My reservoir is like the distance between Aden and Amman in al-Balqa

Al-Balqa Applied University

Faculty of Medicine

Epidemiology and Biostatistics

الوبائيات والإحصاء الحيوي (31505204)

Lecture 3

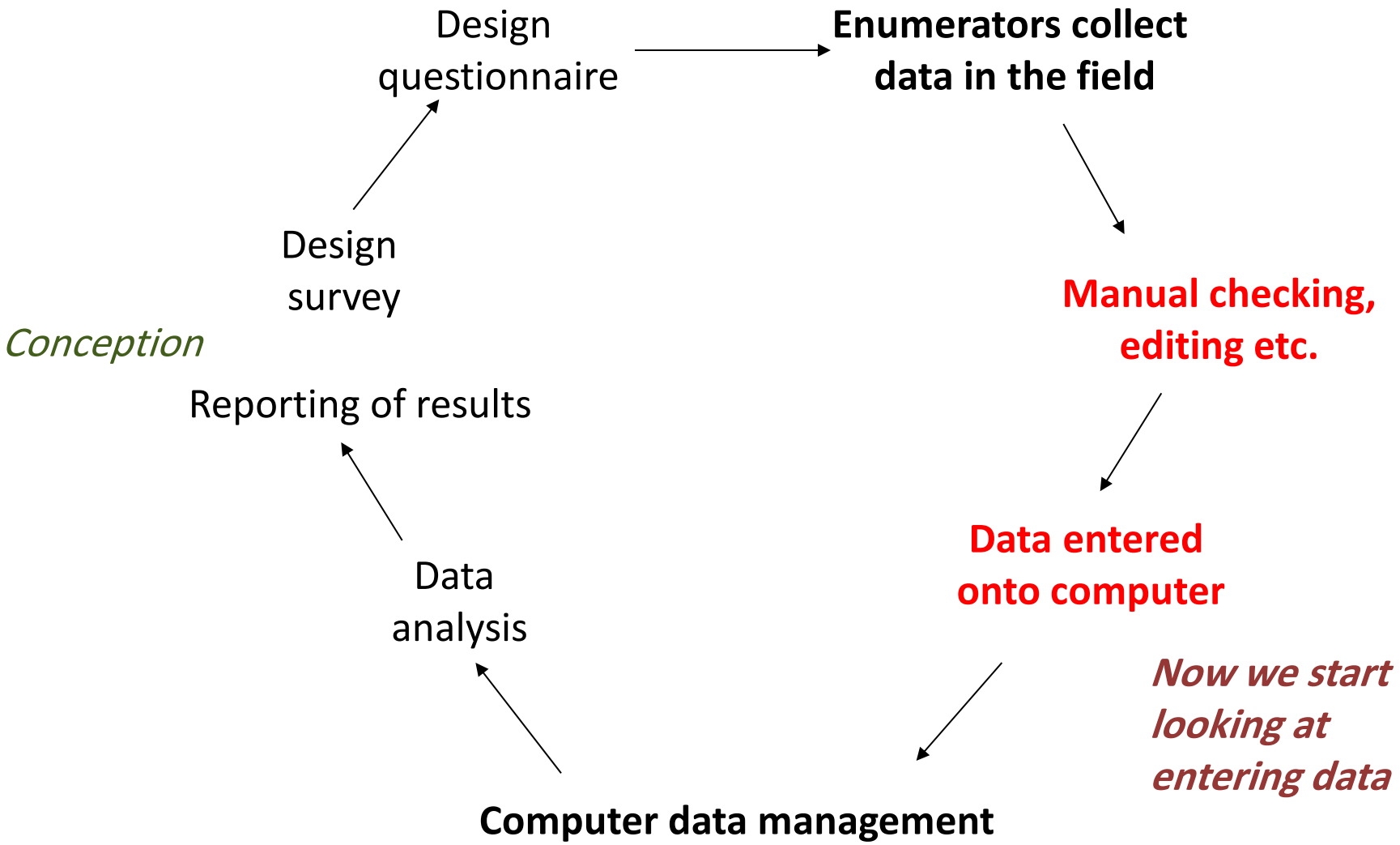
Descriptive statistics

Measures of variability

Graphical display: looking at data

23-6-2019

Data Management Cycle



Data Preparation

- **Data:** The simple concept of data is 0,1, it has no actual meaning (group of numbers or coding). This data when processed gives meaningful " information".

- In a research , the data is collected using questionnaire. Ask participants to fill it up, or you make direct observation and collect data.

- The questionnaire that you collect from consist of questions , and each question asks about something, e.g. age.

Data Preparation ……

- When you applied to the university , you were asked about your name, gender, nationality , etc. , each of them represents a variable.

- **Each question** has got **an answer**; **the answer** maybe :

1. Open ended : It is the most used on researches.
2. Coding : The options come in the form of :

   (A.xxx B. yyy C. zzz) or (1.Xxx 2. Yyy 3.Zzz).

- **Example**: What is your gender? (it is a **question** which means **a variable**) 1. Male 2. Female

  (*The answer is the coding for this question*).

Types of Data

- **The data subdivided into:**

1. **Quantitative Data (Numerical):** They are variables that **can be measured**, counted & have a numeric meaning such as ; age , weight , height.

2. **Qualitative Data (Categorical):** Information which **can not be expressed as a number**. It is something that you **cannot count** , assign a numeric value to it such as: gender , residency , nationality, etc.

Types of Variables

- **Quantitative data**: can be **discrete** taking only certain values or **continuous**, taking any value.

