ISCB20.05–Introduction to Biostatistics

Md. Jubayer Hossain
https://jhossain.me/

December 25, 2020

Lead Organizer, Introduction to Scientific Computing for Biologists
Founder, Health Data Research Organization

HEALTH DATA
RESEARCH ORGANIZATION

Section–1.1: Introduction to Biostatistics

Statistics and Biostatistics: Definitions–1

Definition–1: What is Statistics?

Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied. (Source: `https://en.wikipedia.org/wiki/Statistics`)

Definition–2: What is Statistics?

Simply, Statistics is a branch of mathematics that deals with collecting, organizing, analyzing, and interpreting data. (Source: `https://app.pluralsight.com/library/courses/interpreting-data-descriptive-statistics-python/`)

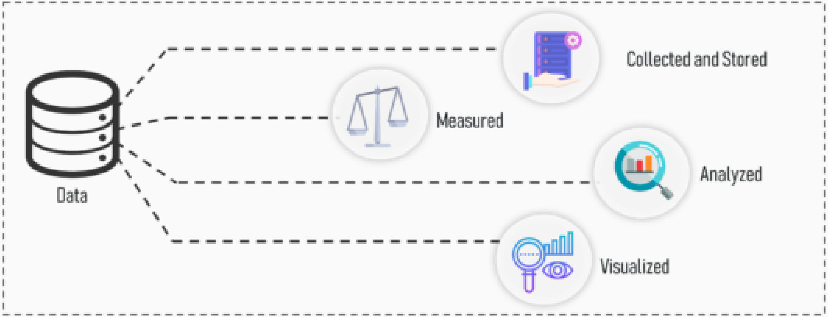
Statistics and Biostatistics: Definitions–2

Definition–3: What is Biostatistics?
Biostatistics is the application of statistics to a variety of topics in biology. In this course, we tend to focus on biological topics in the health sciences as we learn about statistics. (Source: `https://bolt.mph.ufl.edu/6050-6052/`)

Definition–4: What is Biostatistics?
Biostatistics is the application of statistics to problems in the biological sciences, health, and medicine. (Source: `https://ocw.jhsph.edu/index.cfm/go/viewCourse/course/MethodsInBiostatisticsI/`)

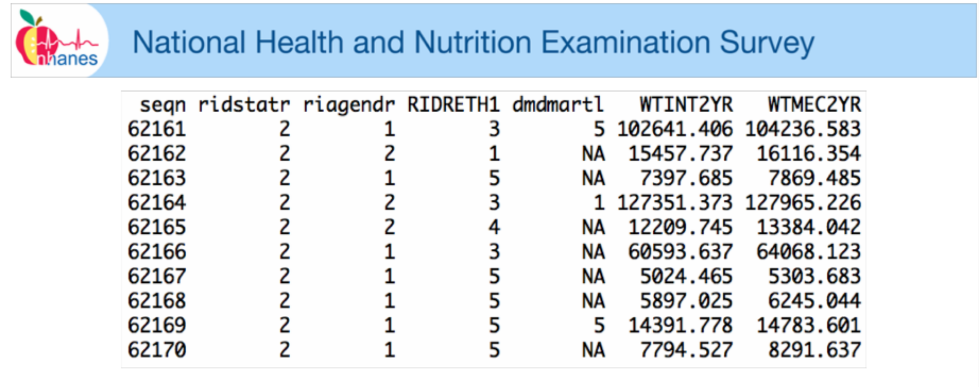
What is Data?

- **Definition–1:** Data is a collection of facts, such as numbers, images, words, measurements, observations, audios, videos or just descriptions of things.
- **Definition–2:** Data is a tool to reach suitable conclusion.



Source: `https://www.edureka.co/blog/statistics-and-probability/`

Data can be Numbers

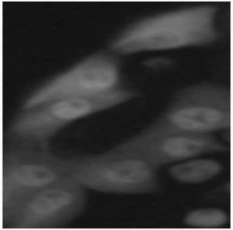


National Health and Nutrition Examination Survey

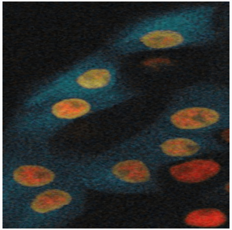
```
 seqn ridstatr riagendr RIDRETH1 dmdmartl   WTINT2YR   WTMEC2YR
62161        2        1        3        5 102641.406 104236.583
62162        2        2        1       NA  15457.737  16116.354
62163        2        1        5       NA   7397.685   7869.485
62164        2        2        3        1 127351.373 127965.226
62165        2        2        4       NA  12209.745  13384.042
62166        2        1        3       NA  60593.637  64068.123
62167        2        1        5       NA   5024.465   5303.683
62168        2        1        5       NA   5897.025   6245.044
62169        2        1        5        5  14391.778  14783.601
62170        2        1        5       NA   7794.527   8291.637
```

