BIOSTATISTICS

Mahmoud Al Hussami, PhD., DSc.

Associate Professor of Epidemiology

Unit One

INTRODUCTION

Biostatistics

- **It** can be defined as the application of the mathematical tools used in statistics to the fields of biological sciences and medicine.
- It is a growing field with applications in many areas of biology including epidemiology, medical sciences, health sciences, educational research and environmental sciences.

Concerns of Biostatistics

- Biostatistics is concerned with collection, organization, summarization, and analysis of data.
- We seek to draw inferences about a body of data when only a part of the data is observed.

Purposes of Statistics

- To describe and summarize information thereby reducing it to smaller, more meaningful sets of data.
- To make predictions or to generalize about occurrences based on observations.
- To identify associations, relationships or differences between the sets of observations.

Data

- Data are numbers which can be measurements or can be obtained by counting.
- Biostatistics is concerned with the interpretation of the data and the communication of information about the data.

Sources of data

Data are obtained from

- Analysis of records
- Surveys
- Counting
- Experiments
- Reports

Variables

1. A variable is an object, characteristic, or property that can have different values.
2. A quantitative variable can be measured in some way.
3. A qualitative variable is characterized by its inability to be measured but it can be sorted into categories.

Random Variables

1. A random variable is one that cannot be predicted in advance because it arises by chance. Observations or measurements are used to obtain the value of a random variable.
2. Random variables may be discrete or continuous.

Discrete Random Variable

1. A discrete random variable has gaps or interruptions in the values that it can have.
2. The values may be whole numbers or have spaces between them.

Continuous Random Variable

1. A continuous random variable does not have gaps in the values it can assume.
2. Its properties are like the real numbers.

Populations and Samples

1. A population is the collection or set of all of the values that a variable may have. The entire category under consideration.
2. A sample is a part of a population. The portion of the population that is available, or to be made available, for analysis.

Population and Sampling

- Sampling: the process of selecting portion of the population.
- Representativeness: the key characteristic of the sample is close to the population.
- Sampling bias: excluding any subject without any scientific rational. Or not based on the major inclusion and exclusion criteria.

Example

- Studying the self esteem and academic achievement among college students.
- Population: all student who are enrolled in any college level.
- Sample: students' college at the University of Jordan.

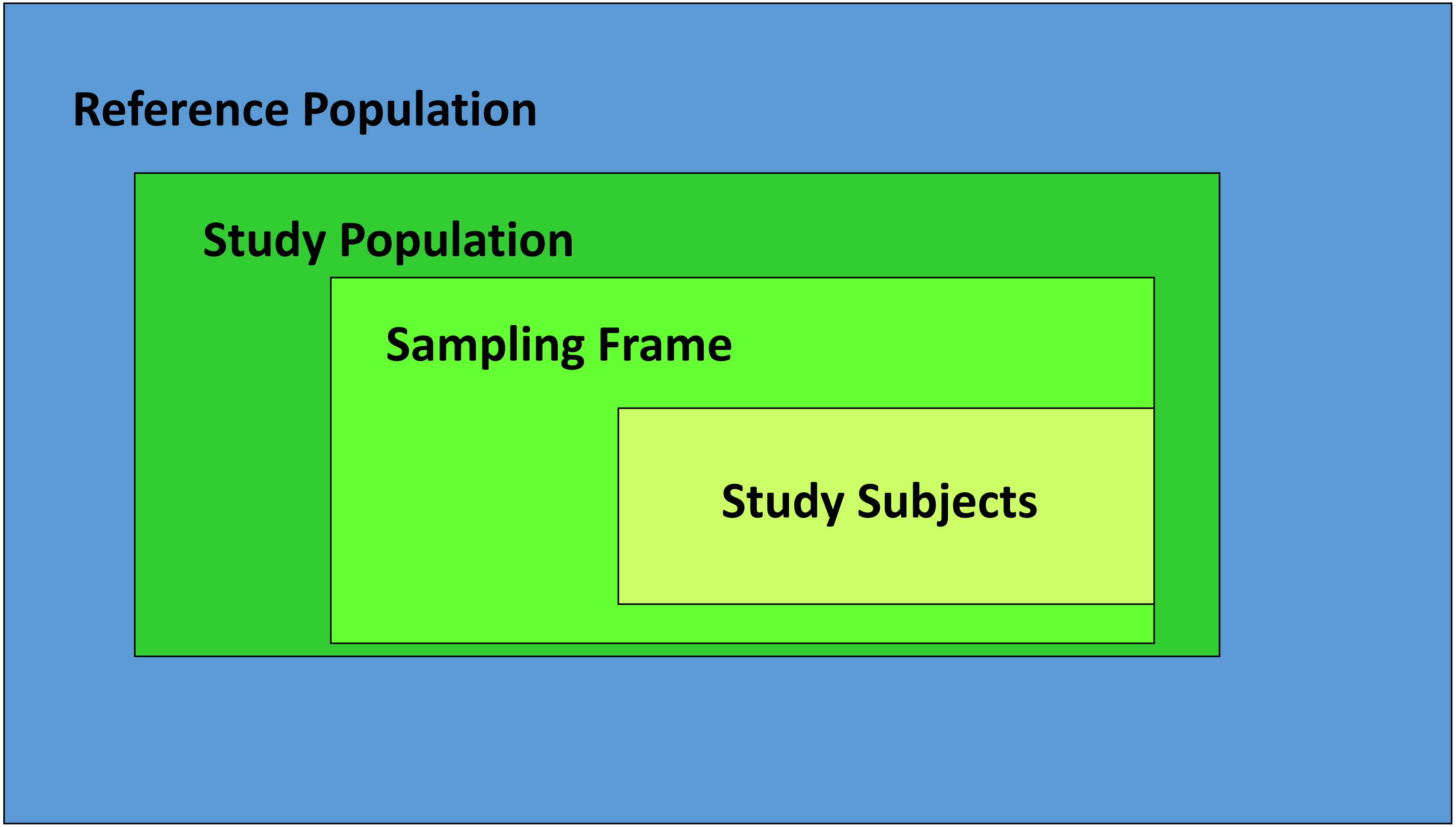
What is sampling?

