Admin · Cleaning · Computing · Recode · Missing? · Missing Solutions · Lab

Outline for Today…

- **Admin –** Info!
- **Data Cleaning Overview**
  - Inc. Data Dictionaries
- **Computing / Indices**
  - Concepts, Guidelines, 2 Demos
- **Recoding**
  - Guidelines, 2 Demos
- **Missing Values**
  - Concepts: *What* is missing?
  - Guidelines: Seven Solutions
  - Examples, Demo, & Warnings
- **Lab –** It's a Phat one! ☺

5

SOC424 @ CSUN - Ellis Godard -- Indices

Where we are….

#	Date	Read (5th)	Due	Area	Lecture Topic	Lab #	Lab Assignment	T Lab	R Lab
1	Tue Aug 26			*Orientation*	Welcome & Orientation	-			
2	Thu Aug 28	1.1 to 1.4			Basic Terms	-		-	-
3	Tue Sep 2	2.1 & 2.5			Measurement Issues	1	Levels / Age 3x	1	-
4	Thu Sep 4	3.1			Data Reduction	2	Heaven & Hell		1, 2
5	Tue Sep 9	-	HW1	*Description*	Display & Analysis (Shapes & SPSS)	3	Total Miles	2, 3	
6	Thu Sep 11	3.2 & 3.5			Central Tendency	4	CT		3, 4
7	Tue Sep 16	3.3 & 3.4			Dispersion	5	Dispersion	4, 5	
8	Thu Sep 18	3.7			Indices & Data Cleaning	6	Music index		5, 6
9	Tue Sep 23	4.2		*Inference*	[illegible]	7	Standardizing Scores	6, 7	
10	Thu Sep 25	4.3	HW2		Zs & Ps	8	Table A		7, 8
11	Tue Sep 30	3.6 & 5.1			Parameters & Pt Estimation	9	Distances	8, 9	
12	Thu Oct 2	2.2 to 2.4 & 4.3			Sampling (Issues, Methods, Effects)	10	Sampling		9, 10
13	Tue Oct 7	4.4 to 4.6			The Central Limit Theorem	11-EC	CLT/World (EC)	10, 11*	
14	Thu Oct 9	5.3		*Estimation*	Confidence Intervals	12	CI for Intervals		11*, 12
15	Tue Oct 14		HW3		CIs for Proportions	13	CI for Proportions	12, 13	
16	Thu Oct 16	6.1 & 6.4			Hypotesting & Zs	14	Writing Hypotheses		13, 14
17	Tue Oct 21				Hypotesting for Large ns	15	Two Tests	14, 15	
18	Thu Oct 23	6.3 & 6.8	HW4		The "t" test, for small ns	16	CI & Test Ages		15, 16
19	Tue Oct 28	5.4			Sample Size Estimation	17	Estimating n Needed	16, 17	
20	Thu Oct 30	7.1, 7.3, & 10.1		*Covariation*	Differences in Means	18	Comparing Means		17, 18
21	Tue Nov 4	7.2	HW5		Differences in Proportions	19	Comparing Proportions	18, 19	
22	Thu Nov 6	12.1			Analysis of Variance	20, 21-EC	ANOVA (+ MODELS EC)		19, 20, 21*
	Tue Nov 11				(Veteran's Day)				
23	Thu Nov 13	9.4 & 9.5			Scatterplots & Correlation	22	Grade Correlations		22
24	Tue Nov 18	9.1 to 9.3	HW6		Regression	23	Regression Lab	20,21*,22,23	
25	Thu Nov 20	10.2 & 11.1			Multiple Regression	24-EC	Multiple Reg (EC)		23, 24*
26	Tue Nov 25	8.1	HW7	*Association*	Crosstabulations	25	TBA (any)	24*, 25	
	Thu Nov 27				(Thanksgiving)				
27	Tue Dec 2	8.2 & p.233			Dependence	26, 27-EC	TBA (SCU) (& 27-EC)	26, 27*	
28	Thu Dec 4	pp.238 to 243	HW8		Association	28-EC	Measures of Assoc (EC)		25,26,27*,28
x	Tue Dec 9				(no lecture - work session only)			28*	
x	Thu Dec 11				(no lecture - work session only)				
x	Thu Dec 18	-	HW9		(no meetings - deadline only - 10am, firm!)				