A. **Discrete variable (count data)** that have only certain fixed values and no intermediate values possible *(number of students in a class room*).

A. **Continuous variable (real-values)** where between any two points. There are at least theoretically infinite number of values (weight , height, etc.).

- **Example:** The number of times a patient is admitted to a hospital is **discrete** (*a patient cannot be admitted 0.8 times*), while a patient's weight is a **continuous** (a patient's weight could take any value within a range).

Types of Variables …

- **Qualitative data (Categorical)**: can be **nominal variable Or ordinal variables.**

A. **Nominal variable (not ordered)/ Name only:** The variables are divided into a number of named categories that **cannot be ordered** one above the other. It has no ordinary sense. No ordering of the categories (*The answer is determined*).

- **Example:** a patient's ethnicity , gender, eye color, names , marital status, blood groups, etc.

❖ **Binary variable**: A variable has two answer options.

- **Example:** yes or no questions, gender questions.

➢ This type of variable makes the data analysis easier.

Types of Variables …

B. Ordinal variable (ordered): The variables are divided into number of name categories that **can ordered from lowest to highest or vice versa.** Has sense of ordering. Categories can be ordered.

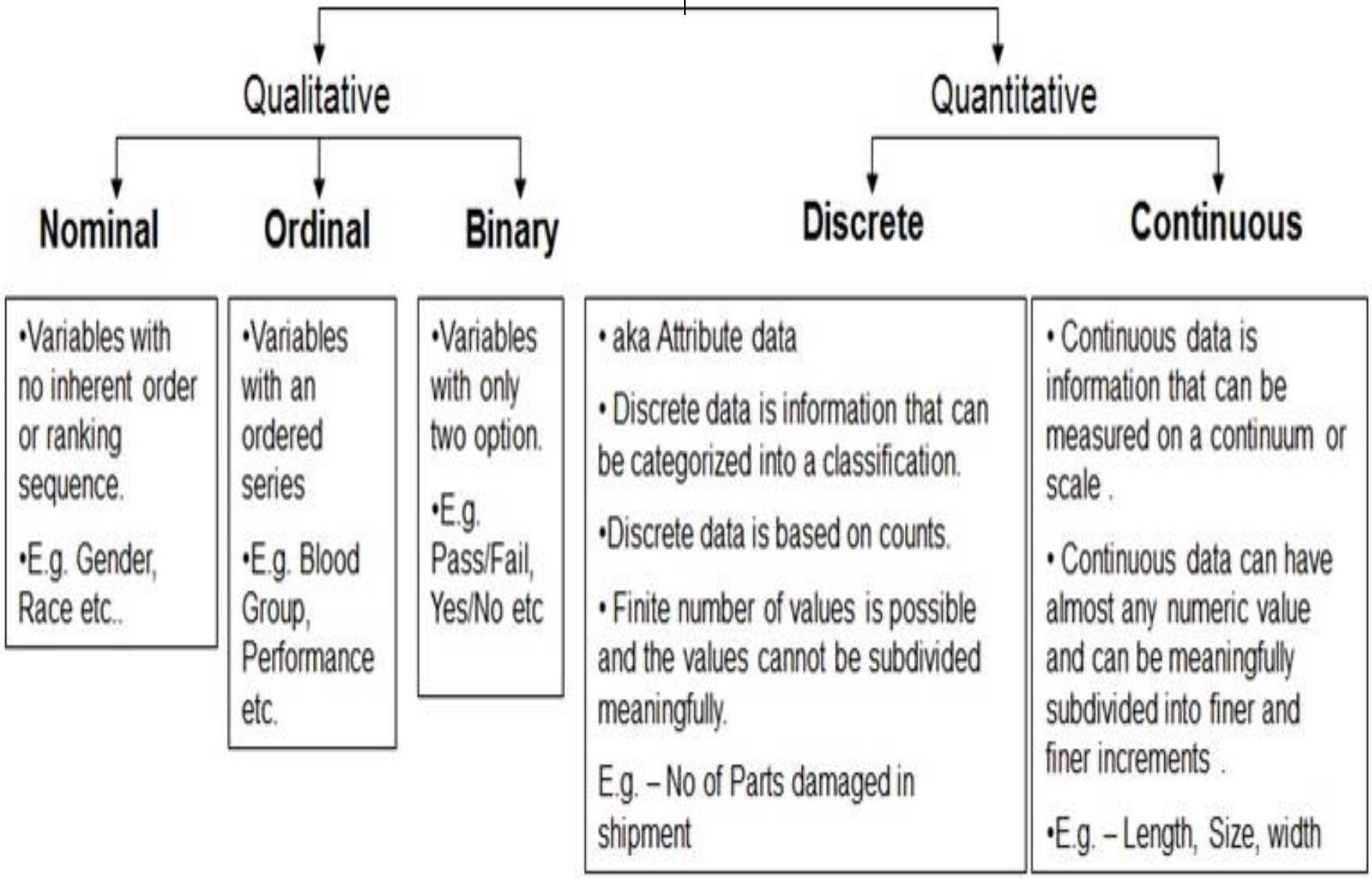
- **Example:** ***Response to treatment*** **,** Educational level (high school, university degree, college degree) can be organized in to an ascending or descending order.

Types of Variables …

- **One variable** could be quantitative or qualitative according to how it's presented.