- Sampling is the selection of a number of study units/subjects from a defined population.

Questions to Consider

- Reference population – to whom are the results going to be applied?
- What is the group of people from which we want to draw a sample (study population)?
- How many people do we need in our sample (Sample Size) ?
- How will these people be selected(Sampling Method)?

Sampling - Populations



Population

- Is a complete set of persons or objects that possess some common characteristic that is of interest to the researcher.
  - Two groups:
    The target population
    The accessible population

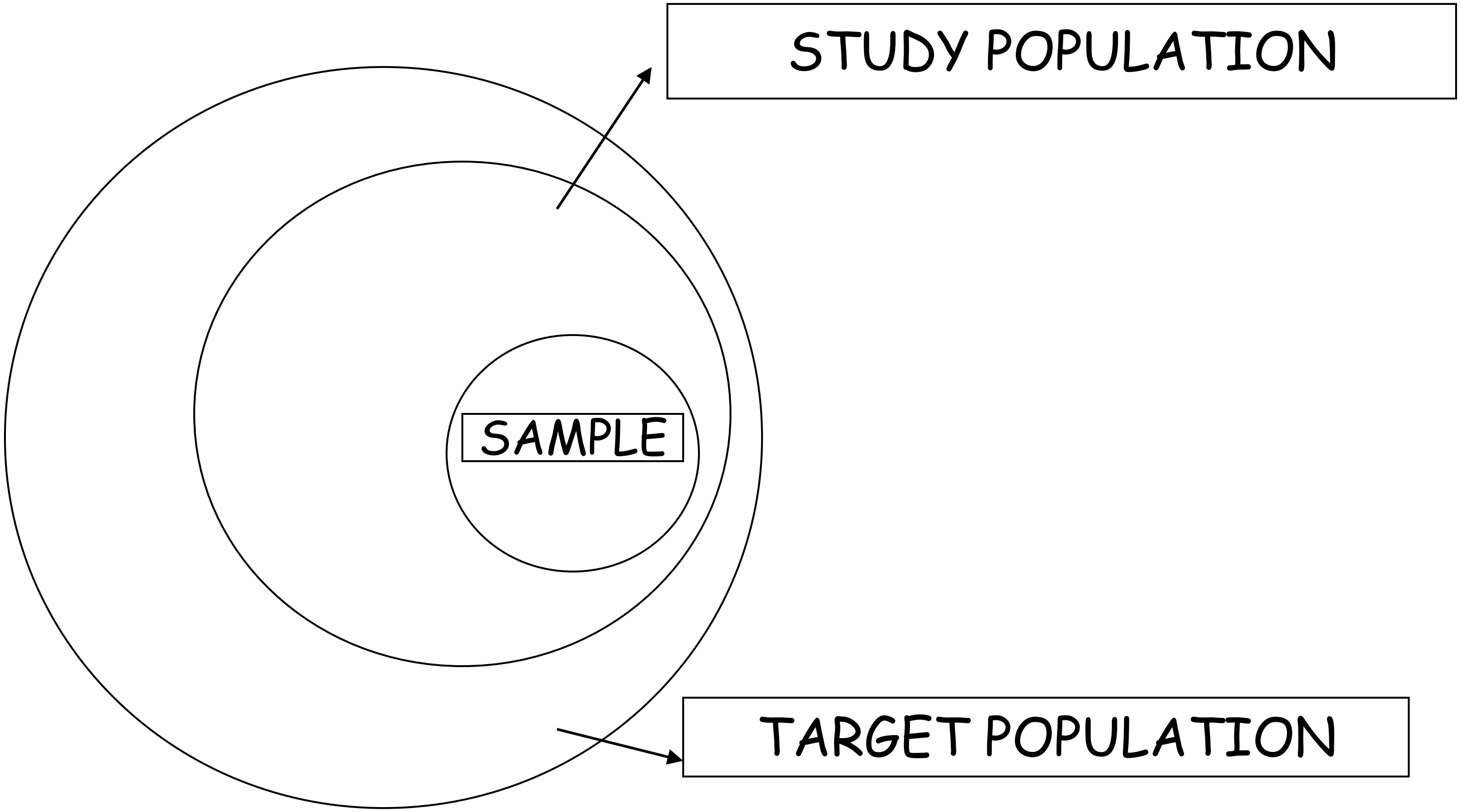
Target Population

- The entire group of people or objects to which the researcher wishes to generalize the findings of a study.
- Target population should meet the criteria of interest to the researcher
- Example: all people who were admitted to the renal unit for dialysis in Al-Basheer hospital during the period of 2014 – 2016.

Accessible Population

- The group of people or objects that is available to the researcher for a particular study

SAMPLING



Sampling

- Element: The single member of the population (population element or population member are used interchangeably)
- Sampling frame is the listing of all elements of a population, i.e., a list of all medical students at the university of Jordan, 2014-2016.

Sampling Methods

- Sampling depends on the sampling frame.
- Sampling frame: is a listing of all the units that compose the study population.

Primary vs Secondary Data

- **Primary data**. It is the data that has been compiled by the researcher using such techniques as surveys, experiments, depth interviews, observation, focus groups.
- **Types of surveys.** A lot of data is obtained using surveys. Each survey type has advantages and disadvantages.
