Basics of Statistics~Definitions

Statistics ~ *Oxford Dictionary*	*"The science of collecting and analysing numerical data in large quantities, especially for the purpose of **inferring**... from those in a representative **sample**."*
Statistics ~ *simple*	The practice of *collecting data from a **small(er)** number of individuals to draw conclusions about a **large(er)** number of individuals*.
Descriptive Statistics	This involves organizing, and summarizing data through *numerical, tabular, and graphical means.* **Estimation Procedures** are *formulas* used to obtain *estimates* (i.e. **descriptive statistics**) of **population** traits from the **sample** data.
Inference Procedures	The *formulas used to quantify the magnitude of **error** in generalizing* from a 'part' (**sample**) to the 'whole' (**population**) e.g., *t-test, *Z-test, *chi-square test, *Wald test, etc. The **Inference "Statistics"** obtained may reflect the **reliability** of the result or the possibility of **error**.
Inference	A conclusion reached based on evidence and reasoning. *In statistics, we learn things from **samples** and infer to the **population**.*
Population	The *entire set of individuals* to which the study are to be extrapolated *(generalized).*
Parameter	This is a *numerical summary* of a **population**.

Basics of Statistics~Definitions (cont)

Sample	A subset of the population
Sampling Plan	The *methodology* for choosing the **sample**.
Statistic	This is a *numerical summary* of a **sample**.

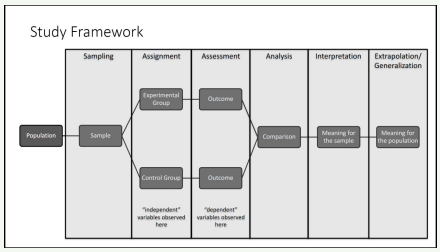
Types of estimates

Point	Interval

Types of Study Design

• Randomized Controlled Trials • Cohort Studies • Case-Control Studies • Cross-Sectional Studies

Framework of a Typical Study



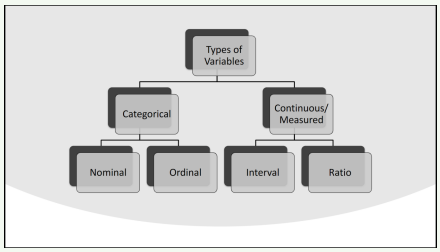
Exploratory Data Analysis

	The first step in dealing with data is to organize your thinking about the data.
Explor-atory data analysis	This is the process of *using statistical tools and ideas to examine data in order to describe their main features*.
Steps of an explor-atory data analysis	• Examine each variable first. • Then study the relati-onships among the variables. • Begin with a graph or graphs. • Add numerical summaries of specific aspects of the data

Variables

The characteristics of the individuals within the population

Variables Tree



All about Variables

Categorical Data or Variables	Data can only take on *discrete values*. It allows for classification of individuals based on some attribute or characteristic.
Nominal Data or Variables	apply in name only – *no inherent ordering* (e.g. *blood type, hair color*).
Ordinal Data or Variables	Data can be *ranked in order* but only take on discrete values (e.g. *satisfaction score, Glasgow Coma Score*).
Continuous/Measured Data or Variables	These values can be *added or subtracted and provide meaningful results*.
Interval Data or Variables	The difference between each number/value is equal (e.g. temperature in Celsius, IQ score). No absolute 0 *(Zero has a meaningful value)*
Ratio Data or Variables	These values are on *an interval scale with an absolute zero with no meaningful value* (e.g. weight, temperature in Kelvin)

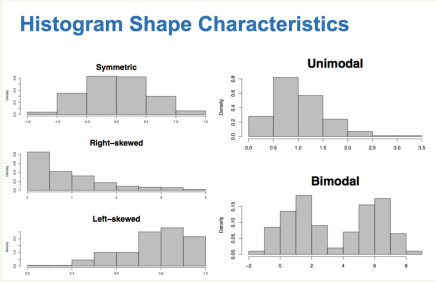
Graphing Variables

Categorical-**Pie Chart**	This shows the distribution of a categorical variable as a “pie” whose slices are sized by the counts or percent for the categories.
Categorical-**Bar Chart**	This represents each category as a bar whose heights show the category counts or percent.
Continuous-Histograms	Count how many (or the percentage of individuals) that fall into each interval)
Continuous-Boxplot	• The Lower Inner Fence (LIF) • LIF = Q1 – 1.5*IQR* • *The Upper Inner Fence (UIF) • UIF= Q3 + 1.5*IQR • The Lower Adjacent Value • The actual data value just inside the LIF • The Upper Adjacent Value • The actual data value just inside the UIF