- So, just saying blood pressure as a number means we classify it as **quantitative variable.** While saying types of blood pressure (explain high, normal, and low) then it's going to turn **a qualitative.**

Types of Variables

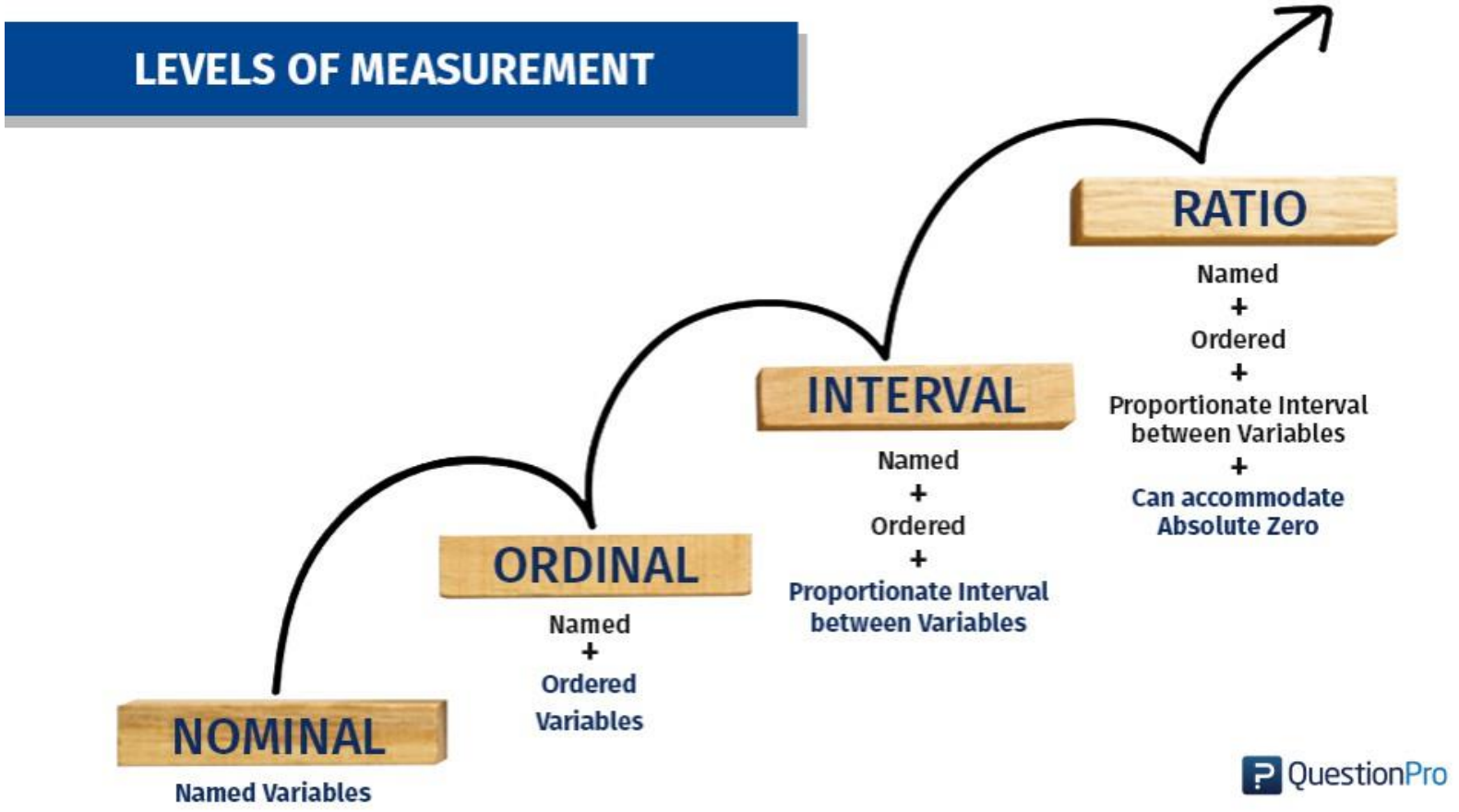


Types of Measurement Scales

1. **Nominal Scale (المقياس الأسمي / التنصيفي) ***
2. **Ordinal Scale) (المقياس الترتيبي **
3. **Interval Scale (مقياس الفتره) ***
4. **Ratio Scale (المقياس النسبي) ****

- The four scale types are ordered in that all later scales have all the properties of earlier scales—**plus** additional properties.

Level of Measurement



Nominal Scale

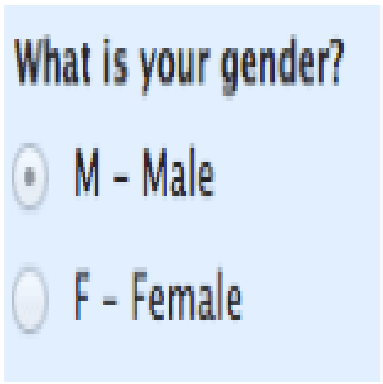
- Not really a ‘scale’ because it does not scale objects along any dimension.
- It simply labels objects.