Source: `https://www.cdc.gov/nchs/nhanes/index.htm`

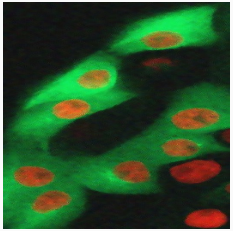
Data can be Images



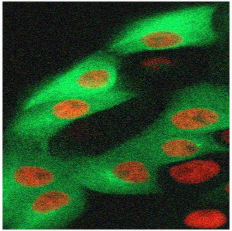
(Filtered Image)



(Filtered Image)



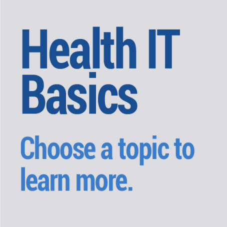
(Noisy Image)



(Noisy Image)

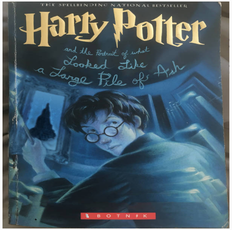
Source: `https://www.cdc.gov/nchs/nhanes/index.htm`

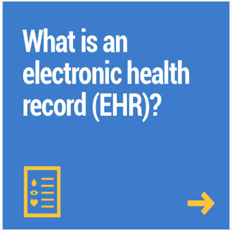
Data can be Words



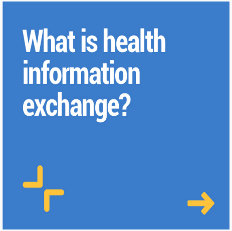
Health IT Basics

Choose a topic to learn more.





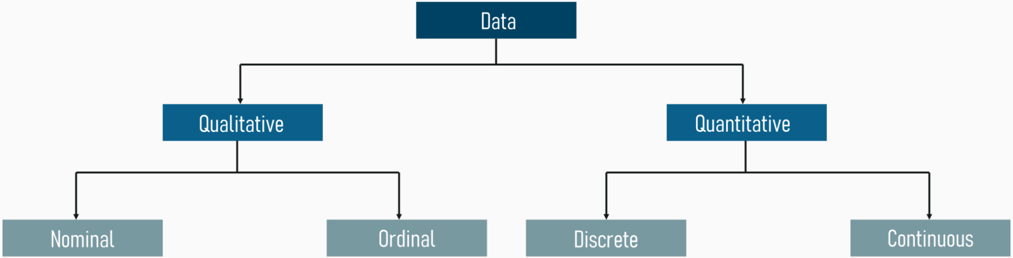
What is an electronic health record (EHR)?



What is health information exchange?

Source: `https://www.healthit.gov/topic/health-it-and-health-information-exchange-basics/health-it-and-health-information-exchange`

Types of Data



Data

Qualitative

Quantitative

Nominal

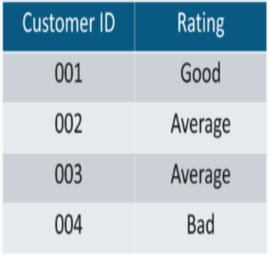
Ordinal

Discrete

Continuous

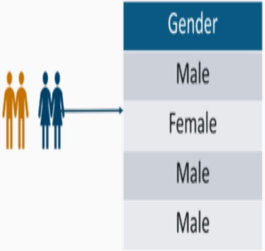
Qualitative or Categorical Data

- Classifies individuals or items into different groups.
- Qualitative data is further divided into two types of data
  - **Ordinal:** groups have an order or ranking.
  - **Nominal:** groups are merely names, no ranking.



Customer ID	Rating
001	Good
002	Average
003	Average
004	Bad

(a) Ordinal Data



Gender
Male
Female
Male
Male

(b) Nominal Data

Source: `https://www.edureka.co/blog/statistics-and-probability/`

Quantitative or Numeric Data

- Numerical, measurable quantities in which arithmetic operations often make sense.
- Quantitative data is also further divided into two types of data
  - **Continuous:** could take on any value within an interval,many possible values.
    - A person's height: could be any value (within the range of human heights), not just certain fixed heights.
    - Time in a race: you could even measure it to fractions of a second.
    - Blood pressure, mmHg.
    - Weight, pounds (kilograms, ounces, etc.)
  - **Discrete:** countable value, finite number of values.
    - The number of students in a class.
    - The results of rolling a die.

