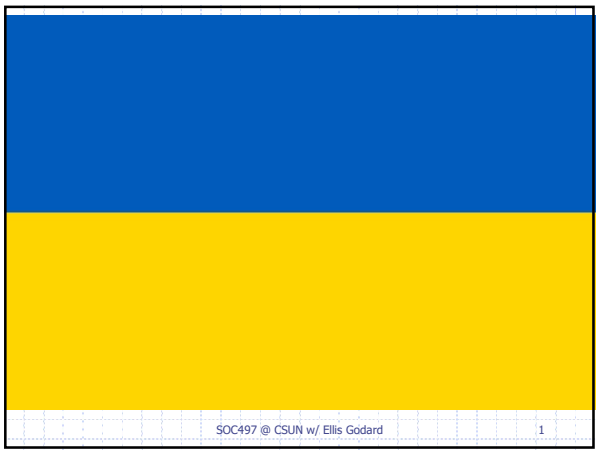

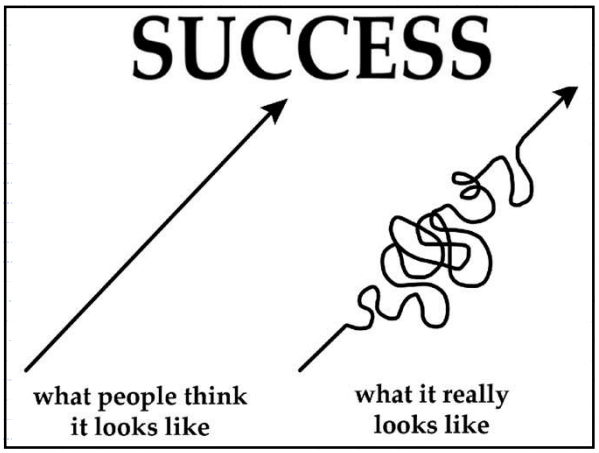

SOC497/L: SOCIOLOGY RESEARCH METHODS

Beyond Valid Values:

Indices & Scales (& Typologies)

Ellis Godard

When are you graduating?

A. Last spring
0%

B. This fall
23%

C. Next spring
60%

D. Later
17%

Reminders: ToDos…

- Still need 1 intake and 4 headshots ☹

Reminders: ToDon'ts…

- wait to draft your survey questions
- wait until night b4 midterm due to start
- forget to ID a secretary on labs (3+)
- use the "measure" column to find LOM
- use all Central Tendency or Dispersion measures for the same variable & dataset @ once
- say "prove" about any scientific idea
- say any output, ever, is "normal" – it isn't!
- call me "Godard" ☺

Slides with these, you've seen before

Outline for Today

Continuing on theme of data cleaning…

- Complex Variables
  - Indices – concepts, steps, examples
  - Scales
    - Meaning, Types, vs. Indices
  - Typologies
- Missing Values
  - 2 kinds, What's "Missing"?, 7 Solutions
- SPSS Demo & Lab

Intro to Complex Variables

- Needing multiple measures
  - Concepts have varied meanings
  - Meanings have multiple operationalizations
  - May need more than 1 measure to assess
- That's a challenge!
  - Manipulating many variables simultaneously may be tricky, difficult, inappropriate, or impossible
- 3 tools to resolve these problems (B says 2?)
  - **Indices:** accumulate scores (compute, count, IF)
  - **Scales:** score patterns or intensity structures
  - **Typologies:** label specific intersections of variables

Q1. Which involves similar patterns across multiple variables?

A. Indices 28%

B. Missing values 25%

C. Scales 25%

D. Typologies 22%

E. Beyonce 0%

Basic Concept for Indices

- Purpose: Instead of using separate variables, combine 'em
- You've done this before:
  - TOTPERWK = MILES * TRIPS * 2
- No limit to procedures or # of variables
  - Additive: ABALL = ABMOM + ABBABY + ABPOOR + ABRAPE
  - Multiplicative: Status = (income x education) / famsize
- Hint (important!):
  - Subtract the # of components with a lowest valid value is 1
  - Otherwise, minimum could be 4 (i.e. 1+1+1+1)
  - But now it will be 0 (1+1+1+1-4), which is meaningful

Advantages of Indices

1. More efficient
   - simplify/ease summary/analysis of data
2. More descriptive
   - more variation (12 dichotomies -> one 0-8)
3. More valid
   - "Better" measurement if difficult underlying concept
     - SES via measures of wealth?
     - Status via measures of prestige?
   - *Best* if concept has multiple inseparable dimensions
     - Love
     - Status
   - *But* assumes dimensions distributed in same places