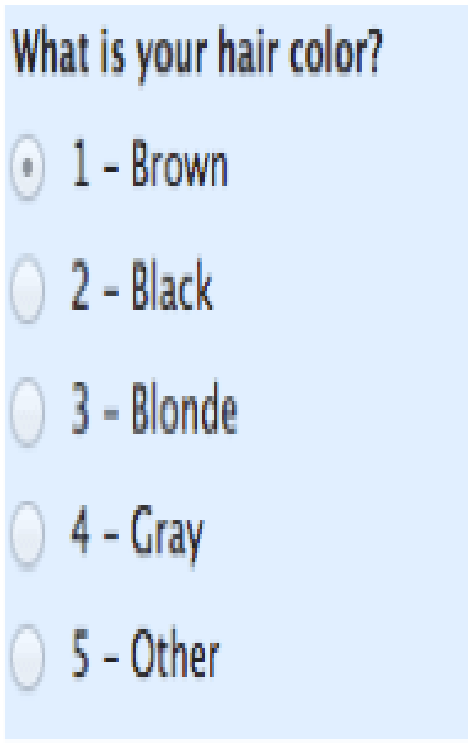
Example: Gender is a nominal scale

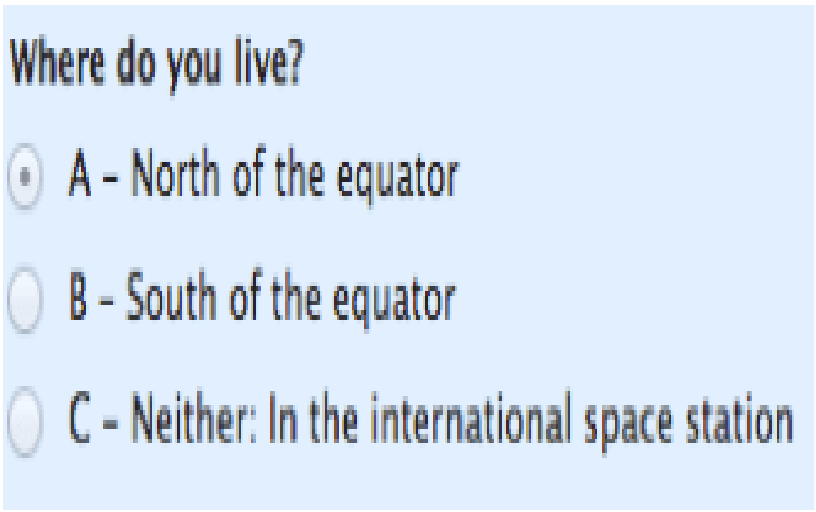
Male = 1

Female = 2

Nominal Scale….







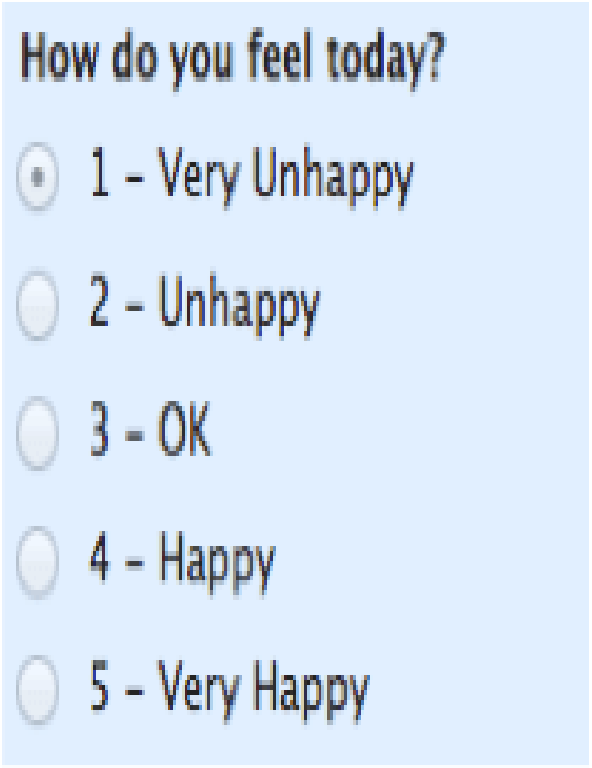
Ordinal Scale

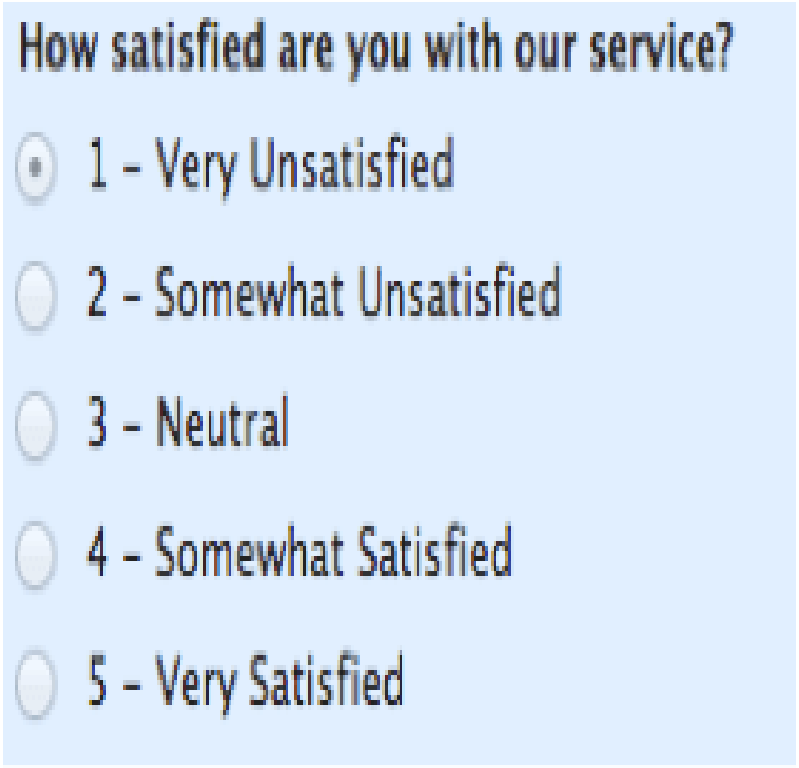
- **Ordinal Scale:** Nominal categories with implied order- Low, medium, high.