Admin · Cleaning · Computing · Recode · Missing? · Missing Solutions · Lab

Recent Quiz Concern

What's the appropriate measure of central tendency for Grade, with the values A, B, C, D, and F?

Answer	Respondents	%
Range	4 respondents	18 %
Mean		0 %
Median ✓	8 respondents	36 %
Variation ratio		0 %
Mode	10 respondents	45 %

- 18% picked Range, which isn't a measure of central tendency
- 45% picked Mode, for an ordinal variable that has a middle
- Only a third knew ABCDF is ordinal & median's the right choice

6

SOC424 @ CSUN - Ellis Godard -- Indices

SOC424 – Statistics w/ Dr. Ellis Godard

Data Cleaning:
Computing, Recoding, & Missing Values

Admin · Cleaning · Computing · Recode · Missing? · Missing Solutions · Lab

Admin Reminders

- **Grading Sheet**
  - Read the first page – *many* grade questions are answered there
    - And don't forget to check the 1st page ☺
  - Your row gray unless you did Intake Form, said you wanted your info on the PDF, gave an approved codename, & emailed a headshot
    - Fa25: *Still* missing some intakes & headshots, after 4 (nay, 8!) weeks
- **File Folders, including:**
  - Handouts: www.csun.edu/~egodard/424/hdo
  - Notes: www.csun.edu/~egodard/424/lect
  - More in 1st day's notes – I again recommended bookmarking!

7

SOC424 @ CSUN - Ellis Godard -- Indices

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Data Dictionaries

- **Click UTILITIES –VAR INFO**
  - To get values list for every variable
- **Click UTILITIES – FILE INFO**
  - To get a "data dictionary" of the whole file

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Data Cleaning

- **Data rarely comes ready to analyze**
  - Often need to alter it for your purposes
  - Either to fit hypotheses, or for statistical procedures
- **Always requires that you document changes**
  - You (*and others*) should know exactly what you did
  - Your work should be understood, and replicable
- **Three ways to adjust (…)**
  - Create a new variable (computing)
    - combining two or more variables into a single measure
  - Deem some values meaningless (missings)
  - Reduce or remove distinctions (recoding)
    - altering the scale on which some variable is measured

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Data Cleaning You'll Do…

- **<u>Computing</u> (previous lab; review today)**
  - Any equation using any variables in the dataset:
    - MEDIAUSE = TVHRS + PHOHRS + CYBHRS + READHRS
    - HOUSWLTH = (FAMINCOM * RESPEDUC) / FAMSIZE
    - MUSIC = BLUES + JAZZ + REGGAE + COUNTRY + …
  - Creates an Index
- **<u>Recoding</u> (next lab; new today)**
  - Data reduction (of level of measurement)
  - Guidelines for Cutoffs
  - Consistent valence (for comparisons or indices)
  - Missing values
    - System missing vs. User-defined
    - Solutions – at least seven…
- **<u>Missings</u> (this lecture, demo, & going fwd)**

<u>NOT</u> related

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Indices (Indexes?) in a Nutshell

- **Logic**
  - Instead of using separate variables, combine 'em
  - You did this, with MILES & TRIPS
- **Procedure**
  - Tell SPSS how to combine 'em
    - E.g. Totmiles = MILES * TRIPS * 2
  - Can use *any* mathematical procedure(s) and # of variables
    - Additive: MEDIAUSE = INSTAHRS + FBKHRS + TKTKHRS
    - Multiplicative: Status = (income x education) / famsize

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Indices – Advantages

1. **More efficient**
   - Simplify/Ease the summary & analysis of data
2. **More descriptive**
   - More variation (one-way miles vs weekly & roundtrip)
3. **More valid / "better" measurement**
   - Esp. if underlying concept is difficult-to-grasp w/ a single variable
   - Love, Class, Status, Deviance, Religiosity

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Index Demonstrations

- **RSHOWS and MEDIAS will be indices**
  - Like TOTMILES = TRIPS * MILES * 2 (first SPSS lab)
    - Transform > Compute; name the Target Variable; write the formula; click OK
  - But <u>*additive*</u> (not multiplicative) – adding up 1s&0s
    - Note that 1s&2s are more common… more on that soon…
- **Always follow up!**
  - Is there a new row in Variable view? A new column in Data View?
  - Scan some rows in Data View to see/illustrate that worked
  - Compare histogram (etc.) of index to histogram of original measures

