. Ok, but dude, change the music 16%

E. OMG, no, stop, boomer, now. I can't even… 6%

The intro/setup question used to be…

Who should there be more jokes about during lecture?

1. African-Americans 0%
2. Armenians 0%
3. Asians & Pacific Islanders 0%
4. Latinos 0%
5. Rednecks 0%



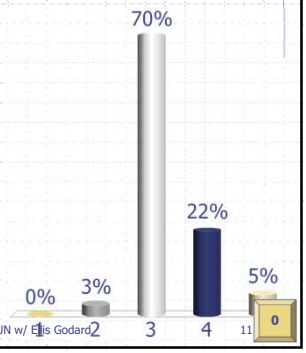
http://xkcd.com/552/

Q1. What's *methodologically* wrong w/ that question?

1. It implies a false premise 5%
2. The responses aren't similar 0%
3. The responses aren't mutually exclusive 3%
4. The responses aren't exhaustive 0%
5. All of the above 92%

Q2. How many basic statistical tests have I reviewed in lecture?

1. 1
2. 2
3. 3
4. 4
5. 5

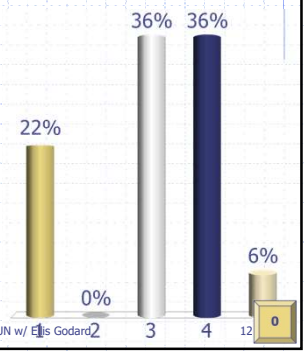


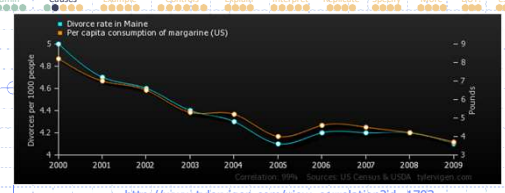
Relationships

- Definition: distribution of DV varies by IV
  - I.e. conditional distributions differ
- Could look at different DVs:
  - e.g. accounting for attitudes about abortion
- Could look at different IVs:
  - What accounts for partyid best (polyviews, educ, relig, or income)?

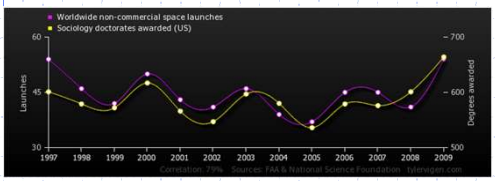
Q3. How many of those have a p value?

1. 1
2. 2
3. 3
4. 4
5. 5





http://www.tylervigen.com/view_correlation?id=1703

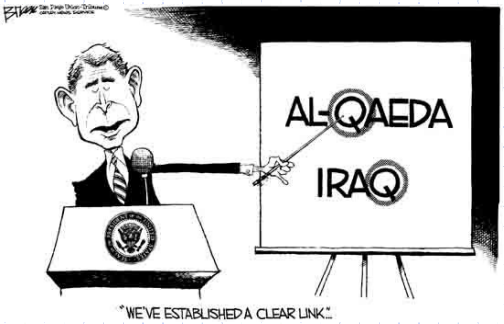


http://www.tylervigen.com/view_correlation?id=805

Outline for Today

- **Heads Up**
  - Lots to cover, meaty finish
- **Causal Relationships**
  - 3 Criteria for Claims of Causation
- **Control Variables**
  - Summary of Regression Models
  - Three-way Crosstabs – 1 new e.g., 1 from HW
- **The Elaboration Method**
  - 4 main interpretations of possible results
  - Uses, Refinements, & More

Covariation is not Causation



Criteria for Claims of Causality

1. Time order
2. Association (e.g. tau or gamma)
3. **Eliminating alternate explanations**

(Possibly a 4th: plausible mechanism or theoretical placement)

Must have all 3

- **Time order not sufficient**
  - Brush teeth before arrive here
  - Turn off lights before long vacation
- **Association not sufficient**
  - More crime in cities w/ more fire engines
- **Eliminate Alternative Explanations**
  - Can't always do, and can never eliminate all
  - Want to at least consider the *possibility*
  - In practice, often try to **control** for whatever can (e.g. for demographics, esp. if any sample bias)
- ***Possible mechanism not sufficient***
  - *Just an argument w/o consideration of data*

How many hours per day does R watch TV?