Indicators of a Common Factor

- Index components should be related
  - Conceptually, not just statistically
  - Components each indicate a *factor* common to all
  - Aggregation of components measures *intensity* or *diversity* of that factor
- Example: whether someone is a good student
  - ask whether respondents agree or disagree w/ these statements:
    - 1. I attend every class
    - 2. I study every night
  - A good student should agree with both statements.
    - Both indicators reflect a good student.
    - Doing either is a good student; doing both is even "better" – more intense, more diverse, thank just doing one

Guidelines for Well-Designed Indices

Most have to do w/ *selection* of the *components*, rather than the process of actually *combining* them

- Step 1: **Item selection**
  - face/logical validity
  - unidimensionality
  - Variance – all say "if patient wants it"?
- Step 2: **Bivariate relationships**
  - Should be **correlated** with each other (r>0.2?)…
  - … but not perfectly, because then only 1 needed
  - All in **same direction**!! (*valence*)

Example w/ Valence Issue

- For all five questions please answer strongly agree, agree, disagree, or strongly disagree.
  - Q1: President Bush should not go to war.
  - Q2: Congress should authorize a war against Iraq.
  - Q3: I am against killing of other humans.
  - Q4: If American troops are sent into battle, it will be a just cause.
  - Q5: Most of my friends do not support a war against Iraq.

Q2. The valence of a question is…

A. Whether it's strong or weak 15%

B. Whether or not it's valid 15%

C. Whether or not it's reliable 6%

D. Whether it's positive or negative 27%

E. All of the above 36%

0

Step 3: Multivariate Relationships

- Babbie shows a good trivariate example
  - Review it – uses crosstabs!
- We'll use Cronbach's alpha of reliability
  - Technical Definition
    - "squared correlation between the score a person obtains on a particular scale and the score the person would have obtained on a scale of all the possible items in that scale's universe"
  - In SPSS: Analyze – Scales – Reliability Analysis
    - Just pick variables and get alpha; don't need to select any other options or stats
  - Approximate interpretation:
    - Higher the value, more reliable the index is
    - % of the time the components correspond to each other

Step 4: Scoring

A. Assign a numerical value to each indicator.
  - Agree = 1 Disagree = 0
  - SA = 5, A = 4, N = 3, D = 2, SD = 1

B. Assess how missing values will be handled
  - 7 strategies – later this lecture
  - If any case is missing a component value, can't calculate index value!

C. Determine what range is desired
  - Prefer 2-4 per component (if more, extremes get sparse), 3-9 overall
  - Otherwise, extremes get sparse
  - Hint: if addings 1s and 2s, subtract the # of measures

D. Determine how the components should be weighted
  - Equal unless compelling reason to do otherwise

Step 5: Validation

Four options…

A. Inspect sample cases
  - Look across some rows - make sense? Combine correctly?

B. Conduct item analysis
  - Statistically measure extent to which composite measure is related to or associated with the included items
    - Pearson's correlation coefficient (0.2 or higher)
    - Cronbach's alpha of reliability (?)

Step 5: Validation, cont'd

C. Seek external validation
- Use other items in dataset
  - ~ construct validity
  - euthanasia index correspond w/ abortion? Death penalty?
- If not mesh, possibilities:
  - Either the index or included items don't measure the concept
  - Sampling problem (next lecture)
- Re-examine index first
  - become more explicit about the concept – compassion? Fear of death?

D. Shortcut: use established measures!

Q3. Components of a well-designed index…

A. are perfectly correlated 3%

B. have at least 5 values each 18%

C. have consistent response valence 36%

D. relate to at least two dimensions 12%

E. all of the above 30%

30

Introducing "Scales"

- Texts
  - Scales as *levels of measurement*
- SPSS
  - Scale as *interval or ratio* measures
- Technical
  - Scale as an *intensity structure*
  - Ranks cases on a continuum
  - Combines related ordinal measures
    - Like a specific kind of index…

Indices vs. Scales

- What they have in common
  - both can be composite measures of variables (built by combining 2 or more measurements)
- How are they different
  - Indexes accumulate different kinds of scores; e.g. miles & trips
    Scales assess patterns across similar measures e.g. varied abortion attitudes
  - Scales involve ordinal *components*; Indices combine anything

