Introduction to Statistics

Outline of Statistics we are going to learn all together:

1. Descriptive Statistics
2. Data Visualization
3. Probability Distribution
4. Hypothesis Testing
5. Regression Analysis

What is Statistics?

Statistics is a collection of methods for **collecting, displaying, analying and drawing conclusions from data**.

When we have to talk about graphs, as a statistician we have to use the language of statistics. For example:

- Average income in Pakistan (Mean)
- Highest score in cricket match (Maximum)
- Fastest Bowler (Maximum)
- Lowest run scored (Minimum)
- 40% teachers in Pakistan are female (Percentage)
- Tomorrow is expecting rain (Likelihood)
- Dollar rate keeps increasing and decreasing (Varience)
- Boys make more mess than girls (t-test)

Types of Data

Data Types # 1

1. **Cross Sectional**: Data collected at one point. For example, how many people are going to match today, tomorrow, and the day after tommorrow? This comparison will be individual for each day.
2. **Time Series**: Data collected over the different time points. For example, how many people watched the match this whole week? This comparison will be observed collectively through out the week.

Data Types # 2

1. **Univariate**: Data contains a single variable to measure entity. For example, how much the plant grows with 1ltr of water?
2. **Multi-variate**: Data contains more than two variable to measure something. For examples, how much the plant grow with 1ltr of water and 1kg fertilizer.

Types of Variables

A variable is a characteristic that can be measured and that can assume different values. Height, age, income, province or country of birth, and type of housing are all examples of variables. Vairables may be classified into two main categories: **categorical and numeric**. Each categroy is then classified in two subcategories: *nominal or ordinal* for **categorical variables**, *discrete or continous* for **numeric variables**.

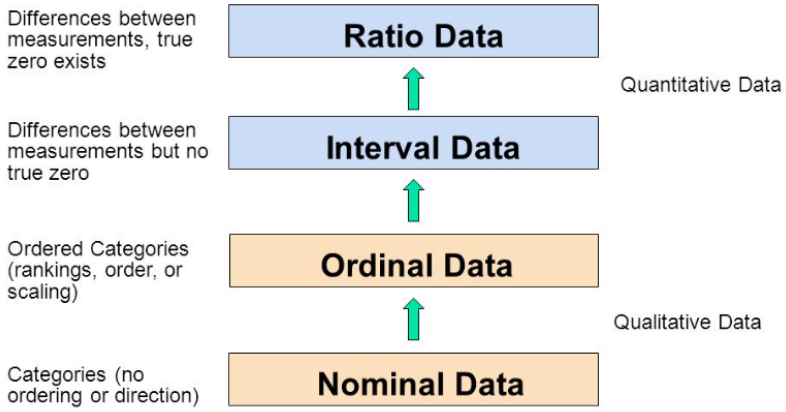
Categrorical Variables

A categorical variable (also called **qualitative variable**) refers to a characteristic that can't be quantifiable. Categorical variables can be either nominal or ordinal.

Numeric Variables

A numeric variable (also called **quantitative variable**) is a quantifiable characteristic whose values are numbers (except numbers which are codes standing up for categories). Numeric variables may be either continous or discrete.

Four Levels of Measurement



1. **Nominal**: A nomial scale describes a variable with categories that do not have a natural order or ranking. We can code nominal variables with numbers if we want, but the order is arbitrary and any calculations, such as computing a mean, median, or standard deviation, would be meaningless. For example, genotype, blood type, zip code, gender, race, political party.

    - Nominal variable can be sub-divided into two variable: a. *Binomial*: Categories in which only two choices are given. For example, do you like travelling or not? How does the food taste? b. *Multinomial*: Categories in which more than two choices are given. For example, what are the means of transport to go school?
2. **Ordinal**: An ordinal scale is one where the order matters but not the difference between values. For examlpe, income level (low income, middle income, high income), education level (high school, BS, MS, PhD).
3. **Interval**: An interval scale is one where there is order and the difference between two values is meaningful. For example, temperature(25-50 Celcius), Math score(75-100), credit score(300-850).
4. **Ratio**: A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0. When the variable equals 0, there is none of that variable. For example, pulse rate, weight, length, temperature in Kelvin(0.0 Kelvin really does mean "no heat"), survival time.