0-1 hours	24.2%
2 hours	29%
3 hours	17.6%
4-6 hours	23.8%
7 or more hours	5.5%
	100% (n=1947)

Afraid to walk at night in neighborhood?

- Yes 42%
- No 58%
- (total 100%)

Crosstab: TV by FEAR

	0-1	2	3	4-6	7
Yes	39 %	39 %	42 %	48 %	62%
No	61 %	61 %	58 %	52 %	38 %

Crosstab: TV by FEAR

	0-1	2	3	4-6	7
Yes	39 %	39 %	42 %	48 %	62%
No	61 %	61 %	58 %	52 %	38 %

$df = 4$ $x^2 = 12.57$ $p = 0.01$

Crosstab: TV by FEAR

Females only (df=4; x^2=5.53; p = 0.24)

	0-1	2	3	4-6	7
Yes	54	58	53	60	73
No	46	42	47	40	27

Males only (df=4; x^2=2.06; p = 0.72)

	0-1	2	3	4-6	7
Yes	22	21	25	28	31
No	78	79	75	72	69

Control Variables

- Sometimes part of the design
  - Experiments separate test cases from control group
- For Sociologists, typically *after* data collection
  - Approximate experimental w/ statistical control
  - Divide sample on basis of control variable
    - Two or more subsamples (groups, values, etc)
    - One subset has the control variable, one does not.

Controlled *Relationships*

- In addition to the overall ("pooled") results
  - Like what you've done in 364, lab, and HW
- Re-compute findings separately for each
  - Crosstab, t-test, ANOVA, regression, etc.
  - Do that test (IV>DV) for each value of the CV
  - These are "partial tables" (vs "pooled") with statistics about the "partial relationships"
- Compare partial relationships
  - w/ each other
  - w/ the original

Multivariate Relationships

- If there IS one, then the relationship between the DV & IV *changes* across values of the CV
- Now, in addition to having *conditional* distributions, have *partial relationships* – sets of conditional distributions that vary according to some other (third) distribution
  - crosstabs make frequencies marginal in order to look at differences within them
  - Three-ways make crosstabs marginal in order to look at differences across them

Elaboration Model

- Paradigm/Method/Model by Earl Babbie (497)
  - A.k.a. Interpretation Method, Lazarsfield Method, Columbia School
    - See esp. Robert K. Merton & Paul F. Lazarsfeld (eds). *Continuities in Social Research: Studies in the Scope and Method of 'The American Soldier'*. NY: Free Press.
- Used to understand how a relationship between two variables changes through simultaneous introduction of additional variables
- Four (4!) basic forms:
  - Explanation, Interpretation, Replication, Specification

Practical Steps in Elaboration

- **Univariate first (always)**
  - Esp. for missings, center, spread, & shape
- **Bivariate X -> Y**
  - This is the core piece of any explanation
- Bivariate X -> Z and Y-> Z
  - The control variable must be related to the other 2
- Multivariate (correlations matrix)
  - This can quickly flesh out which possible control variables simply aren't related in the dataset
- **Trivariate X->Y | Z** and Z - > Y | X

Admin Causes Example Controls Explain Interpret Replicate Specify More BUT! e.g. Etc

4 general outcomes

- Vary by whether the control variable is
  - antecedent (prior in time to both X and Y)
  - intervening (comes between them in time)
- Vary by whether they look at
  - partial relationships (subsets)
  - pooled (both)
- For each
  - Example, identification, & interpretation

Admin Causes Example Controls Explain Interpret Replicate Specify More BUT! e.g. Etc

eg1: Fire Engines Cause Damage?

Modal percentages suggest that sending more trucks causes more damage.

Damage done	One engine	2 or more
$10,000 +	20%	**60%**
< $10,000	**80%**	40%
Tot	100%	100%
N	(2000)	(1000)

Admin Causes Example Controls Explain Interpret Replicate Specify More BUT! e.g. Etc

No… *Fires* Cause Damage

of Fire Engines and Amount of Damage, controlling for # of Alarms

	One Alarm Fires		Two Alarm Fires	
	1 engine	2 or more	1 engine	2 or more
$10,000 +	30%	30%	**60%**	**60%**
< $10,000	**70%**	**70%**	40%	40%
Tot	100%	100%	100%	100%
N	(200)	(1000)	(800)	(1000)

Controlling for the size of the fire (as measured by the number of alarms), the relationship between engines and damage is ***spurious***: Bigger fires cause more damage and demand more trucks.

