Lec: 1

Biostatistics

What is biostatistics?

***Biostatistics*: ** is the application of statistical principles to questions and problems in medicine, public health, or biology.

What studying biostatistics is useful for?

• Design and analysis of research studies.

•Describe and summarize the data we have.

•Analyze data to measure the association or difference.

• To conclude if an observation is of real significance or just due to chance.

• To understand and evaluate published scientific research papers.

The statistical analysis journey:

The statistical analysis journey goes through the following steps:

• Transforming the research idea into a research question.

•Choosing the proper study design and selecting a suitable sample.

• Performing the study and collecting data.

• Analyzing data (using the appropriate test).

• Getting and interpreting the p-value.

• Reaching a conclusion (answer) regarding the research question.

The statistical analysis journey:

The statistical analysis journey goes through the following steps:

• Transforming the research idea into a research question.

•Choosing the proper study design and selecting a suitable sample.

• Performing the study and collecting data.

• Analyzing data (using the appropriate test).

• Getting and interpreting the p-value.

• Reaching a conclusion (answer) regarding the research question.

Data and statistics

The purpose of most studies is to collect **data** to obtain information about a particular area of research. Our data comprise **observations** on one or more variables; any quantity that varies is termed a **variable**.

For example, we may collect basic clinical and demographic information on patients with a particular illness.

The variables of interest may include the sex, age and height of the patients.

Types of data variables

A **data variable** is "something that varies" or differs from person to person or group to group.

Data variables are the items that we collect data about.

Examples for data variables are sex, age, weight, marital status, satisfaction rate, etc.

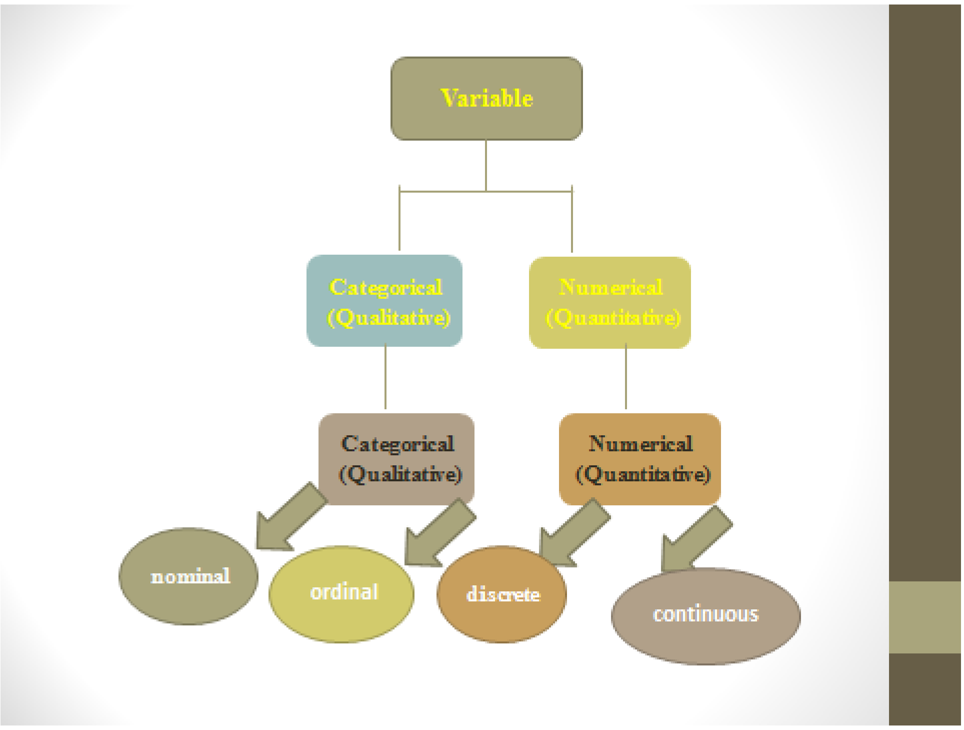
When dealing with data, it is important to recognize the type of each data variable for the following reasons:

- **Summarizing data: describing a variable in mean with standard deviation or in frequency with percentage depends on the type of data variable.**
- **Graphical presentation:** choosing the proper graph to represent the data depends on the type of data variable.
- **Analyzing data:** choosing the suitable statistical tests also depends on the type of data variables.

Data variables are classified generally into the following 2 types:

A. Categorical variables: which are either nominal or ordinal?

B. Numerical Variables: which are either discrete or continuous?



A. Categorical variables:

They are also known as qualitative or nominal data; they have NO unit of measurement.

Individuals are described as belonging to any of the categories of this variable.

Examples:

1. Satisfaction status: (satisfied, neutral, not satisfied }

2. Sex: (female, male)

3.Colors: (red, green, blue, pink)

3. Nationality: (all countries)

Categorical variable are:

- ❑ **Nominal data**

- ❑ ***Ordinal data***

- ❑ ***1. Nominal variables*****:** those are categorical variables that have no intrinsic order.

