Data and Variable

Kongmany Chaleunvong

GFMER - WHO - UNFPA - LAO PDR

Training Course in Reproductive Health Research

Vientiane, 24 September 2009

Data

- Knowledge is what we know well.
- Information is the communication of knowledge.
- In every knowledge exchange, there is a sender and a receiver.
- The sender makes common what is private, does the informing, the communicating. Information can be classified as **explicit and tacit** forms.
- The explicit information can be explained in structured form, while tacit information is inconsistent and fuzzy to explain. Know that data are only crude information and not knowledge by themselves.

Data…

- Data is known to be crude information and not knowledge by itself.
- The sequence from data to knowledge is: **from Data to Information, from Information to Facts, and finally, from Facts to Knowledge**.
- Data becomes information, when it becomes relevant to your decision problem.
- Information becomes fact, when the data can support it. Facts are what the data reveals.
- However the decisive instrumental (i.e., applied) knowledge is expressed together with some statistical degree of confidence.

Data…

- Fact becomes knowledge, when it is used in the successful completion of a decision process.
- Once you have a massive amount of facts integrated as knowledge, then your mind will be superhuman in the same sense that mankind with writing is superhuman compared to mankind before writing.
- The following figure illustrates the statistical thinking process based on data in constructing statistical models for decision making under uncertainties

VARIABLE

A **VARIABLE** is a characteristic of a person, object or phenomenon which can take on different values. These may be in the form of numbers (e.g., age) or non-numerical characteristics (e.g., sex).

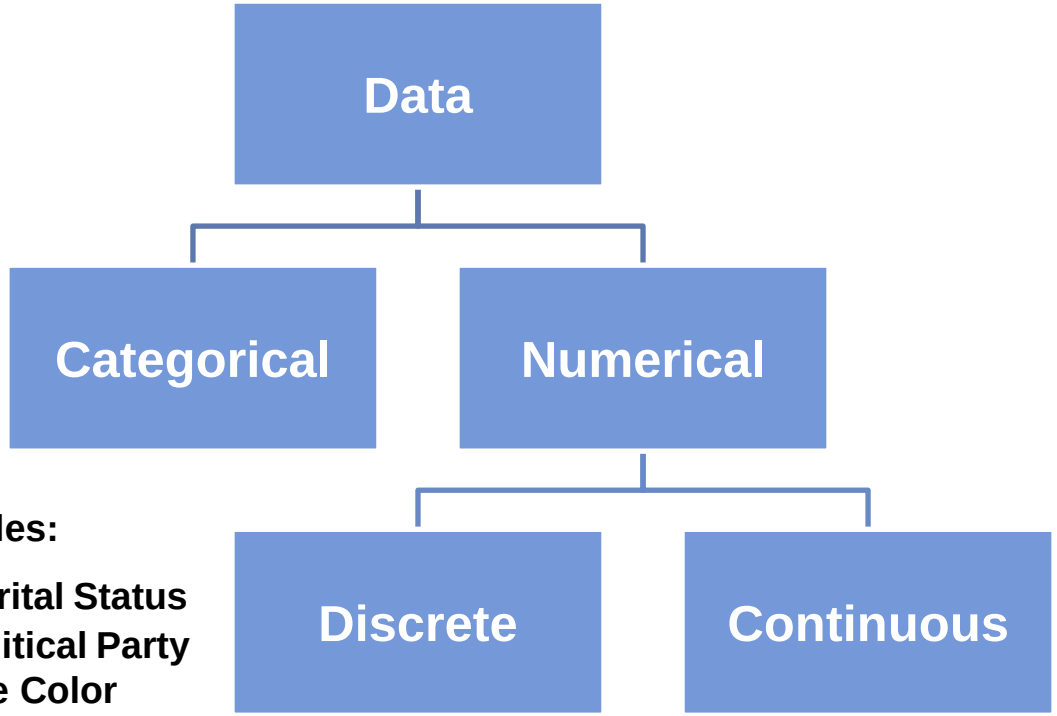
- A simple example of a variable in the form of numbers is 'a person's age'. The variable 'age' can take on different values since a person can be 20 years old, 35 years old and so on. Other examples of variables are:
- weight (expressed in kilograms or in pounds);
- home - clinic distance (expressed in kilometers or in minutes walking distance);
- monthly income (expressed in dollars, Kip); and
- number of children (1, 2, etc.).