Admin Causes Example Controls Explain Interpret Replicate Specify More BUT! e.g. Etc

Make sure CV is related to DV *and* IV

Here, 1-alarm fires are overwhelmingly sent 1 engine, while only 2-alarm fires are likely to involve over $10K in damage

IV	1 Alarm	2 Alarms
1 Engine	**83%**	56%
2 or more	17%	**43%**
Tot	100%	100%
N	(1200)	(1800)

DV	1 Alarm	2 Alarms
< $10K	**60%**	30%
$10K+	40%	**70%**
Tot	100%	100%
N	(1200)	(1800)

Admin Causes Example Controls Explain Interpret Replicate Specify More BUT! e.g. Etc

Explanation

- Partials significantly less than in original (may be 0) & test variable is antecedent
  - e.g. w. fire trucks & damage done, CV = size of fire; w. storks & birthrates, CV = urban vs rural
  - e.g. polviews > partyid; income = CV
  - Country club membership predicts political party affiliation …
    - *but* income makes original relationship spurious (it disappears when we isolate rich and poor).
- Test/control variable "explains away" original relationship, which is called "spurious"

Admin Causes Example Controls Explain Interpret Replicate Specify More BUT! e.g. Etc

eg2: Racial Differences in IQ?

Modal percentages suggest that whites have higher IQs than blacks.

IQ	Black	White
< 100	**80%**	40%
110 +	20%	**60%**
Tot	100%	100%
N	(200)	(1000)

Well… Racial Differences in Education

Race and IQ, controlling for Amount of Education

	Not HS Graduate		High School Graduate	
	Black	White	Black	White
< 100	**60 %**	**60 %**	30 %	30 %
110 +	40 %	40 %	**70 %**	**70%**
Tot	100%	100%	100%	100%
N	(200)	(1000)	(800)	(1000)

Controlling for whether or not respondents graduated from high school, the relationship between race and IQ is ***interpretation***:
Whites are more likely to graduate from high school.

Interpretation Examples

- Examples
  - e.g. examine race->inc controlling for education
    - find that race doesn't have as big an impact for those with more education
  - Children from "broken" homes more likely to become delinquent than those from intact homes
    - But if introduce "supervision" as a control variable, no difference between delinquency rates:
    - Supervised children get into trouble at the same rate whether they come from a broken home or not. The same holds true for unsupervised children.

Make sure CV is related to DV & IV

Here, whites are much more likely to graduate high school, and those who graduate HS are more likely to score at least 110 on an IQ test

IV	Black	White
Not HS	**50%**	20%
HS +	50%	**80%**
Tot	100%	100%
N	(2000)	(1000)

DV	Not HS	HS +
< 100	**60%**	30%
100 +	40%	**70%**
Tot	100%	100%
N	(1200)	(1800)

eg3: Are retired people maladjusted?

Modal percentages suggest that retirees are more likely to be maladjusted.

Adjustment	Employed	Retired
Adjusted	**78%**	58%
Maladjusted	22%	**42%**
Tot	100%	100%
N	(172)	(98)

Interpretation

- Similar to explanation, but intervening
  - If we filter the original relationship through a control, the original relationship disappears
  - CV *clarifies* the original relationship
    - rather than denying, as in spuriousness
    - shows that relationship functions through some particular process or mechanism
  - A&F call this a "chain relationship"

Well… Maybe the community matters?

Employment status and Level of adjustment, controlling for community

	Elmira, VA		New York, NY	
	Employed	Retired	Employed	Retired
Adjusted	**78%**	**58%**	**73%**	**60%**
Maladjusted	22%	42%	27%	40%
Tot	100%	100%	100%	100%
N	(172)	(98)	(161)	(200)

Controlling for whether respondents are in the big apple or a little hamlet, the relationship between employment and adjustment is ***replicated***:
Retirees are more likely to be maladjusted, regardless of their community.