- Numbers are used to place objects in order.

- **But**, there is no information regarding the differences (intervals) between points on the scale.

Ordinal Scale ….

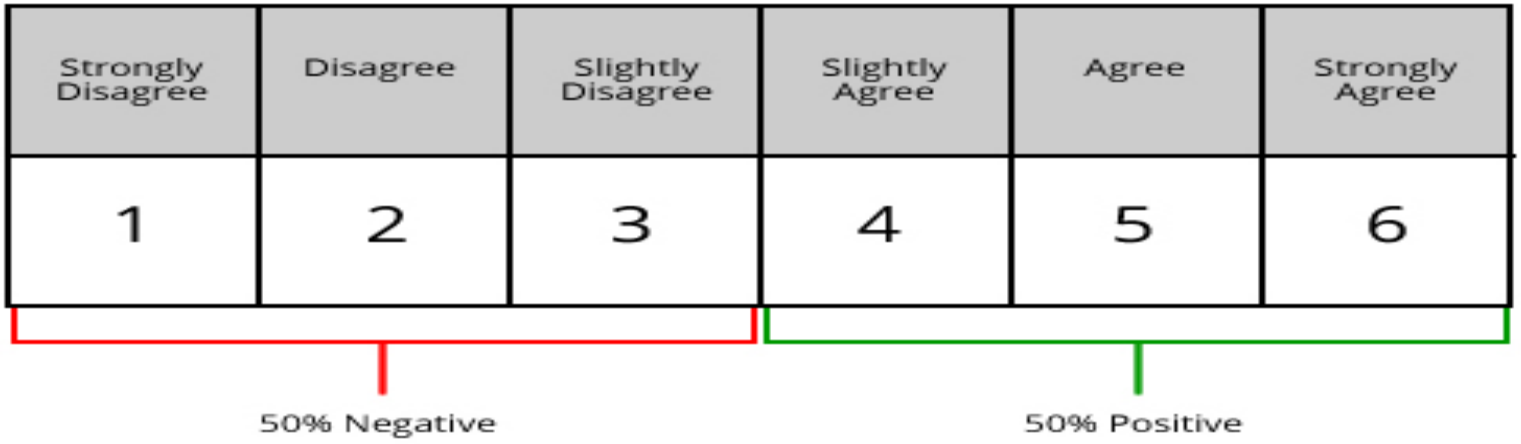




Likert Scale

Question: Compared to others, what is your satisfaction rating of the National Practitioner Data Bank?

1	2	3	4	5
Very Satisfied	Somewhat Satisfied	Neutral	Somewhat Dissatisfied	Very Dissatisfied



Interval Scale

- **Interval scale (**Numeric scales)**:** An interval scale is a scale on which equal intervals between objects, represent equal differences.
- The interval differences are meaningful. But, we can't defend ratio relationships.

- Differences *can* be compared; no true zero. **Ratios *cannot* be compared.**
- **Example: Temperature in Celsius.**

The difference between 10 and 20 degrees is the same as between 80 and 90 degrees but, we can't say that 80 degrees is twice as hot as 40 degrees.

Interval Scale….

- **Interval scales** are nice because the realm of statistical analysis on these data sets opens up. For example, *central tendency* can be measured by mode, median, or mean; standard deviation can also be calculated.

Ratio Scale

Ratio scale: Order and distance implied. Differences *can* be compared; has a true zero. **Ratios *can* be compared**.

Examples: Height, weight, blood pressure

- Ratios are meaningful.
- We can say that 20 seconds is twice as long as 10 seconds.

Summary of data-types and scales

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Two types of variables

1. **Dependent variables**: The variable that is used to describe or measure the problem under study. It is the **center of the study**.

2. **Independent variables**: The variable that are used to describe or measure the factors that are assumed to cause or at least to influence the problem.

- **Example:** If we are studying the blood pressure on a group of people take into consideration their age and environment, so the **center of study** is Hypertension, other variable are called **independent**.

- Whether a variable is dependent or independent is determined by the statement of the problem and study objectives.

Quantitative Analysis Software

- SAS (http://www.sas.com)
- SPSS (http://www-01.ibm.com/software/analytics/spss/)
- STATA (http://www.stata.com/)
- Microsoft Excel (!)
- Many others