Types of Variables

- **Categorical** (qualitative) variables have values that can only be placed into categories, such as “yes” and “no.”

- **Numerical** (quantitative) variables have values that represent quantities.

Types of Variables



Examples:

- **Marital Status**
- **Political Party**
- **Eye Color**

(Defined categories)

Examples:

- **Number of Children**
- **Defects per hour**

(Counted items)

Examples:

- **Weight**
- **Voltage**

(Measured characteristics)

Scales of Measurement

- Scales of measurement include:
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- The scale determines the amount of information contained in the data.
- The scale indicates the data summarization and statistical analyses that are most appropriate.

Nominal

- A **nominal scale** classifies data into distinct categories in which no ranking is implied.
- Data are labels or names used to identify an attribute of the element.
- A nonnumeric label or a numeric code may be used.

Example

- Variable sex: Male, Female
- Variable Occupation: Farmer, Officer, Student …

Ordinal

- An **ordinal scale** classifies data into distinct categories in which ranking is implied
- The data have the properties of nominal data and the order or rank of the data is meaningful.
- A nonnumeric label or a numeric code may be used.

Example:

- Variable Product satisfaction: Satisfied, Neutral, Unsatisfied

Variable Student Grades: A, B, C, D, F

Interval and Ratio

- An **interval scale** is an ordered scale in which the difference between measurements is a meaningful quantity but the measurements do not have a true zero point.

- A **ratio scale** is an ordered scale in which the difference between the measurements is a meaningful quantity and the measurements have a true zero point.
- This scale must contain a zero value that indicates that nothing exists for the variable at the zero point.

Interval and Ratio data are always numeric.

Interval and Ratio

Numerical Variable	*Level of Measurement*
Temperature (in degrees Celsius or Fahrenheit)	Interval
Standardized exam score (e.g., ACT or SAT)	Interval
Height (in inches or centimeters)	Ratio
Weight (in pounds or kilograms)	Ratio
Age (in years or days)	Ratio
Salary (in American dollars or Japanese yen)	Ratio

Qualitative and Quantitative Data

- Data can be further classified as being qualitative or quantitative.
- The statistical analysis that is appropriate depends on whether the data for the variable are qualitative or quantitative.
- In general, there are more alternatives for statistical analysis when the data are quantitative.

Statistical Data

- Categorical data is by nature always grouped.
- Classes for categorical data are usually single valued.
- Numerical Data can be gathered as grouped or converted after gathering.
- Numerical data is usually grouped for graphical presentation.
- Classes for numerical data are usually a range of values.

Example Data

Subject	Sex	Age(year)	Marital status	Education	Coronary	Smoking
1	Male	32	Single	Illiterate	No	No
2	Female	25	Married	Primary school	No	Yes
3	Male	50	Divorce	Illiterate	Yes	Yes
4	Female	49	Married	Primary school	No	No
5	Male	30	Married	Primary school	No	No
6	Male	47	Married	Vocational	Yes	Yes
7	Male	38	Married	High school	No	No
8	Male	35	Married	Illiterate	Yes	Yes
9	Female	28	Single	Vocational	No	Yes
10	Male	45	Married	Element Primary	No	No
11	Female	41	Divorce	High school	Yes	No
12	Female	37	Married	Vocational	No	Yes

DATA IN FREQUENCY DISTRIBUTIONS

Sex

Category	Frequency	Percentage
Female	5	41.5
Male	7	58.3

Marital status

Category	Frequency	Percentage
Divorce	2	16.7
Married	8	66.6
Single	2	16.7

Education

Category	Frequency	Percentage
Illiterate	3	25.0
Primary school	4	33.3
High school	2	16.7
Vocational	3	25.0

Coronary

Category	Frequency	Percentage
Yes	4	33,3
No	8	66.7

Normal distribution Not normal distribution

Age n=12 (Mean=38.1; SD=8.43) (Median=37.5; Min=25; Max-50)

SE=2.435403; 95%CI=32.7-43.4)

Category	Frequency	Percentage
<= 30 Yrs	3	25.0
31 - 45 yrs	6	50.0
>= 46 yrs	3	25.0