Variable Types

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

Objectives of this workshop

1. Overview of:

   - Variable types.
   - Data format/layout for common statistical analyses.
   - Common data proforma/raw data & how to turn into an analysis-ready format.

2. Practical:

   - Data entry in spreadsheet e.g. MS Excel.
   - Preparing Google Form for data entry.

Introduction

- Statistical analysis requires data.
- Data must be in suitable form for analysis.
- Else GIGO! **G**arbage **I**n **G**arbage **O**ut.

Introduction

- Important to turn our raw data into a format that computer/software can understand.
- Usually, there are standard ways of data entry.
- Quite standard among statistical software.

Introduction

- Dependent on the planned analysis.
- Data entry (i.e. those who prepare the template for data entry) requires knowledge of basic statistics...
- and also knowledge of how computer read data for analysis.

Variable types

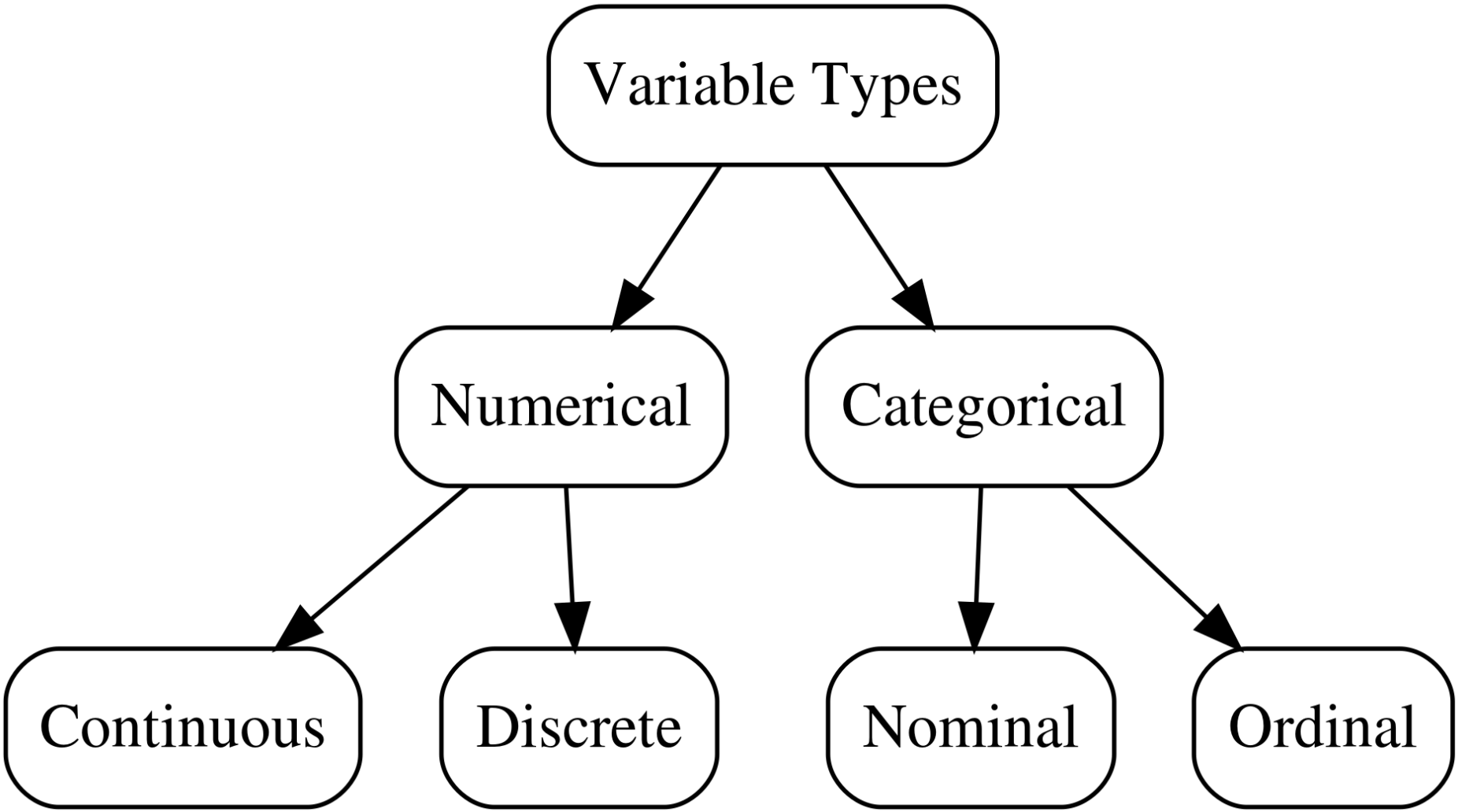

Figure 1: Variable Types

Numerical

Quantitative variable.