Tables-Continuous/categorical variables

Advantages :)	• Gives the reader a compact and structured synthesis of information • Shows a lot of detail in a small amount of space
Disadvantages :(	• Because the reader only sees numbers, the table may not be readily understood without comparing it with other tables

Skew



Skewness

Right or positively skewed distributions will yield skewness values > 0	Left or negatively skewed distributions will yield skewness values < 0

Kurtosis

This is often made in comparison to a Bell Shaped/Normal distribution	Kurtosis is a measure of how "peaked" or "flat" a distribution is
– If a distribution has an excessive amount of observations close to the mean (meaning fewer observations in the tails) it will have a more peaked appearance (Platykurtic Kurtosis < 3)	– If a distribution has an excessive amount of observations which are further away from the mean (ex. in the tails) it will have a flattened appearance as if it really has no tails at all (Leptokurtic Kurtosis > 3)

So, about sampling...

Probability sampling	Random selection-equal chance for selection for every member of the **population**
Non-probability sampling	Convenience or voluntary self-selection increases the likelyhood of selection of one or more participants
Parametric tests	In these tests, reasonable and evidence-supported assumptions must be made about the distribution. They can be used to make strong statistical inferences when data are collected using probability sampling.
Non-parametric tests	Very few assumptions are made, if any, about the population distribution. They are more appropriate for non-probability samples, but they result in weaker inferences about the population.

So, about sampling... (cont)

Significance level (alpha)	The risk of rejecting a true null hypothesis that you are willing to take, usually set at 5%.
Statistical power	The probability of your study detecting an effect of a certain size if there is one, usually 80% or higher.
Expected effect size	A standardized indication of how large the expected result of your study will be, usually based on other similar studies.

So, about sampling...

Probability sampling	Random selection-equal chance for selection for every member of the **population**
Non-probability sampling	Convenience or voluntary self-selection increases the likelyhood of selection of one or more participants
Parametric tests	In these tests, reasonable and evidence-supported assumptions must be made about the distribution. They can be used to make strong statistical inferences when data are collected using probability sampling.
Non-parametric tests	Very few assumptions are made, if any, about the population distribution. They are more appropriate for non-probability samples, but they result in weaker inferences about the population.
Significance level (alpha)	The risk of rejecting a true null hypothesis that you are willing to take, usually set at 5%.

So, about sampling... (cont)

Statistical power	The probability of your study detecting an effect of a certain size if there is one, usually 80% or higher.
Expected effect size	A standardized indication of how large the expected result of your study will be, usually based on other similar studies.

So, about sampling...

Probability sampling	Random selection-equal chance for selection for every member of the **population**
Non-probability sampling	Convenience or voluntary self-selection increases the likelyhood of selection of one or more participants
Parametric tests	In these tests, reasonable and evidence-supported assumptions must be made about the distribution. They can be used to make strong statistical inferences when data are collected using probability sampling.
Non-parametric tests	Very few assumptions are made, if any, about the population distribution. They are more appropriate for non-probability samples, but they result in weaker inferences about the population.
Significance level (alpha)	The risk of rejecting a true null hypothesis that you are willing to take, usually set at 5%.
Statistical power	The probability of your study detecting an effect of a certain size if there is one, usually 80% or higher.

So, about sampling... (cont)

Expected effect size	A standardized indication of how large the expected result of your study will be, usually based on other similar studies.

So, about sampling...

Probability sampling	Random selection-equal chance for selection for every member of the **population**
Non-probability sampling	Convenience or voluntary self-selection increases the likelyhood of selection of one or more participants
Parametric tests	In these tests, reasonable and evidence-supported assumptions must be made about the distribution. They can be used to make strong statistical inferences when data are collected using probability sampling.
Non-parametric tests	Very few assumptions are made, if any, about the population distribution. They are more appropriate for non-probability samples, but they result in weaker inferences about the population.
Significance level (alpha)	The risk of rejecting a true null hypothesis that you are willing to take, usually set at 5%.
Statistical power	The probability of your study detecting an effect of a certain size if there is one, usually 80% or higher.
Expected effect size	A standardized indication of how large the expected result of your study will be, usually based on other similar studies.