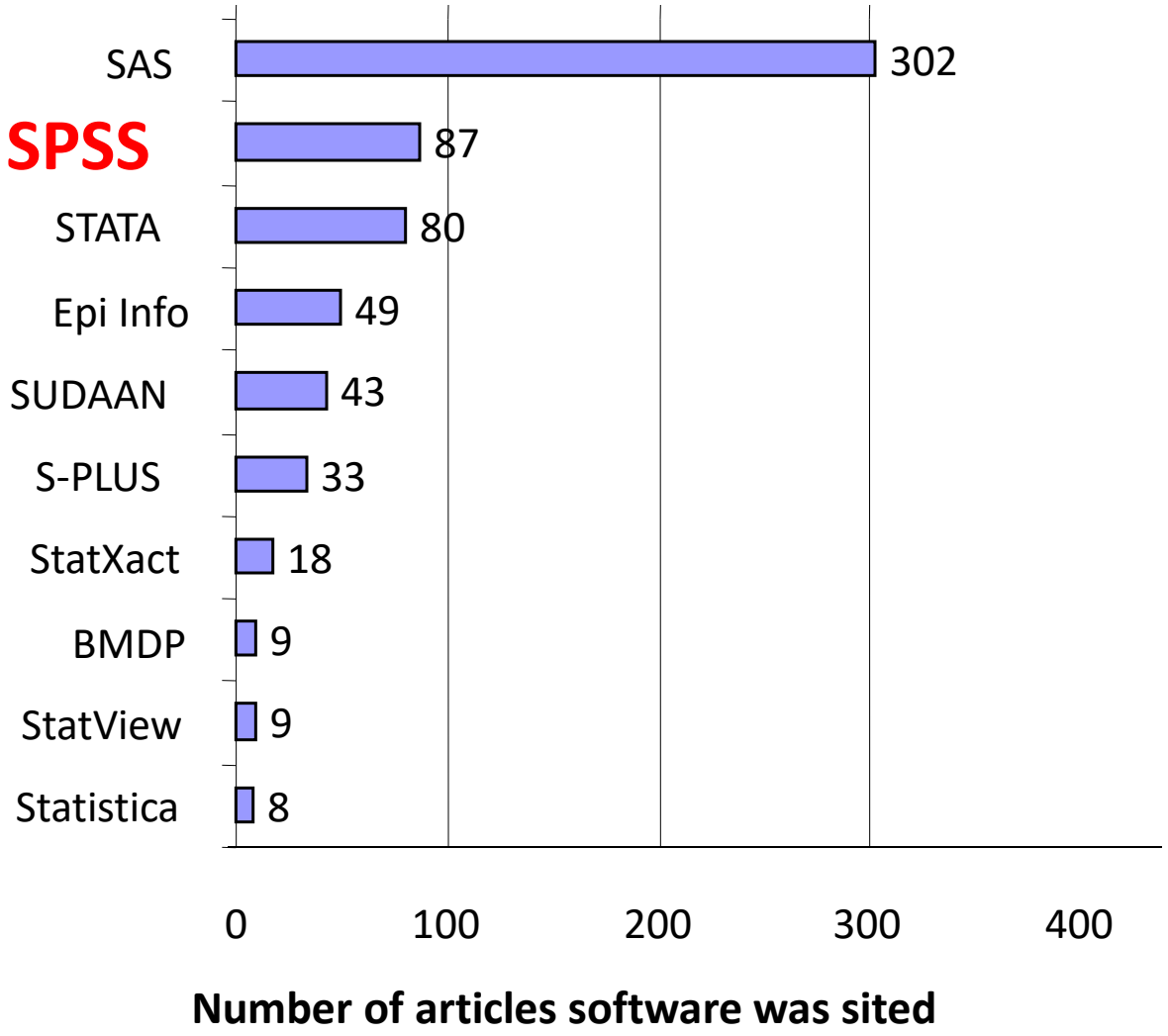
SPSS

Statistical Package for the Social Sciences

" الحزمة الاحصائية للعلوم الاجتماعية "

One of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions.

Statistical Software Packages Most Commonly Cited in the NEJM and JAMA between 1998 and 2002



SPSS interface

- **SPSS Windows has 3 windows:**

- **Data Editor :** Viewer or Draft Viewer which displays the output files.

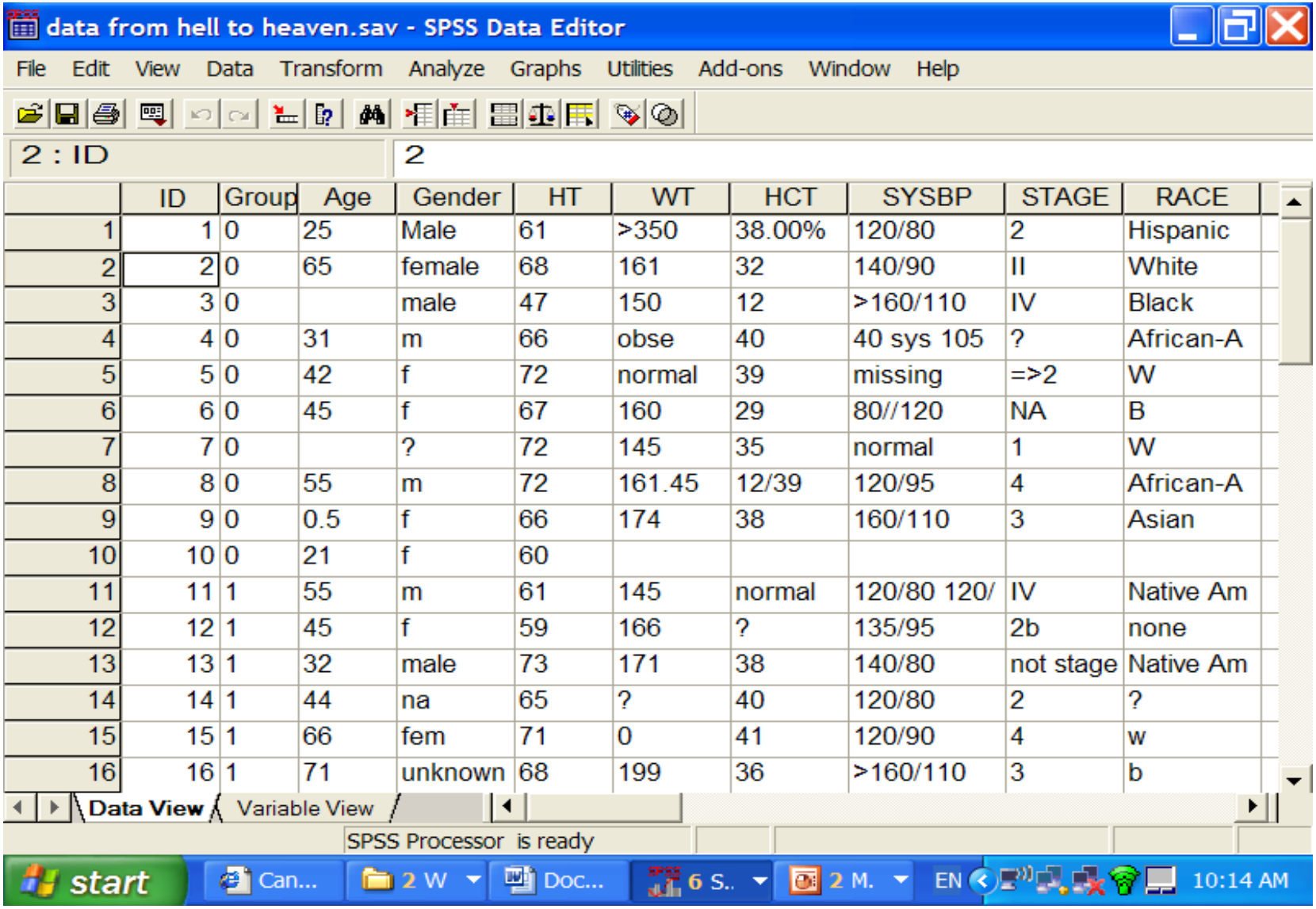
Syntax Editor, which displays syntax files.