  - **Mail**: lowest rate of response; usually the lowest cost
  - **Personally administered**: can “probe”; most costly; interviewer effects (the interviewer might influence the response)
  - **Telephone**: fastest
  - **Web**: fast and inexpensive

Primary vs Secondary Data

- **Secondary data.** The data that has been compiled or published elsewhere, e.g., census data.
  - The trick is to find data that is useful. The data was probably collected for some purpose other than helping to solve the researcher's problem at hand.
  - Advantages: It can be gathered quickly and inexpensively. It enables researchers to build on past research.
  - Problems: Data may be outdated. Variation in definition of terms. Different units of measurement. May not be accurate (e.g., census undercount).

Survey Errors

- Response Errors. Data errors that arise from issues with survey responses.
  - subject lies – question may be too personal or subject tries to give the socially acceptable response (example: “Have you ever used an illegal drug? “Have you even driven a car while intoxicated?”)
  - subject makes a mistake – subject may not remember the answer (e.g., “How much money do you have invested in the stock market?”
  - interviewer makes a mistake – in recording or understanding subject’s response
  - interviewer cheating – interviewer wants to speed things up so s/he makes up some answers and pretends the respondent said them.
  - interviewer effects – vocal intonation, age, sex, race, clothing of interviewer may influence response. An elderly woman dressed very conservatively asking young people about usage of illegal drugs may get different responses than young interviewer wearing jeans with tattoos on her body and a nose ring.