- ❑ **Examples:**

- ❑ Sex: (female, male), can also be presented as (male, female)

- ❑ Blood groups: (A, B, AB, O) can also be presented as (A, B, O, AB) or any other order.

- ❑ Nationality: can be presented in any way; there is no order for the countries.

If the nominal variable has only two groups as

Sex (male, female),

An answer to a question (Yes, No),

Or a disease status (diseased, not diseased),

We call it a **dichotomous** variable, or a **binomial** variable.

Sometimes, categorical variables are coded in numbers like:

1 for females and 2 for males, or 0 for No, and 1 for yes, and so on.

Even if they are coded or represented as numbers, they are still categories, and the data type is categorical.

The number here is just a code.

2- ***Ordinal variables*****:** those are categorical variables that have an order, and that order has a meaning.

Examples:

BMI status: (underweight, normal, overweight, obese, extremely obese)

Agreement level:

(Strongly disagree, disagree, undecided, agree, and strongly agree)

Even if this variable is coded in numbers from 1 to 5, it is still an ordinal variable that is categorical and not numerical.

B. Numerical variables:

Those variables are either measured or counted, represented in numbers, and have a measurement unit.

Examples:

- Height (in cm)

- Weight (in kg)

- Blood glucose level (in mg/dL)

- Number of kids in the family (4 kids, 2 kids, one kid, etc.)

Numerical variables are either discrete or continuous.

1- Discrete variables:

They take only integer numbers (no decimals) such as 0,1,2,3,4…

They usually represent a count of something.

Examples:

- Number of kids in a family.

- Number of stents inserted into the coronaries.

- Number of patient visits to the hospital.

The unit of measurement represents what we are counting (as kid, stent, visit, respectively)

2. Continuous variables:

They can take any real numerical value, including decimals (as 14.55, 48.8, 178.2).

They involve measurement and have measurement units.

Examples:

- Weight (in kg)

- Height (in cm)

- Blood glucose level (in mg/dL)

How to differentiate between types of data variables:

Step 1: Is there a unit of measurement?

If No, it is categorical, and if Yes, it is numerical.

Step 2:

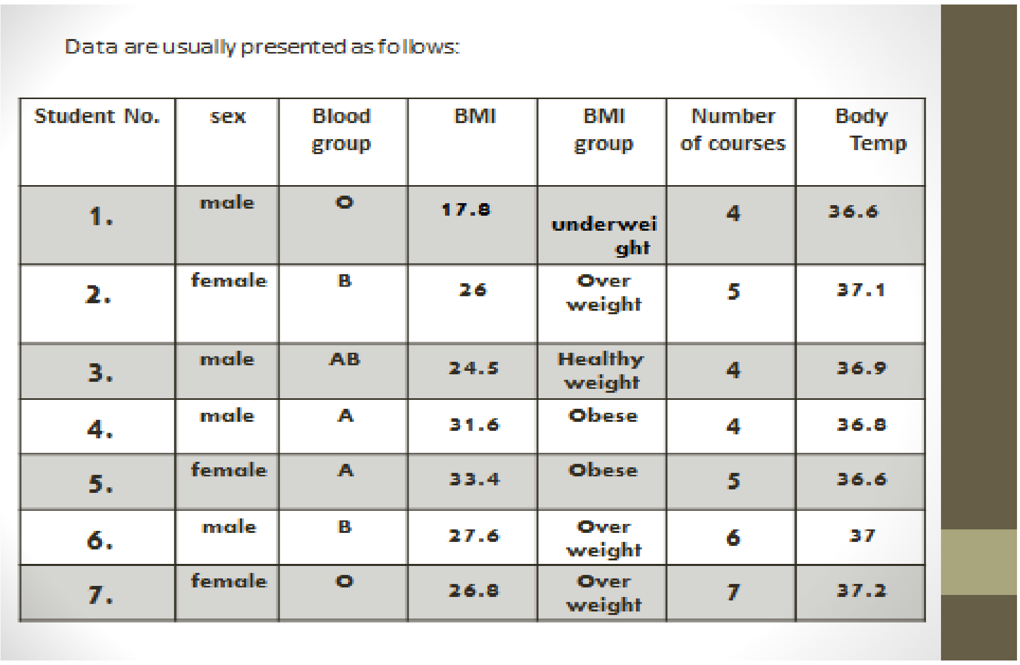
For the categorical variables: Is there an order?

If No, it is nominal, and if Yes, it is ordinal.

For the numerical variables: Is it counted or measured?

If counted, it is discrete, and if measured, it is continuous.

