Turning Raw Data into Analysis-ready Data The Basics of Data Entry & Management

Wan Nor Arifin

Unit of Biostatistics and Research Methodology, Universiti Sains Malaysia.

email: wnarifin@usm.my

26 June 2018

- Common raw data, proforma and response options.
- How to turn that into an analysis-ready format.

- Continuous: Age, SBP etc.
- Count: number of admission, number of dengue cases etc.

Nominal:

- Common: Gender, Race.
- Knowledge question in KAP: Yes/Know/Don't know.
- ► Visual Analog Scale (continuous line): No Pain ↔ Worst Pain Ever!

Ordinal:

- Common: Education level, Cancer staging, Disease severity.
- Likert-type scale in KAP: Strongly Disagree/Disagree/Neutral/Agree/Strongly Disagree.
- Pain Rating Scale: No Pain 1-2-3-4-5-6-7-8-9-10 Worst Pain Ever!

Samples

MAKLUMAT SOSIODEMOGRAFIK

1.	Umur:	tahun						
2.	Jantina:		Lelaki		Perempuan			
3.	Bangsa:		Melayu Cina India		Bumiputera Sabah Bumiputera Sarawak Lain-Lain:			
4.	Taraf perkahwinan:		Belum Berkahwin Berkahwin Janda		Balu Duda			
5.	5. Jika berkahwin, nyatakan bilangan anak :							
6.	6. Pendapatan kasar bulanan: RM							
7.	 7. Tahap pendidikan: Tidak bersekolah Sek. Rendah (darjah 1-6) Sek. Men. Rendah (Tingkatan 1-3) Sek. Men. Atas (Tingkatan 4-5) Tingkatan 6/Sijil/Diploma Lain-Lain, vatakan 							

Figure 1: Sociodemographic Information

Samples

BAHAGIAN C: SIKAP

Arahan: Sila nyatakan tahap persetujuan anda bagi kenyataan berikut dengan menandakan $(\sqrt{)}$ pada skala yang sesuai.

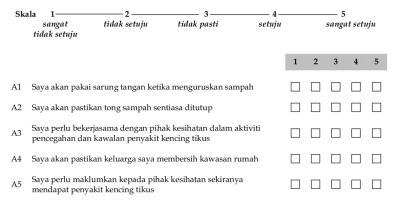


Figure 2: Likert-type Scales

Samples

BAHAGIAN B: PENGETAHUAN

Arahan: Tandakan ($\sqrt{}$) pada pilihan jawapan anda. Jawab semua soalan.

1.	Pernahkah anda mendengar tentang p	enyakit kencing tikus?	🗌 Ya	🗌 Tidak
2.	Jika ya, anda mendengarnya melalui?	 Televisyen/Radio Surat Khabar Doktor Lain-Lain, nyatakan_ 	☐ Majalah ☐ Ceramah K ☐ Poster/Pan	

	Pernyataan	Betul	Salah	Tidak Tahu
K1	Penyakit kencing tikus juga dikenali sebagai penyakit leptospirosis			
K2	Penyakit kencing tikus disebabkan oleh kuman			
K3	Penyakit kencing tikus adalah penyakit haiwan yang boleh menjangkiti manusia			
K4	Penyakit kencing tikus boleh dikesan melalui ujian darah			

Figure 3: Mixed Scales

Wan Nor Arifin (USM)

Turning Raw Data into Analysis-ready Data

- Generally, one question/item, one variable.
- Except, list of presence/absence of disease multiple options possible. Common question in medical research, but a very common mistake in data entry → unable to analyze data.
- Variable names
 - CAPITAL letters.
 - "_" in place of space in names, DISEASE_STATUS, FOLLOWUP_1 etc.



- Setup data entry sheet e.g. MS Excel properly for data entry.
- Variable types:
 - Numerical
 - ★ Continuous set cell with decimal points.
 - ★ Count set cell with no decimal point.
 - Categorical
 - ★ Decide on coding.
 - * Set up the codebook, i.e. what the codings/numbers stand for:

1	Registrat	tion Number:					
2	Name: _						
4	Gender:	\Box Male \Box Fe	male				
5	Race: 🗆	Malay 🗆 Chir	nese 🗆 Ir	ndian 🗆	Thai		
6	Number	of admission:			times.		
0	Chronic	Kidney Diseas	e Stage:	$\Box 1 \Box$	2 🗆 3A	🗆 3B 🗆	4 🗆 5

Codebook:

- **1** RN: Patient's registration number.
- NAME: Patient's full name.
- AGE: Age on present admission in years.
- **9** GENDER: Gender (MALE = 1, FEMALE = 2).
- **3** RACE: Race (MALAY = 1, CHINESE = 2, INDIAN = 3, THAI = 4).
- O ADM: Total number of admission to the hospital since 2000.
- CKD: CKD Stage on present admission (1 = 1, 2 = 2, 3A = 3, 3B = 4, 4 = 5, 5 = 6).

Table 1: Data in columns.

RN	NAME	AGE	GENDER	RACE	ADM	CKD
HRPZ/020391	Ali	20	1	1	5	4
HRPZ/015561	Ganesan	45	1	3	3	5
HRPZ/017754	Cheong	42	1	2	2	3
HRPZ/015432	Botsakorn	49	2	4	7	6

- In MS Excel:
 - set up the cell to accept only certain data type e.g. number, text, date, specific decimal place, dropdown menus etc.
 - Built in Google Form when we set up the form.

- Practical sessions.
- Data entry in:
 - Spreadsheet program e.g. MS Excel, LibreOffice Calc etc.
 - Google Form.

Thank you!