Types of Samples

- **Nonprobability Samples** – based on convenience or judgment
  - *Convenience* sample - students in a class, mall intercept
  - *Judgment* sample - based on the researcher's judgment as to what constitutes "representativeness" e.g., he/she might say these 20 stores are representative of the whole chain.
  - *Quota* sample - interviewers are given quotas based on demographics for instance, they may each be told to interview 100 subjects – 50 males and 50 females. Of the 50, say, 10 nonwhite and 40 white.
- The problem with a nonprobability sample is that we do not know how representative our sample is of the population.

Probability Samples

- **Probability Sample**. A sample collected in such a way that every element in the population has a known chance of being selected.

- One type of probability sample is a **Simple Random Sample**. This is a sample collected in such a way that every element in the population has an equal chance of being selected.

- How do we collect a simple random sample?
  - Use a table of random numbers or a random number generator.

TABLE 10–2. Random Numbers

21	71	89	96	97
82	59	22	78	12
76	93	64	79	28
20	60	70	34	51
93	58	36	93	90
68	63	19	21	91
18	32	36	27	71
58	80	[✓ 58]	67	50
66	25	[✓ 20]	31	62
17	25	[✓ 07]	94	18
02	29	[✓ 30]	15	92
55	06	[✓ 25]	09	26
38	11	[✓ 01]	47	93
42	47	[✓ 73]	25	84
82	04	[✓ 23]	08	88
37	24	[✓ 51]	98	05
94	58	85	86	71
37	92	[✓ 27]	20	58
29	64	[✓ 13]	05	24
85	48	[✓ 37]	37	21
20	56	91	53	66
33	23	13	82	54
62	11	[✓ 29]	17	37
01	57	73	53	97
34	19	[✓ 75]	62	16
81	10	[✓ 55]	36	36
92	50	32	68	82
37	33	43	20	08
10	50	18	85	27

Probability Samples

- Other kinds of probability samples (beyond the scope of this course).
  - *systematic random sample*.
    - Individuals are chosen at regular intervals from the sampling frame
    - Ideally we randomly select a number to tell us the starting point
    - every 5th household
    - every 10th women attending the conference
    - Sampling fraction = $\frac{\textbf{Sample size}}{\textbf{Study population}}$

$$\text{Interval size=} \frac{\textbf{study population}}{\textbf{Sample size}}$$

  - *stratified random sample*.
    - The population is sub-divided based on a characteristic and a simple random sample is conducted within each stratum
  - *cluster sample*
    - First take a random sample of clusters from the population of cluster. Then, a simple random sample within each cluster.

Variables

represent information that must be collected in order to meet the objectives of a study

- Measurable characteristic of a person, object or phenomenon which can take on different values
  - Weight
  - Distance
  - Monthly income
  - Number of children
  - Color
  - Outcome of disease
  - Types of food
  - Sex
- Variables allow clear definition of the core problem and influencing factors by introducing the concept of value

Types of Variables

- Continuous (e.g., height)
- Discrete (e.g., number of children)
- Categorical (e.g., marital status)
- Dichotomous (e.g., gender)
- Attribute variable vs. Active variable
  - Attribute Variable: Preexisting characteristics which researcher simply observes and measures. E.g. blood type.
  - Active Variable: Researcher creates or manipulates this. E.g. experimental drug.

Types of Variables (cont.)

Independent variable—the presumed cause (of a dependent variable)