- **The Data Editor has two parts:**

- **Data view window**, which displays data from the active file in spreadsheet format.
  - The place to enter data
  - **Columns**: variables
  - **Rows**: records

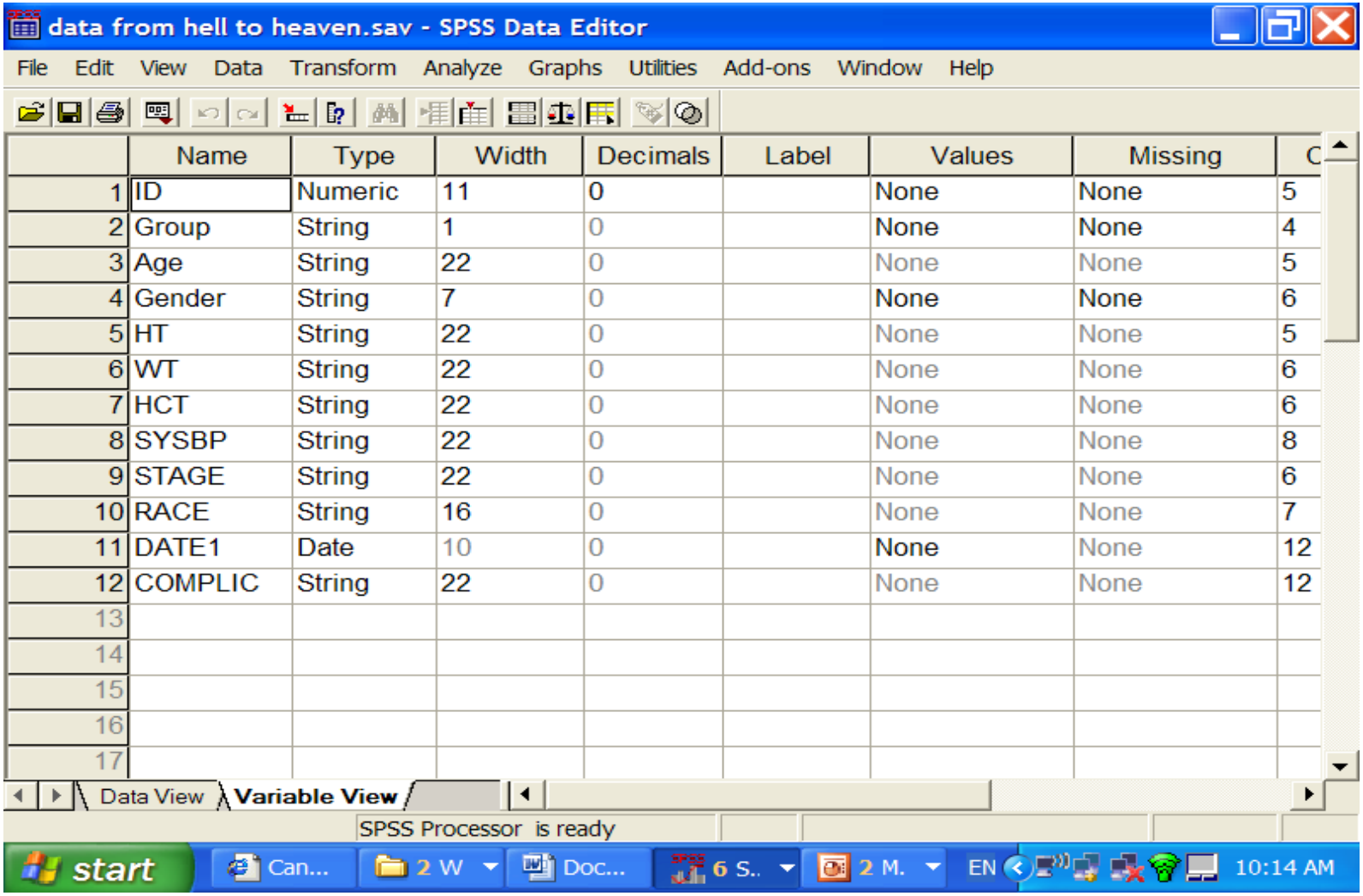
- **Variable View window**, which displays metadata or information about the data in the active file, such as variable names and labels, value labels, formats, and missing value indicators.
  - The place to enter variables
  - List of all variables
  - Characteristics of all variables