Scales: A Special Case

- What level of measurement are they?
  - Textbooks often say interval
  - Seem ordinal to me (differences not equal)
  - Unsettled debate – see the "schemapiric view"
    - S Stevens, Science 30, Aug 1968, V 161, No 3844, p849-856, "Measurement Stats and the Schemapiric view"
- When in doubt, treat as both
  - E.g. consider the mean *and* median
  - Each procedure has assumptions; compare results
  - Like triangulation – alt perspective = deeper inquiry

Likert: The Holy Grail of Scales

- Informally
  - any measure with 4 to 7 categories
- Technically
  - an index of things w/ 5 or so categories
  - Series of measures w/ different criteria

Intensity Scaling

- Bogardus Social Distance Scale
  - Measures willingness to participate in social relations
  - ethanize stranger > spouse (distance)
- Guttman Scaling
  - Hard vs. easy indicators of the same concept
    - cold, HIV, coma, full-blown AIDS, severed head
  - Some items may prove more extreme indicators
    - Support life in prison for causing death while DUI
    - Support life for 2nd degree murder
    - Support life for 1st degree murder
      - Scoring would be 3, 2, 1, top to bottom

Low/High Precision Scaling

- Semantic Differential
  - Choosing between two opposite positions
    - Emotional.....Rational
    - Sensitive.....Competitive
    - Relaxed ..... Active
- Thurstone Scales
  - Attempt to define intervals on an ordinal scale
  - Uses a panel of experts to judge items and score them
  - They evaluate based on their own criteria

Tips about Scales (heh)

- Useful in explanatory work
- Require careful checks of validity & reliability
  - those checks are interesting themselves
  - can write article just on methodology
  - could do in paper
    - E.g. 2 measures of income consistent? (check w/ educ)
    - or two ways of asking polit Qs – in gss modules

When to use which?

- Scales are generally superior to indexes
- But often difficult (even impossible) to construct from available or achievable data
  - Scales require planning in advance
  - Indices are usually constructed post hoc
- Indexes are more frequently used
  - But their construction is *not* obvious or straightforward
  - Must be explicitly explained & defended

Introducing Typologies

- Labeling the intersections of 2 or more variables
- Usually nominal composites
  - Conceptual Example: three types of terrorists (Based on historical time connection, typical tactics used during attacks, motivation, and willingness to commit mass victimization.)
    - Early anarchists.
    - New terrorists.
    - Post-modern terrorists
- Could be ordinal...
  - Cellular Example: six types of students

	A	B / C	D / F
More able	Achievers	Underachievers	Failures
Less able	Overachievers	Average students	Challenged

Reminders: Valid Percent

Here, 50% of cases have "system missing" values (that is, no value – literally missing!).

Unless you want to treat not answering as a category, you should focus on valid percents

Which one of the following was the cause of the latest problem in your relationship?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Time for the relationship	8	13.8	27.6	27.6
	Different values	8	13.8	27.6	55.2
	Lack of commitment	1	1.7	3.4	58.6
	Honesty	2	3.4	6.9	65.5
	Jealousy	1	1.7	3.4	69.0
	Communication	5	8.6	17.2	86.2
	Other problems	3	5.2	10.3	96.6
	No problems	1	1.7	3.4	100.0
	Total	29	50.0	100.0	
Missing	System	29	50.0		
Total		58	100.0		

Kinds of Missing Values

- System missing
  - no data on that variable for that case
  - Indicated in SPSS w/ a period instead of a #
- User-defined missing
  - Values typically excluded from computations
    - DK, NA, NAP
  - Values not used in a particular comparison
    - E.g. if focus on 2 categories of a nominal variable

What is "Missing"?

- Missing Cases?
  - Have values that are already declared as missing
  - Listed in 1 or more rows of freq table labeled "Missing", *below* the section labeled "Valid"
  - Use valid percent!
    - Should differ from percent column
  - Note in discussions of sample size, possible biases, etc.
    - Next lecture…
- Missing Variables?
  - Doesn't mean anything
  - A variable that's not in the dataset??

What is "Missing"?