Replication

- No change. Partials same as pooled.
  - E.g. if race->educ not change by region
- Adds support to the original Ha
  - Original relationship not conditional on CV
  - X->Y replicated under test conditions

Uses of the Method

- **Resolve:** interpret 3way data (if have it, or assigned)
- **Solve:** new problems, inc. test & develop theory (explanation in the original sense of a purpose)
- **Investigate:** exploratory considerations about possible relationships among multiple variables, eg. in prep for regression or an index
  - Note: Some of you are interested in many variables.
  - Can look at elaborations to pick interesting results
  - But, scientific & ethical problems with reporting selective results

Eg 4, Real World: Crime

- Brown et al studied the use of adjudication for crimes committed by youth to see if it lowered their chances of getting into prison as adults
  - X = adjudication program
  - Y = adult prison or not
- Wanted to know if the program made a difference for some youth but not others
  - control variable Z = age when in program
- Found that age *specified* the relationship; they *specified the conditions under which* the relationship held:
  - X-Y p < .0001
  - Z-Y p = .0184
  - X-Y | Z P < .0001 for 13+ but p=.0304 for 7-12

eg5: Social class and Civil Rights?

Modal percentages don't suggest a relationship: The working class is slightly more likely to support the Civil Rights Movement (CRM), but support is low overall and the differences across rows are small.

Support for CRM ↓	Middle Class	Working Class
High	37%	45%
Low	**63%**	**55%**
Tot	100%	100%
N	(120)	(120)

Specification

- 1 partial is stronger than the other, which is less or zero (disappears)
  - E.g. educ-> inc varies by sex (payoff is higher for women than for men)
- The test control variable specifies conditions under which the original relationships holds up

Well… Maybe race matters?

Social Class and CRM support, controlling for Race

	Black		White	
CRM Support	Middle C	Working C	Middle C	Working C
High	**70%**	50%	30%	20%
Low	30%	50%	**70%**	**80%**
Tot	100%	100%	100%	100%
N	(20)	(100)	(100)	(20)

Only when we control for the race of respondents do we see that blacks are much more likely to be working class *and* to support CRM; The original relationship had been ***distorted*** **or** ***suppressed*** by the influence of race.

Suppressor Variables

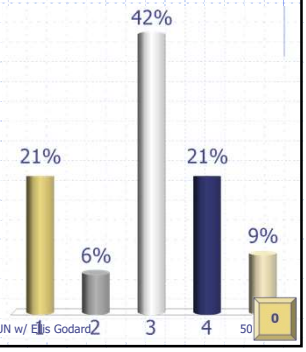
- relationship not apparent until controlled
  - (4th row, table 16-5)
- hard to find; few go in search of one
  - (e.g. age, p.399)
- May also have no original relationship until explanation or specification
  - suppressed explanation or suppressed specifier

Distorter Variables

- A variable that reverses the direction of a zero-order relationship.
  - For example, what if TV watching made women more afraid to walk at night in their neighborhood – but made men *less* afraid, *reversing* the effect of the IV
  - For example; when studying the starting salaries of men and women, the researcher was surprised to find the women receiving higher starting salaries, on average, than their male counterparts.
  - The distorter variable was "time of hire." Many of the women had been hired relatively recently, when salaries were higher overall than in the earlier years when many of the men had been hired. Only recently had women began to be hired.

Q4. How many ***basic*** forms/results of elaboration are there?

1. 1
2. 2
3. 3
4. 4
5. 5



Summary of Control Results

- **Replication** (no effect of the control)
  - relationship is the same within different categories of the control variable
  - The control changes nothing
  - (actually 4 different varieties)
- **Specification** (a differentiating control)
  - original relationship differs within categories of the control variable
  - Relationship different for different values of CV
- **Interpretation** (intervening variable):
  - CV intervenes between X&Y, & weakens relationship
  - 3rd variable follows the IV and precedes the DV
- **Explanation** (a spurious relationship)
  - 3rd variable precedes and is a cause of both the IV and DV
  - CV precedes X & Y and "explains away" relationship
- **Suppressor variable**
  - relationship stronger *within CV categories* than the original relationship was
  - Association of X and Y only seen when control for Z
- **Distorter variable**
  - Relationship reverses

Note: Not just crosstabs!