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Demo 1: Indices *with* Dummies

- **Computing-Rshows.sav**
  - Measures of whether Rs watched 8 reality shows
    - Each is a "dummy variable" – a binary toggle, 0 or 1
    - 0 if they didn't watch it, and 1 if they did, for each of 8
- **RSHOWS is an index, of all 8**
  - Computed by adding those binary (dichotomous) measures
    - RSHOWS = SURVIVOR + BIGBRTHR + REALWRLD…
    - What will the total be for someone who watched none of them?
      - RSHOWS = 0+0+0+0+0+0+0+0 = 0
    - What will the total be for someone who watched all 8?
      - RSHOWS = 1+1+1+1+1+1+1+1 = 8

14 SOC424 @ CSUN - Ellis Godard -- Indices

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Demo 3: Recoding (combining *values*)

- **Create data set with variable UNITS**
- **Sample first row & enter data**
  - How many class units are you taking this semester? (Leave 1 missing)
- **Look at descriptive statistics**
  - Shape, central tendency, dispersion
  - Histogram, frequency table, & summary table (reverse order)
- **Recode into 4 categories**
  - TRANSFORM > RECODE INTO **DIFFERENT** VARIABLE
  - 0=none, 1-5=light, 6-9=medium, 10-12=regular, 13&up=heavy
  - Handout on website – key = "Old & New Values", one new value @ a time
- **Follow-up:**
  - New column in Variable view, new row in Data view
  - Check a few cases (rows) to be sure it worked
  - Check descriptive statistics/output for the new variable

17 SOC424 @ CSUN - Ellis Godard -- Indices

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Demo 2: Indices *without* Dummies

- **Computing-Medias.sav**
  - Measures of whether Rs uses 8 social media platforms
    - This time, **1** if they didn't, **2** if they did (instead of 0&1)
    - These are NOT dummy variables – "none" isn't all 0s anymore; it's all 1s!
- **MEDIAS is still an index, like RSHOWS**
  - Computed by adding 8 binary (dichotomous) measures
  - But now *just* adding them doesn't give a meaningful value:
    - If someone used none, adding eight 1s would get 8 (1+1+1…), but "8" isn't "none"
    - And if someone used all 8, adding 2+2+2+2+2+2+2+2 makes 16?? Confusing!
    - We need the extremes (and the range, and every *value)* to be meaningful
  - So, instead, subtract 8, because 8 out of 8 start at 1 instead of 0
    - MEDIAS = FACEBOOK +TWITTER + INSTAGRAM… **- 8**
    - For someone who uses all 8, MEDIAS = 2+2+2+2+2+2+2+2 - 8 = 8
    - For someone who uses none, MEDIAS = 1+1+1+1+1+1+1+1 - 8 = 0

In a nutshell: When adding non-dummy index components (variables), **subtract the number of them for which the lowest valid value is 1**

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Reminders: Valid Percent

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		

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

18 SOC424 w/ Dr. Ellis

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Recoding a Variable

- **NOT "recoding" – that's a *different word***
- **NOT the same as computing**
  - That was combining *variables* (miles*trips*2 or abany+abpoor); this is combining *values* of *one* variable (1-8 <HS; 12 = HS; etc)
- **NOT part of missing values (that's next; you'll see…)**
- **Three potential guidelines for recoding**
  - **Meaningful Division**
    - e.g. grade levels & likely degree (HS, BA)
  - **Convenient Division**
    - graphical or distributive; focus on "humps"
  - **Consistent Division**
    - E.g. even thirds? Fourths? - use cumulative percent column

16 SOC424 @ CSUN - Ellis Godard -- Indices

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Two *Kinds* of Missing Values

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

SOC497 @ CSUN w/ Ellis Godard

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

What is "Missing"?

- **Missing Cases?**
  - Count @ top of output, with Valid too
  - Have values that are already declared as missing
  - Listed in row(s) 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.
- **Missing Variables?**
  - A variable that's not in the dataset??
  - That's not countable – it's infinite
  - Not something you should or could worry about

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

7 Solutions, cont'd...

- **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 an index of the measures is 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 others
      - But what if they've smoked pot and tobacco, drank beer, and tried shrooms, but not LSD or coke; which of the other three (meth, PCP, and ketamine) does it make sense to say they've had?