Data are usually presented as follows:



Data are usually presented as follows:

Student No.	sex	Blood group	BMI	BMI group	Number of courses	Body Temp
1.	male	O	17.8	underweight	4	36.6
2.	female	B	26	Over weight	5	37.1
3.	male	AB	24.5	Healthy weight	4	36.9
4.	male	A	31.6	Obese	4	36.8
5.	female	A	33.4	Obese	5	36.6
6.	male	B	27.6	Over weight	6	37
7.	female	O	26.8	Over weight	7	37.2

Types of data variables in this dataset are:

o Sex: nominal (dichotomous), categorical

o Blood group: nominal, categorical

o BMI: continuous, numerical

o BMI group: ordinal, categorical

o Number of courses: discrete, numerical

o Body temperature: continuous, numerical

Some more ideas:

• Some textbooks classify numerical data into interval variables and ratio variables.

Ratio variables : are variables that have true zero, such as weight. When we say the weight is zero, this means the complete absence of weight, and a weight of 30 kgs is twice as heavy as 15 kgs.

While in **interval variables** as temperature, there is no true zero. A temperature of 00C does not mean the absence of heat, and a temperature of 300C is not twice as hot as 150C.

• When data is ordinal in nature with a large number of levels as a pain score measured on a 10 levels scale, it can be treated as a discrete variable.

• Some variables that are continuous in nature are sometimes measured as discrete; age is an example as it is usually reported as the number of years instead of the exact age.

Levels of data measurement:

It is possible to change the type of data variable into another one, but only in one direction:

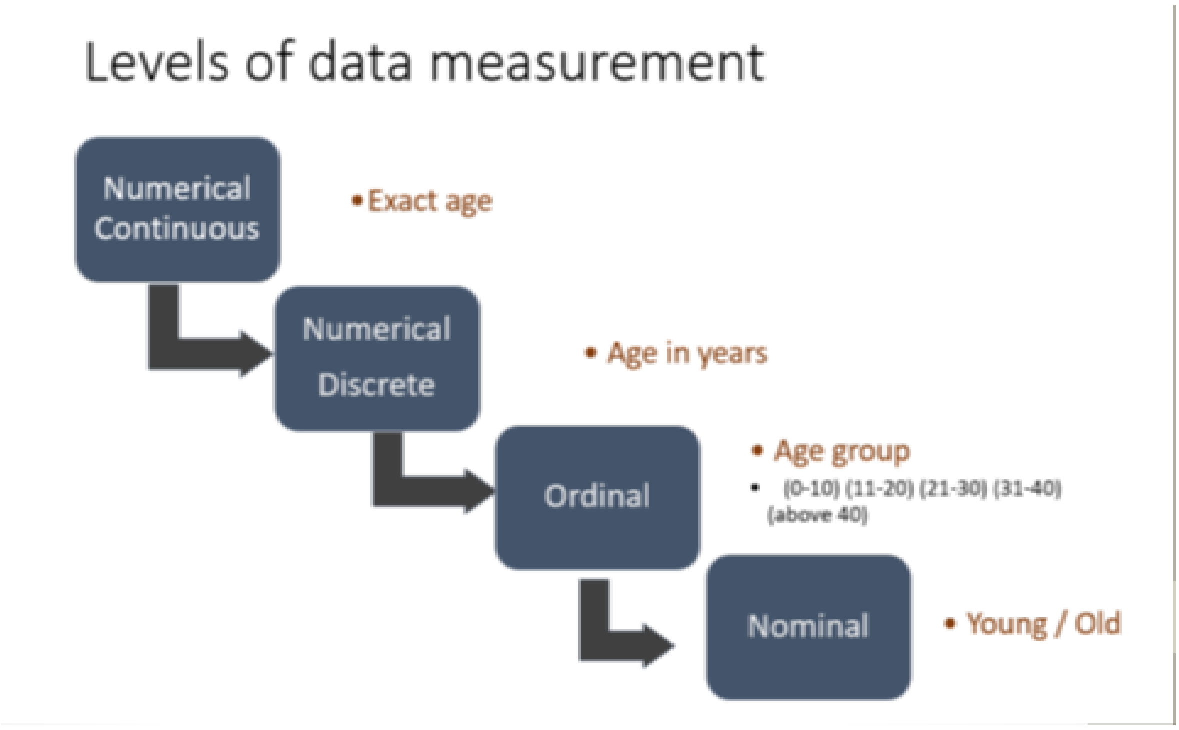
Numerical continuous → numerical discrete → ordinal → nominal

• We can change the age from a numerical variable to an ordinal variable if we categorize it into different age groups.

• Also, we can change the age from an ordinal variable as age groups into a nominal variable of two levels (young, old).

• But if we collect the data in a categorical form, we cannot transform it into a numerical form.

O Whenever possible, collect your data at the highest level, numerical continuous or numerical discrete, as it is more accurate and can be categorized easily later on.



Levels of data measurement

Numerical Continuous
- Exact age

Numerical Discrete
- Age in years

Ordinal
- Age group
- (0-10) (11-20) (21-30) (31-40) (above 40)

Nominal
- Young / Old