OK to compute...	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Median and percentiles	No	Yes	Yes	Yes
Add or subtract	No	No	Yes	Yes
Mean, standard deviation, standard error of the mean	No	No	Yes	Yes
Ratios, coefficient of variation	No	No	No	Yes

Measures of Central Tendency

There are three main measures of central tendency: **mean, median and mode**. Each of these measures describes a different indication of the typical or central value in the distribution.

Mean

Mean is the average or the most common value in a collection of numbers. It is the sum divided by the number of observations.

Mean and Its Properties

- Mean is meaningful for interval and ratio data.
- Outliers change the mean of the data, therefore, median is useful.

Median

Median is the middle number in a sorted (ascending or descending) list of numbers.

Median and Its Properties

- It is unique for each dataset.
- Outliers don't have any effect on median.
- Ratio, interval and ordinal data has best use with median.
- Typically, there is 50% data is on the right and 50% on left side of median.

Mode

Mode is the value the occurs most frequently in the data.

Notions and Terms in Statistics

There are four big terms in statistics:

1. **Population:** A population is the entire group that we want to draw conclusions about. Population research has more power and more acceptable results. For example, advertisements for Data Science jobs in the Netherlands.
2. **Sample:** A sample is the specific group that we collect data from. The size of the sample is always less than the total size of the population. Samples research is used to reduce the cost of data collection. For example, top 50 search results for advertisements for Data Science jobs in teh Netherlands on May 1, 2020.
3. **Parameter:** A parameter is a number that summarizes some aspect of the population as a whole.
4. **Statistic:** A statistic is a number computed from the sample data.

    - *population* --> *sample* --> *data* --> *statistics* --> *parameters*
    - The source of data always comes from population or from samples.
    - When we analysis data from population and sample, the results will almost be same but there will be higher chances of error in sample as compared to population.

- **Descriptive statistics** is the branch of statistics that involves organizing, displaying, and describing data.

- **Inferential statistics** is the branch of statistics that involves drawing conclusion about a population based on information contained in a sample taken from that population.

- A **measurement** is a number or attribute computed for each member of a population or of a sample. The measurements of sample elements are collectively called the sample data.

- **Qualitative data** are measurements for which there is no natural numerical scale, but which consist of attributes, labels, or other non-numerical characteristics.
- **Quantitative data** are numerical measurements that arise from a natural numerical scale.

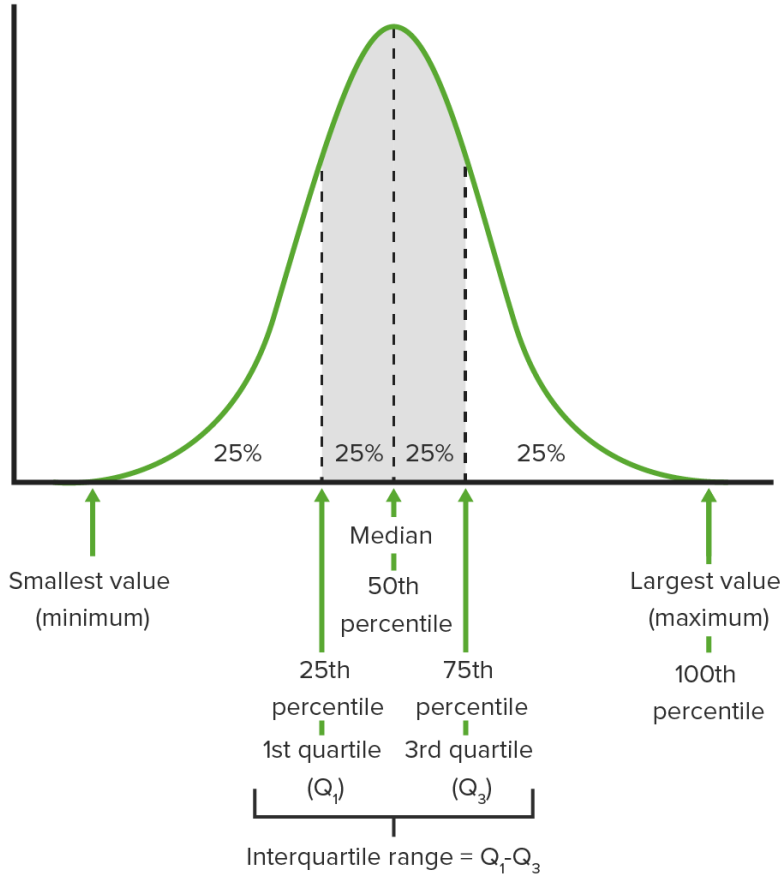
Statistical Symbols

Symbol	Meaning	Refers to a value in a
n	Sample size	Sample
$\overline{x}$	Sample mean	Sample
μ	Population mean	Population
s	Sample standard deviation	Sample
σ	Population standard deviation	Population
$\hat{p}$	Sample proportion	Sample
p	Population proportion	Population
p	Probability of success	Binomial distribution
x, z	Represents a specific, observed data value	Sample
X, Z	Represents a random variable and all possible values it could be	Population

Measure of Dispersion (Variability, Scatter or Spread)

What is Dispersion?