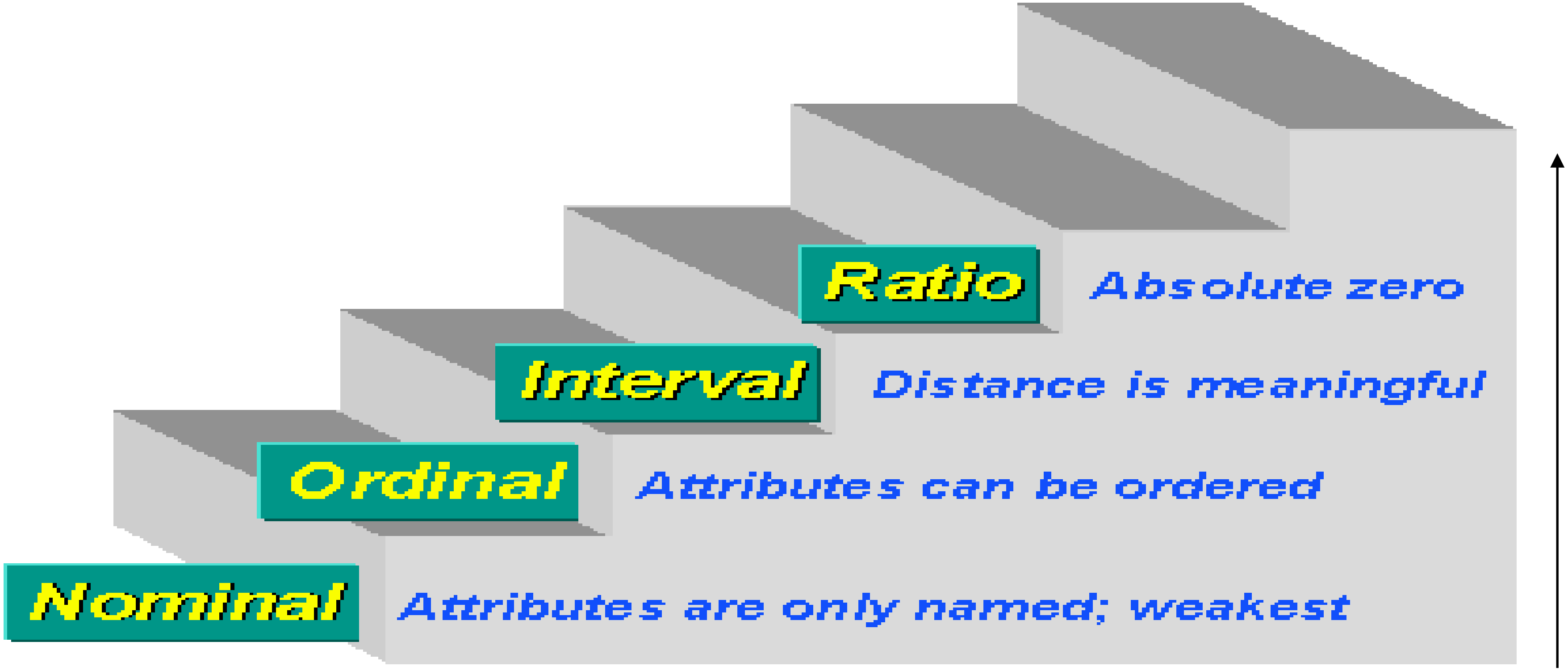
Dependent variable—the presumed effect (of an independent variable)

Example: Smoking (IV) → Lung cancer (DV)

Levels of Measurement

- Nominal
- Ordinal
- Interval
- Ratio

Levels of Measurement



Ratio Absolute zero

Interval Distance is meaningful

Ordinal Attributes can be ordered

Nominal Attributes are only named; weakest

Nominal Level of Measurement

- Categories that are distinct from each other such as gender, religion, marital status.
- They are symbols that have no quantitative value.
- Lowest level of measurement.
- Many characteristics can be measured on a nominal scale: race, marital status, and blood type.
- Dichotomous.
- Appropriate statistics: mode, frequency
- We cannot use an average. It would be meaningless here.

Ordinal Level of Measurement

- The exact differences between the ranks cannot be specified such as it indicates order rather than exact quantity.
- Involves using numbers to designate ordering on an attribute.
- Example: anxiety level: mild, moderate, severe. Statistics used involve frequency distributions and percentages.
- Appropriate statistics: same as those for nominal data, plus the median; but not the mean.

Interval level of Measurement

- They are real numbers and the difference between the ranks can be specified.
- Equal intervals, but no “true” zero.
- Involves assigning numbers that indicate both the ordering on an attribute, and the distance between score values on the attribute
- They are actual numbers on a scale of measurement.
- Example: body temperature on the Celsius thermometer as in 36.2, 37.2 etc. means there is a difference of 1.0 degree in body temperature.
- Appropriate statistics
  - same as for nominal
  - same as for ordinal plus,
  - the mean

Ratio level of Measurement

- Is the highest level of data where data can categorized, ranked, difference between ranks can be specified and a true or natural zero point can be identified.
- A zero point means that there is a total absence of the quantity being measured.
- All scales, whether they measure weight in kilograms or pounds, start at 0. The 0 means something and is not arbitrary (SUBJECTIVE).
- Example: total amount of money.

What Type of Dats To collect?