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

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) to see any declared
    - 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 those values is not the only solution…

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

What *You* Should Do

- **#1 - "Exclude that value" – almost *always!***
  - **Do NOT recode** unless you have a good clear reason to, & some strategy for what the old and new values would be.
  - **Recoding is NOT part of handling missing values;** *changing* the values doesn't tell SPSS to *exclude* values
- **Look for value labels in "Values" that aren't valid**
  - DK, RF, NA, NAP, others?
- **If there are any, make sure that…**
  - they're indicated in the "Missing" column,
  - the frequency distributions distinguishes "Valid" & "Missing" lists,
  - only valid values are in the "Valid" list,
  - any others (DK, etc.) are in the "Missing" list, and
  - the "Percent" and "Valid Percent" are different.

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

7 Solutions for Missing Values

- **Must explicitly describe & defend strategy chosen**
- **3 options confront missing data directly**
  - Exclude that value, esp. if few cases affected
    - Most likely route for DK, NA, NAP, RF values
    - Find those labels in Values column, and put them in Missing column
    - Do NOT just remove the labels FROM Values – that hides what thye *mean*
    - That's **what you almost always do** in this class
  - **If do something else (recode?), better have a good reason!**
    - Seriously! You will probably *not* recode. *(You didn't… right??)*
  - Exclude that variable, esp. if too many cases
    - Insufficient observations measured to study that measure
  - Treat missings as a response category or variable
    - Could be an interesting study on its own – could focus *solely* on the missing cases, with a missing value as the DV!

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Questions to Ask Yourself (& see FAQ!)

- In the "Values" column in Variable View, are there any non-valid values (DK, NA, RF, NAP, etc.)?
  - If so, are in "Missing" column?
- In the frequency table(s), are there separate groupings for "Valid" and "Missing" values?
  - There might not *be* any missing cases
- In the frequency table(s), are there any values listed under "Valid" that should be listed in the "Missing" group?
  - If so, you didn't add them in the "Missing" column
- Do NOT recode w/o a good, clear reason
  - Including strategy for old and new values
  - Almost certainly NOT what you should be doing for missing values

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Strategically Waning Support

- **My help w/ missings will drop 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
  - If you don't know the basics after 3 months, uh oh! 8^p

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

In previous labs…

- **Developing parallel skills**
  - SPSS
    - Created data file
    - Created a simple index (w/ COMPUTE)
    - Frequency distribution & histogram
    - Measures of central tendency & dispersion
  - Statistics
    - Understanding variables
    - Choosing and interpreting output
- **Want to continue to develop both**
  - New SPSS Skills, & Better Statistical analysis

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Demo 4: Declaring Missing Values

- **Computing-rshows.sav again**
- **Look at frequency distributions for YRBORN and PCTYPE**
- **Both have a value of 99 for "not answered"**
- **Must declare that "missing"**
  - We don't have a "real" value – not a ***valid*** value
  - We want SPSS to ignore that value, and that case

Admin Cleaning Computing Recode Missing? Missing Solutions Lab

For Your Next Lab: music.sav

- **12 measures of how much Rs like each of 12 musical genres**
  - How much they like each, on a 1-5 scale ->
  - But also 3 "missing values" (NAP, DK, NA)
- **You will "clean" data in 2 ways**
  - For each variable, Recode 4 values into 2
  - Then compute a new index that combines the 12 new dichotomies
  - Note: You can *ignore* missings here *if* you follow handout instructions!
- **Handout on Canvas – MusicLab.pdf**
  - *If* you follow the instructions, you'll get an index from 0-12
  - As you work, think about what the values mean – what's a 0? 12? 4?

```
0.00 = "NAP"
1.00 = "LIKE VERY MUCH"
2.00 = "LIKE IT"
3.00 = "MIXED FEELINGS"
4.00 = "DISLIKE IT"
5.00 = "DISLIKE VERY MUCH"
8.00 = "DK  MUCH ABOUT IT"
9.00 = "NA"
```


Admin Cleaning Computing Recode Missing? Missing Solutions Lab

Demo 5/b: Missings w/ Outlier (if time permits)

- Look at a histogram & mean for AGE
- **Note & Address the 99 (DK) as missing**
- Look at a histogram & mean for AGE
- **Declare the value of 32 as missing for AGE**
- Look at a histogram & mean for AGE again
- **Assess effect of outlier on mean *(3rd minus 2nd)***