Dispersion is how much data spread around its mean. The spread of a data set can be described by a range of descriptive statistics including standard deviation, variance, and interquartile range and many more. The range between the minimum and the maximum is called dispersion of data.

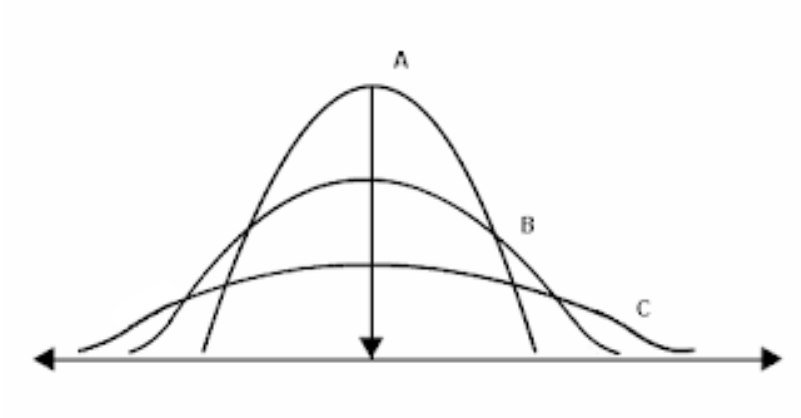


Role of STD and Mean?

The higher the value for the std from the mean, the more spread out the values are in a sample. Conversely, the lower the values for the std from the mean, the more tightly packed together the values are in a sample.

Std is important because it tells us how spread out the values are in a given dataset. For instance:

A: Mean = 30.1, B: Mean = 30.2, C: Mean = 30.4
A: STD = 4.76, B: STD = 10.99, C: STD = 13.85



Reliability of STD and Mean

Mean only gives us a small picture, therefore, means are incomplete withoud dispersion(std).

Our data driven policies depend on 50% on mean and 50% on std. That's why mean with std is more useful than only mean by itself.

What is Standard Error?

The standard error (SE) of the mean measures how far the sample mean (average) of the data is likely to be from the true population mean. The standard error is always smaller than the standard deviation.

Fundamentals of Visualization

For graphs there are two types of visualization depends on the variable type:

For categorical variable

- Count (plot type): Male vs. Female, True vs. False

For continuous variable

- Scatter plot
- Statistical proportions (means and their comparison)

Choosing a Statistical Method

Main Objectives of Choosing a Right Statistical Method

1. Gain knowledge about how to choose right statistical method.
2. Do's and don't's of statistics.
3. Reliable results.
4. Paper revisions with proof of statistical test.
5. Making Data Visualization based on statistical method.
6. Interpreting results based on final report/conclusion.

What are tests and their types?

There are two types of statistical tests:

1. **Parametric Tests**: These tests have more reliable results because first we have to meet the assumptions that the data comes from a normal distribution.
2. **Nonparametric Tests**: These tests have less reliable results because they don't have to meet the assumptions. These tests often used in industrial level. These tests count the rank of the data.

Steps Before Data Analysis

These steps have to be considered before starting data analysis:

Steps-1: Check for Normality Test

We should check the normality of the data using **Shapiro-Wilk test** (specific and reliable) or using **Kolmogorov-Smirnov test** (general and less reliable).

Step-2: Check for Homogeneity Test

Check whether the variance of the variable in two or more populations (or subgroups of a population) are same or not. We use **Levene'e test** to find the equality of variance for a variable calculated for two or more groups.

Step-3: Know the Purpose of Test

We make data analysis based on the **purpose** of our research question. For example, how many family members like to get to pizza and how many like burger so we can order food based on majority of votes.