- The goal of the researcher is to use the highest level of measurement possible.
  - Example: Two ways of asking about Smoking behavior. Which is better, A or B?

(A)Do you smoke? □Yes □No

(B)How many cigarettes did you smoke in the last 3 days (72 hours)?

(A) Is nominal, so the best we can get from this data are frequencies. (B) is ratio, so we can compute: mean, median, mode, frequencies.

Parameter and Statistic

- Parameter is a descriptive measure computed from the data of the population.
  - The population mean, μ, and the population standard deviation, σ, are two examples of population parameters.
  - If you want to determine the population parameters, you have to take a census of the entire population.
  - Taking a census is very costly.
  - Parameters are numerical descriptive measures corresponding to populations.
  - Since the population is not actually observed, the parameters are considered unknown constants.
- Statistic is a descriptive measure computed from the data of the sample.
  - For example, the sample mean, $\bar{x}$, and the standard deviation, s, are statistics.
  - They are used to estimate the population parameters.

Statistics

- It is a branch of applied mathematics that deals with collecting, organizing, & interpreting data using well-defined procedures in order to make decisions.
- The term parameter is used when describing the characteristics of the population. The term statistics is used to describe the characteristics of the sample.
- Types of Statistics:
    - Descriptive Statistics. It involves organizing, summarizing & displaying data to make them more understandable.
    - Inferential Statistics. It reports the degree of confidence of the sample statistic that predicts the value of the population parameter

Descriptive Statistics

- Those statistics that summarize a sample of numerical data in terms of averages and other measures for the purpose of description, such as the mean and standard deviation.
  - Descriptive statistics, as opposed to inferential statistics, are not concerned with the theory and methodology for drawing inferences that extend beyond the particular set of data examined, in other words from the sample to the entire population. All that we care about are the summary measurements such as the average (mean).
  - This includes the presentation of data in the form of graphs, charts, and tables.

Descriptive Statistics

- Measures of Location
  - Measures of Central Tendency:
    - Mean
    - Median
    - Mode
  - Measures of noncentral Tendency-Quantiles:
    - Quartiles.
    - Quintiles.
    - Percentiles.
- Measure of Dispersion (Variability):
  - Range
  - Interquartile range
  - Variance
  - Standard Deviation
  - Coefficient of variation
- Measures of Shape:
  - Mean > Median-positive or right Skewness
  - Mean = Median- symmetric or zero Skewness
  - Mean < Median-Negative of left Skewness

Statistical Inference

- Is the procedure used to reach a conclusion about a population based on the information derived from a sample that has been drawn from that population.

Inferential Statistics

- Inferential statistics are used to test hypotheses (prediction) about relationship between variables in the population. A relationship is a bond or association between variables.
- It consists of a set of statistical techniques that provide prediction about population characteristics based on information in a sample from population. An important aspect of statistical inference involves reporting the likely accuracy, or of confidence of the sample statistic that predicts the value of the population parameter.

Inferential Statistics

- Bivariate Parametric Tests:
  - One Sample t test (t)
  - Two Sample t test (t)
  - Analysis of Variance/ANOVA (F).
  - Pearson’s Product Moment Correlations (r).
- Nonparametric statistical tests: Nominal Data:
  - Chi-Square Goodness-of-Fit Test
  - Chi-Square Test of Independence
- Nonparametric statistical tests: Ordinal Data:
  - Mann Whitney U Test (U
  - Kruskal Wallis Test (H)

Research Hypothesis

- A tentative prediction or explanation of the relationship between two or more variables
- It's a translation of research question into a precise prediction of the expected outcomes
- In some way it's a proposal for solution/s
- In qualitative research, there is NO hypothesis

Research Hypothesis

- States a prediction
- Must always involve at least two variables
- Must suggest a predicted relationship between the independent variable and the dependent variable
- Must contain terms that indicate a relationship (e.g., more than, different from, associated with)

Hypotheses Criteria

- Written in a declarative form.
- Written in present tense.
- Contain the population
- Contain variables.
- Reflects problem statement or purpose statement.
- Empirically testable.