- The idea behind the elaboration method is **not tied to any particular statistic**
  - It is an analytical tool about complex conceptualized relationships
  - Could employ ("do") elaboration w/ crosstabs, gammas, t-tests, ANOVA, regression, et al.
- Crosstabs just make it easier for you to see the *partial relationships*

Warnings about Elaboration

- Theoretical caution
  - Avoid ex post facto hypothesizing
    - after the fact
  - Be careful about claims
    - Multiple causes are possible and likely
- Methodological freedom
  - No strong guidelines for what makes a partial difference significant
  - Some statistical tests may help that
    - but we won't go there

Other Methodological Limits

- Must examine *four* crosstabs (really 5)
  - For CV to affect, must be related to DV *and* IV
  - original (DxI), checks (IxC,DxC), partials (DxIxC)
- Method is more complex if
  - control isn't dichotomous
  - >1 control
- Model Needs refinements
  - Empirical variations unaccounted in original
  - Partials could = stronger, weaker, disappear, or reverse

Summary Portraits from Elab.

- Ideal type where relationship strongest
  - Paint image of person
  - Picking from characteristics specified
- Here, party id & political views most strongly related for white females who are married and went to college; and weakest for black single males who did not complete college

Example from 424's HW8

- Investigating a possible relationship between how much respondents pray and how many partners they've had
- Found that the relationship varied by gender:
  - For Men, x^2's p=0.017 and gamma was below .3 (i.e. the relationship was weak for me)
  - For Women, x^2's p=0.000 (more evidence against independence) and gamma was higher than 0.3 (moderate, i.e. stronger)
  - We can thus elaborate on the way Partners and Prayer are related, specifying conditions under which the relationship is stronger or more attenuated

Lab Exercise (optional, extra credit)

- Pick any DV, IV, & CV, from any SPSS dataset
- Run a 3-way crosstab (put CV in 3rd box)
  - Get Col %s, Chi-square, & Gamma or Lambda!
- Describe the results in terms of elaboration
  - Is there an "original"/"pooled" relationship? (sig?)
  - Are the "partial" relationships significant?
  - Do the "partials" differ from the original?
  - Which form of elaboration have you evidenced?
  - Explain the relationship among the 3 variables as clearly as possible!

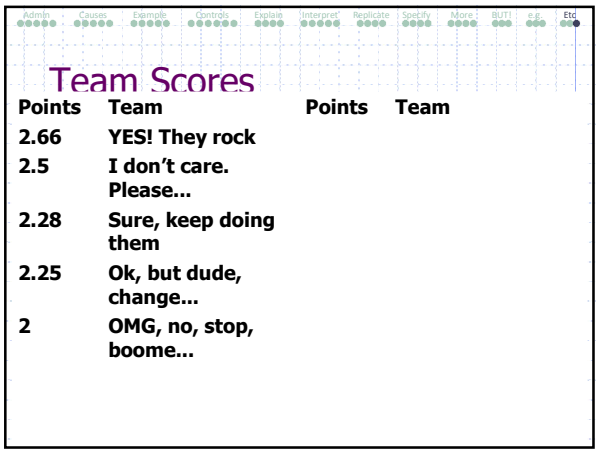
Another Example: Politics

- From POLCONTROL.SAV (GSS 80/90) – POLVIEWS & PARTYID, w/ various controls

		Chisq's p	Gamma	G's p
X-Y		.000	.322	.000
| marital	Y	.000	**.349**	.000
	N	.000	.275	.001
| sex	M	.004	.224	.002
	F	.000	**.399**	.000
| degree	< coll	.000	.216	.000
	coll +	.000	**.670**	.000
| race	W	.000	**.351**	.000
	B	**.030**	.088	**.608**

Q5. Which can be elaborated?

1. Crosstabs 10%
2. T-tests 0%
3. ANOVAs 0%
4. Regression 0%
5. All of the above 90%


Admin
Causes
Example
Controls
Explain
Interpret
Replicate
Specify
More
BUT!
e.g.
Etc
Team Scores
Points
Team
Points
Team
2.66
YES! They rock
2.5
I don't care. Please...
2.28
Sure, keep doing them
2.25
Ok, but dude, change...
2
OMG, no, stop, boome...