SPSS Data View



	ID	Group	Age	Gender	HT	WT	HCT	SYSBP	STAGE	RACE
1	1	0	25	Male	61	>350	38.00%	120/80	2	Hispanic
2	2	0	65	female	68	161	32	140/90	II	White
3	3	0		male	47	150	12	>160/110	IV	Black
4	4	0	31	m	66	obse	40	40 sys 105	?	African-A
5	5	0	42	f	72	normal	39	missing	=>2	W
6	6	0	45	f	67	160	29	80//120	NA	B
7	7	0		?	72	145	35	normal	1	W
8	8	0	55	m	72	161.45	12/39	120/95	4	African-A
9	9	0	0.5	f	66	174	38	160/110	3	Asian
10	10	0	21	f	60					
11	11	1	55	m	61	145	normal	120/80 120/	IV	Native Am
12	12	1	45	f	59	166	?	135/95	2b	none
13	13	1	32	male	73	171	38	140/80	not stage	Native Am
14	14	1	44	na	65	?	40	120/80	2	?
15	15	1	66	fem	71	0	41	120/90	4	w
16	16	1	71	unknown	68	199	36	>160/110	3	b

SPSS Variable View



	Name	Type	Width	Decimals	Label	Values	Missing	C
1	ID	Numeric	11	0		None	None	5
2	Group	String	1	0		None	None	4
3	Age	String	22	0		None	None	5
4	Gender	String	7	0		None	None	6
5	HT	String	22	0		None	None	5
6	WT	String	22	0		None	None	6
7	HCT	String	22	0		None	None	6
8	SYSBP	String	22	0		None	None	8
9	STAGE	String	22	0		None	None	6
10	RACE	String	16	0		None	None	7
11	DATE1	Date	10	0		None	None	12
12	COMPLIC	String	22	0		None	None	12
13								
14								
15								
16								
17								

Data Entry into SPSS

- **There are 2 ways to enter data into SPSS:**

1. Directly enter in to SPSS by typing in Data View.

2. Enter into other database software such as Excel then import into SPSS.

General guidelines for data entry

- Give each variable a valid name (8 characters or less with no spaces or punctuation, beginning with a letter not a numeric number). Short, easy to remember word names.

- Avoid the following variable names: *TEST, ALL, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH*. These are used in the SPSS syntax and if they were permitted, the software would **not be able** to distinguish between a command and a variable.

- Each variable name must be unique; duplication is not allowed. Variable names are not case sensitive. The names NEWVAR, NewVar, and newvar are all considered identical.

General guidelines for data entry……..

- Encode categorical variables. **Convert letters and words to numbers**.
- **Avoid mixing symbols with data**. Convert them to numbers.
- Give each patient a unique, sequential case number (ID). Place this ID number in the first column on the left.
- **Do not** make columns **wider then 8 characters**, unless absolutely essential.

General guidelines for data entry…….

- **Each variable** should be in its own column.

Avoid this:

Animal
Control1
Control2
Experiment1
Experiment2

Change to:

Animal	Group
1	0
2	0
3	1
4	1

❖ **Do not combine variables in one column**.

❖ It is recommended to use 0/1 for 2 groups with 0 as a reference group.

General guidelines for data entry……..

- All data for a project should be in **one spreadsheet**. Do not include graphs or summary statistics in the spreadsheet.

- Each patient should be entered **on a single line or row**. Do not copy a patient's information to another row to perform subgroup analysis.

- **Put ordinal variables into one column** if they are mutually exclusive

Avoid:

Pain		
Mild	Moderate	Severe
1	0	0
0	1	0
0	0	1

Preferred:

Pain
1
2
3

General guidelines for data entry……..

- **For yes/no questions, enter** "0" for no and "1" for yes. Do not leave blanks for no. Do not enter "?", "*", or "NA" for missing data because this indicates to the statistical program than the variable is a **string variable**.

- String variables cannot be used for any arithmetic computation.

General guidelines for data entry……..

- However when data are repeatedly collected over a patient, it's recommended to have patient-day observation on a simple line to ease data management.
- SPSS has a nice feature to convert from the longitudinal format to horizontal format. When the number of repeats are few 2 or 3, horizontal format may be preferred for simplicity.

Longitudinal data entry

Date	ID	SYSBP
1/2/2005	1	130
1/3/2005	1	120
1/4/2005	1	120
3/1/2005	2	110
3/2/2005	2	140

Horizontal data entry

ID	SYSBP1	SYSBP2	SYSBP3
1	130	120	120
2	110	140	

Broad Categories of Statistics

- Statistics can broadly be split into two categories ***Descriptive Statistics and Inferential Statistics:***

1. **Descriptive statistics** deals with the meaningful presentation of data such that its characteristics can be effectively observed.

2. **Inferential statistics** on other hand, deals with drawing inferences and taking decision by studying a subset or sample from the population.

Descriptive Biostatistics

- The best way to work with data is to **summarize and organize** them.

- Numbers that have **not been summarized and organized are called raw data.**

Definition

- **Data is any type of information.**
- **Raw data** is a data collected as they receive.
- **Organize data** is data organized either in ascending, descending or in a grouped data.

Descriptive Statistics

1. **Frequency Distribution.**
2. **Measure of Central Tendency.**
3. **Measure of Dispersion.**