Binary Data

- Yes/No
- Polio: Yes/No
- Cure: Yes/No
- Sex: Female/Male(0 or 1)

Types of Variables

- **Independent Variable(IV)**
  A variable whose value does not change by the effect of other variables and is used to manipulate the dependent variables. It is often denoted as X.
- **Dependent Variable(DV)** A variable whose value change when there is any manipulation in the values of independent variables. Is is often denoted as Y

X **Causes** Y

X (effect) → Year of Experience → Independent

Y (cause) → Salary → Dependent

Other Names for IV and DV

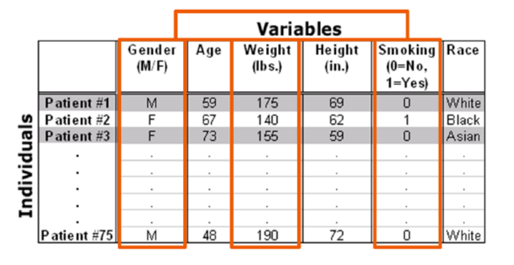
Other Names for Independent Variables

- Explanatory Variables (they explain an event or outcome)
- Predictor Variables (they can be used to predict the value of a dependent variable)

Other Names for Dependent Variables

- Response Variables (they respond to a change in another variable)
- Outcome Variables (they represent the outcome you want to measure)

A Typical Dataset


Variables
Individuals

	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (0=No, 1=Yes)	Race
Patient #1	M	59	175	69	0	White
Patient #2	F	67	140	62	1	Black
Patient #3	F	73	155	59	0	Asian
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Patient #75	M	48	190	72	0	White

- **Variables** contain the information about a particular characteristic for all individuals in a dataset.
- An **observation** in statistics is a value of something of interest you're measuring or counting during a study or experiment: a person's height, a bank account value at a certain point in time, or number of animals.

Sources of Data

Primary Sources of Data

- Collection of data from source of origin.
- Conducting interviews, experimentation.
- Provide first hand information.

Secondary Sources of Data

- Collection of data from agency which already has collected data and processed it.
- Conducting interviews, experimentation

Pros and Cons of Primary Data

Pros

- Can be collected to answer your specific research question.
- You have control over the sampling and measurement methods.

Cons

- More expensive and time-consuming to collect.
- Requires training in data collection methods.

Pros and Cons of Secondary Data

Pros

- Easier and faster to access.
- You can collect data that spans longer timescales and broader geographical locations.

Cons

- No control over how data was generated.
- Requires extra processing to make sure it works for your analysis.

Data Collection Methods

- Interviews
- Questionnaires and surveys
- Observations
- Focus groups
- Oral histories

See More `https://www.jotform.com/data-collection-methods/`

Terms used in Data Collection

- **Variable –** values which changes
- **Observations –** values from variables are referred as observations.
- **Statistical Investigation –** search for information conducted by using statistical methods.
- **text –** who conducts statistical inquiry.
- ITEM 5

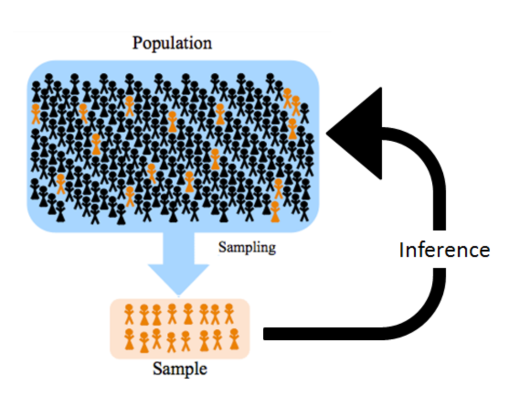
Important Agencies for Secondary Data

- https://dghs.gov.bd/index.php/en/data
- https://framinghamheartstudy.org/
- https://www.data.gov/
- https://healthdata.gov/
- https://www.who.int/ictrp/network/trds/en/
- https://data.cdc.gov/
- https://www.kaggle.com/

Terminologies In Statistics: Population and Sample

Population: The population is the entire group that you want to draw conclusions about.

Sample: The sample is the specific group of individuals that you will collect data from.



Source: `https://online.stat.psu.edu/stat500/`

Terminologies In Statistics: Sampling Frame and Sample Size

Sampling Frame

The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).