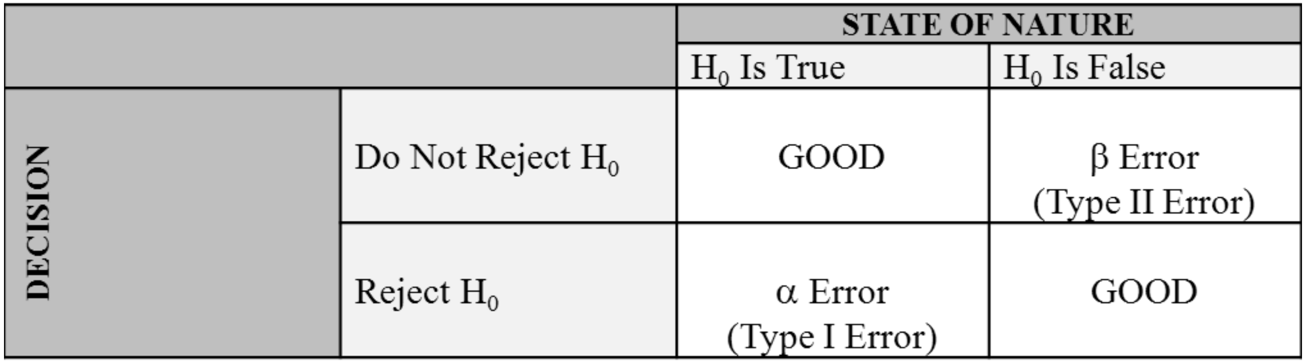
Hypothesis Testing

- A hypothesis is made about the value of a parameter, but the only facts available to estimate the true parameter are those provided by the sample. If the statistic differs (and of course it will) from the hypothesis stated about the parameter, a decision must be made as to whether or not this difference is *significant*. If it is, the hypothesis is rejected. If not, it cannot be rejected.
- H_0: The null hypothesis. This contains the hypothesized parameter value which will be compared with the sample value.
- H_1: The alternative hypothesis. This will be "accepted" only if H_0 is rejected.

  Technically speaking, we never accept H_0 What we actually say is that we do not have the evidence to reject it.

Two Types of Errors: Alpha and Beta

- Two types of errors may occur: α (alpha) and β (beta). The α error is often referred to as a Type I error and β error as a Type II error.
  - You are guilty of an alpha error if you reject H_0 when it really is true.
  - You commit a beta error if you "accept" H_0 when it is false.



		STATE OF NATURE	
		H_0 Is True	H_0 Is False
DECISION	Do Not Reject H_0	GOOD	β Error (Type II Error)
	Reject H_0	α Error (Type I Error)	GOOD

Types of Errors

If You……	When the Null Hypothesis is…	Then You Have…….
Reject the null hypothesis	True (there really are no difference)	Made a Type I Error
Reject the null hypothesis	False (there really are difference)	☻
Accept the null hypothesis	False (there really are difference)	Made Type II Error
Accept the null hypothesis	True (there really are no difference)	☻

Two Types of Errors: Alpha and Beta

TRADEOFF!

1. There is a tradeoff between the alpha and beta errors. We cannot simply reduce both types of error. As one goes down, the other rises.
2. As we lower the α error, the ß error goes up: reducing the error of rejecting H_0 (the error of rejection) increases the error of “Accepting” H_0 when it is false (the error of acceptance).

Tradeoff in Type I / Type II Errors: Examples

- Quality Control.

A. A company purchases chips for its smart phones, in batches of 50,000. The company is willing to live with a few defects per 50,000 chips. How many defects?

B. If the firm randomly samples 100 chips from each batch of 50,000 and rejects the entire shipment if there are ANY defects, it may end up rejecting too many shipments (error of rejection). If the firm is too liberal in what it accepts and assumes everything is "sampling error," it is likely to make the error of acceptance.

C. This is why government and industry generally work with an alpha error of .05

Steps in Hypothesis Testing

1. Formulate H_0 and H_1. H_0 is the null hypothesis, a hypothesis about the value of a parameter, and H_1 is an alternative hypothesis.

   e.g., H_0: μ=12.7 years; H_1: $\mu \neq$12.7 years

2. Specify the level of significance (α) to be used. This level of significance tells you the probability of rejecting H_0 when it is, in fact, true. (Normally, significance level of 0.05 or 0.01 are used)

3. Select the test statistic: e.g., Z, t, F, etc.

4. Establish the critical value or values of the test statistic needed to reject H_0. DRAW A PICTURE!

5. Determine the actual value (computed value) of the test statistic.

6. Make a decision: **Reject H_0** or **Do Not Reject H_0**.

I Hope This Will Help

شكراً