- Continuous.
  - Continuity in value.
  - Fractions/decimals possible.
  - Weight 67.8kg, 1/2kg.
  - SBP 120. Temperature 35.5°C etc.
- Discrete.
  - Count. Number of something.
  - 10 patients. 20 hospitals etc.
  - Fractions/decimals impossible.
  - Can't be 1/2 patient! or 0.25 hospital.

Categorical

Qualitative variable.

- Nominal.
  - No order.
  - e.g. Gender: Male/Female. Diabetes: Yes/No etc.
- Ordinal.
  - Order.
  - e.g. Cancer staging: I/II/III. Education level: Primary/Secondary/Tertiary etc.

Categorical

Nominal & Ordinal variables need coding! Must turn into numbers that computer can understand.

- Yes/No = 1/0 $\rightarrow$ 2 levels.
- Male/Female = 1/2. Can also be coded as 1/0, i.e. Male = Yes/No.
- Stage I/II/III = 1/2/3? $\rightarrow$ 3 levels. 1/2/3 needs extra coding, called dummy variables. Can be automated in most statistical software, to turn 1/2/3 $\rightarrow$ 1/0s.

Some software is able to read data as it is e.g. Male/Female, I/II/III; no need to assign codes to the categories. But, not a good practice because it is error-prone and software-dependent practice.

Common layouts

Depend on the statistical analyses:

- Grouped data - independent groups
- Paired/repeated data - same subjects

Independent groups

Revision:

- Comparison of statistics between **independent/unrelated** groups.
- e.g. Treatment vs Control groups, Diabetic vs Non-diabetic etc.
- Comparison of **means**:
  - Independent t-test (2 groups).
  - ANOVA (3 groups or more).
- Comparison of **proportions/percentages**:
  - Chi-squared test/Fisher's exact test.

Independent groups

Table 1: Independent t-test/ANOVA

FBS	Rx_Group
5.8	1
7.6	1
...	...
10.5	2
8.8	2
...	...
7.6	3
6.8	3
...	...

Independent groups

Table 2: Chi-squared test/Fisher's exact test

Gender_Male	IHD_Yes
1	1
1	0
...	...
0	0
0	1
...	...

Independent groups

Table 3: Chi-squared test/Fisher's exact test

Smoker_Yes	Severity_of_Chest_Infection
1	1
1	2
1	3
...	...
0	1
0	2
0	3
...	...

Paired/Repeated data

Revision:

- Comparison between statistics that belong to the same subjects.
- e.g. Before-after Treatment, Baseline-1st Followup-2nd Follow up-. . . etc.
- These are called paired, repeated measurements.
- Comparison of **means**:
  - Paired t-test (Pre-Post; 2 repetitions).
  - Repeated Mesures ANOVA (Pre-Post 1-Post 2-Post 3; 3 or more repetitions).
- Comparison of **proportions**:
  - McNemar's test (2 repetitions, Yes/No).
  - Cochran's Q (3 or more repetitions, Yes/No).
  - More e.g. Marginal Homogeneity test, Generalized Estimating Equations (GEE).

Paired/Repeated data

Table 4: Paired t-test

FBS_Pre_Rx	FBS_Post_Rx
5.8	5.9
8.9	7.8
...	...

Paired/Repeated data

Table 5: Mc Nemar's test

Lesion_Pre_Rx	Lesion_Post_Rx
0	1
0	0
...	...
1	0
1	1
...	...

Next session...

- We will have a look at the common raw data, proforma and response options and how to turn that into an analysis-ready format.

Thank you!