Sample Size

The number of individuals in your sample depends on the size of the population, and on how precisely you want the results to represent the population as a whole.

Sample Size Calculator

Surveymonkey–`https://www.surveymonkey.com/mp/sample-size-calculator/`

Characteristics of a Good Sample–1

- **Goal-oriented:** A sample should be goal oriented. It should be oriented to the research objectives and fitted to the survey conditions.

Characteristics of a Good Sample–1

- **Goal-oriented:** A sample should be goal oriented. It should be oriented to the research objectives and fitted to the survey conditions.
- **Acurate representative of the population:** A sample should be an accurate representative of the population from which it is taken.

Characteristics of a Good Sample–1

- **Goal-oriented:** A sample should be goal oriented. It should be oriented to the research objectives and fitted to the survey conditions.
- **Acurate representative of the population:** A sample should be an accurate representative of the population from which it is taken.
- **Proportional:** A sample should be proportional. It should be large enough to represent the population properly. In general, the larger the sample size, the more accurately and confidently you can make inferences about the whole population.

Characteristics of a Good Sample–1

- **Goal-oriented:** A sample should be goal oriented. It should be oriented to the research objectives and fitted to the survey conditions.
- **Acurate representative of the population:** A sample should be an accurate representative of the population from which it is taken.
- **Proportional:** A sample should be proportional. It should be large enough to represent the population properly. In general, the larger the sample size, the more accurately and confidently you can make inferences about the whole population.
- **Random Selection:** A sample should be selected at random. This means that any item in the group has a full and equal chance of being selected and included in the sample.This makes the selected sample truly representative in character.

Characteristics of a Good Sample–1

- **Goal-oriented:** A sample should be goal oriented. It should be oriented to the research objectives and fitted to the survey conditions.
- **Acurate representative of the population:** A sample should be an accurate representative of the population from which it is taken.
- **Proportional:** A sample should be proportional. It should be large enough to represent the population properly. In general, the larger the sample size, the more accurately and confidently you can make inferences about the whole population.
- **Random Selection:** A sample should be selected at random. This means that any item in the group has a full and equal chance of being selected and included in the sample.This makes the selected sample truly representative in character.
- **Economical:** A sample should be economical.The objective of the survey should be achieved with minimum cost and effort.

Characteristics of a Good Sample–2

- **Practical:** A sample should be practical. The sample design should be simple. It should be capable of being understood and followed in the fieldwork.

Characteristics of a Good Sample–2

- **Practical:** A sample should be practical. The sample design should be simple. It should be capable of being understood and followed in the fieldwork.
- **Actual information provider:** A sample should be designed so as to provide actual information required for the study and also provide an adequate basis for the measurement of its own reliability.

Characteristics of a Good Sample–2

- **Practical:** A sample should be practical. The sample design should be simple. It should be capable of being understood and followed in the fieldwork.
- **Actual information provider:** A sample should be designed so as to provide actual information required for the study and also provide an adequate basis for the measurement of its own reliability.

Types of Statistics

“There are two kinds of statistics, the kind you look up and the kind you make up”
–Rex Stout

- **Descriptive Statistics –** Identify important elements in a dataset.
- **Inferential Statistics –** Explain those elements via relationships with other elements.

Descriptive Statistics

Descriptive statistical methods provide an exploratory assessment of the data from a study.

- Descriptive statistical methods provide a exploratory data analysis.
  - Frequency Distribution Table
  - Graphs / Charts
  - Summary
- Descriptive statistical methods divide into 3 categories.
  - **Univariate analysis** summarize only one variable at a time.
  - **Bivariate analysis** compare two variables.
  - **Multivariate analysis** compare more than two variables.

Inferential Statistics

Assess the strength of evidence for/against a hypothesis; evaluate the data

- Inferential statistical methods provide a confirmatory data analysis
  - Generalize conclusions from data from part of a group (sample) to the whole group (population)
  - Assess the strength of the evidence
  - Make comparisons
  - Make predictions
- Inferential statistical methods divide into 2 categories.
  - **Hypothesis Testing:** Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.
  - **Model Fitting:** Model fitting is a measure of how well a statistical learning model generalizes to similar data to that on which it was trained. A model that is well-fitted produces more accurate outcomes.

References

- https://bolt.mph.ufl.edu/6050-6052/
- https://online.stat.psu.edu/stat500/
- https://online.stat.psu.edu/stat100/
- https://online.stat.psu.edu/stat200/

Thank You

☺