- **Missing Values?**
  - Various abbreviations
    - DK (don't know), NA (No answer), NAP (Did not apply), RF (Refusal), et al
  - Should be in "missing" column of variable view
    - Click cell then elipses ("…" in a grey box
    - List single value, up to 3, range, or range plus 1
    - Tells SPSS to exclude cases w/ those values from any statistical analysis or data displays
  - But ignoring values is not the only option…

7 Solutions for Missing Values

- Must explicitly describe & defend whatever strategy you select
- **3 options that confront missing data:**
  - **exclude that value**, esp. if few (most likely for DK,NA,NAP)
  - exclude that variable (esp if too many cases) as having insufficient observations measured
  - treat as a response category or variable (esp other values)
- **Four options "impute" a replacement:**
  - assign a random value (very risky)
  - impute middle or mean (slightly less risky)
  - interpret/imply answer from another variable (still risky)
  - assign proportion of what *do* have (works if index reliable)
    - Imagine an index with nine items, and a respondent who only answers six, 4 Yes and 2 No -> assign 2 Y's & 1 No for the others

Q4. Which do you *not* need to do?

A. Be aware of the number of missing cases 3%

B. Ensure missing values are in Missing col. 6%

C. Check for invalid values in "Valid" section 23%

D. Report the number of missing variables 19%

E. Ignore the "measure" column 48%

Strategically Waning Support

- **My help w/ missings drops slowly over the semester**
- 1st month – **Handholding:**
  - Explaining the idea & walking you through it, over & over
  - Showing 2 ways to check, 2+ ways to fix (esp. MISSING column)
- 2nd month – **Helping:**
  - Hoping that you'll check for missing values
  - Pointing out missing values you haven't dealt with
  - Expecting you to remember how to deal w/ them (check handout)
- 3rd month – **Hinting:**
  - Assuming you can find & resolve them; frustrated if you don't ☹
  - Asking "Check all your values" or "What year in college is 9?"
  - Reminding you that you've be dealing with them for 2 months
- 4th month – **Hands off:**
  - No answers, suggestions, or prodding – gotta have it down by now!
  - If you don't know the basics after 3 months, uh oh! 8^p

Lab Assignment Continuity…

- Last lab: Grouping hero types
  - Family members?
  - Entertainment figure?
  - Historical persons?
- This lab: Combining abortion attitudes
  - First, combine variables (compute)
  - Second, analyze attributes (elem stats)
  - Must go beyond last data lab –
    - tell story, summarize the data, describe the sample!

Lab Demo-ish: Indices

- **Rshows.sav**
  - Measures of whether Rs watched 8 reality shows
    - 1 if they did, 0 if they didn't
- **RSHOWS variable is an index**
  - Like TOTMILES = TRIPS * MILES * 2 (first SPSS lab)
    - Transform > Compute; name the Target Variable; write the formula; click OK
  - Computed from dichotomies about specific shows
    - RSHOWS = SURVIVOR + BIGBRTHR + REALWRLD…
    - Someone who watched all 8, RSHOWS = 1+1+1+1+1+1+1+1 = 8
    - Someone who watched none, RSHOWS = 0+0+0+0+0+0+0+0 = 0
  - But if the values had been 1&2 (instead of 0&1):
    - RSHOWS = SURVIVOR + BIGBRTHR + REALWRLD… - 8
    - Someone who watched all 8, RSHOWS = 2+2+2+2+2+2+2+2 - 8 = 8
    - Someone who watched none, RSHOWS = 1+1+1+1+1+1+1+1 - 8 = 0

Lab Demonstration: Checks

- Is there a new row in Variable view?
- A new column in Data View?
- Scan some rows in Data View
  - Check to see/illustrate that the index "worked" for a few cases
- Compare histogram (etc.) of index to original measures

Lab Exercise: Indices

- Use the abort.sav dataset from the website
- Look at frequency distributions and/or histograms of these seven measures:
  - ABANY, ABDEFECT, ABHLTH, ABNOMORE, ABPOOR, ABRAPE, ABSINGLE
- Create an index (additive; equal weights)
- Submit a freq. table & histogram of the index
- Write a few sentences describing the shape, central tendency, and dispersion of this index
  - *Not* just elementary report of data in phrases
  - Use complete sentences
  - Use the summary statistics to *describe the sample!*

Q5. Which of these is *not* a formula for an index?

A. Totmiles = Miles * trips * 2 6%

B. Rshows = Surviv + RealWrld + BigBro 21%

C. Abort = abany + abrape + abincest 6%

D. SES = rincome*educ/famsize 6%

✓ E. Any would "count" as an index 61%

Team Scores

Points	Team	Points	Team
1.79	Next spring		
1.46	This fall		
1.45	Later		