There are two types of purposes:

1. **Comparison**: When we see difference it is called camprison. If our purpose is comparison then it has to be between two groups atleast. For example, Male vs. Female, Grouping individuals based on ethnicity.
2. **Relationship**: When we see connection in our purpose then it is called **relationship**. In this type of purpose we seek connection, correlation, causation, and prediction. For example, does fertilizer application increase crop growth? Can food predict weight of a group of individuals?

Step-4: Data Type

The fourth step before making data analysis is to know the **type of data** we are working with. For example, categorical, continuous, ordinal, discrete etc.

Steps-5: Choose a Statistical Test

The last step is choosing a statistical test. There are **three main families** of statistical test:

Family-1	Family-2	Family-3
Purpose: Comparison	**Purpose**: Comparison	**Purpose**: Relationship
Data: Categorical only	**Data**: Categorical & Continuous	**Data**: Continuous only

Three Families of Parametric Test

Family-1 (When and where to use?)
Purpose: Comparison
Data: Categorical only

With categorical variables, we can't calculate a mean or standard deviation. Therefore, we can not use **Family-1** for parametric test.

Family-2 (When and where to use?)
Purpose: Comparison
Data: Categorical and Continuous

1. **T-Test**: The t-test tells us how significant the difference between groups are. In other words it lets us know if those differences (*measured in means*) could have happened by chance. There are three types of t-test:
   a. **One-sample t-test**: A one sample t-test tests the mean of a single group against a "known mean". For example, how many boys

are taller and shorter than 5.9ft?
b. **Paired-samples t-test:** A paired sample t-test compares the means of two measurements taken from the same group. For example, boys test marks in Maths and Stats.
c. **Unpaired-samples t-test:** An unpaired samples t-test compares the mean for two different groups. For example, boys and girls marks in english.

2. **ANOVA (Analysis of Variance)**: When there are 3 or more groups are involved then we preform ANOVA test instead of t-test. An ANOVA test is a way to find out if survey or experiement results are significant. In other words, they help us figure out if we need to reject the **null hypothesis** or accept the **alternate hypothesis**. Basically, we are testing groups to see if there is a difference between them. For example, a group of psychiatric patients are trying three different therapies; counseling, medication and biofeedback. We want to see if one therapy is better than the others. There are two types of ANOVA that are commonly used:
a. **One-way ANOVA:** This type of ANOVA has one independent variable. For example, the independent variable might ba a brand of drink.
b. **Two-way ANOVA:** This type of ANOVA test has two independent variables. For example, the independent variables might be brand of drink and how many calories it has.

Family-3 (When and where to use?)
Purpose: Relationship
Data: Continuous only

1. **Correlation:** Correlation means association, more precisely it is a measure of the extent to which two variables are related. Therefore, when one variable increases as the other variable increases, or one varaible decreases while the other decreases. For example, a positive correlation would be height and weight. ***Pearson's correlation*** is used for this type of test.
2. **Regression:** Regression is used to determine the strength of the relationship between one dependent variable (usually denoted by Y) and one or a series of independent variables (depending on what regression problem we are working with). This helps us to find the data points that are not measured yet. For example, missing values can be predicted with regression.

Important Things About Parametric Tests

The assumptions about data is that it is normally distributed or follow Gaussian distribution. If we do not follow the assumptions and break the trusts of ***Three test families*** then the derived results won't be reliable.

If the Assumptions are Not Normalized

1. Test the data with **Shapiro-Wilk** or **Kolmogorov-Smirnov** test.
2. Check the variance of the variable using **Levene's** test.
3. Nomalize data using one of following techniques:

- Standardization
- Min-max scaling
- Log transformation

4. Despite all that if the data is still not normalized then use alternative which is nonparametric tests.

Three Families of Nonparametric Tests

Family-1 (When and where to use?)
Purpose: Comparison
Data: Categorical only

1. **Chi-squared test**: This test is used if two categorical variables are related in some population. It can be used with any number of levels or groups. For example, a scientist wants to know if education level and marital status are related for all people in the country. There are two types of chi-squared test:
   a. Chi-squared test of homogeneity
   b. Chi-squared test of independence

Family-2 (When and where to use?)
Purpose: Comparison
Data: Categorical and Continuous

1. Instead of *one-sample test* we use **one-sample Wilcoxon signed rank test** for nonparametric test.
2. We use **Wilcoxon test** for nonparametric instead of *paired t-test*.
3. **Mann Whitney's U-test** instead of *unpaired t-test*.
4. **Kruskal-Wallis test** is used as the alternate of *ANOVA test*.

Family-3 (When and where to use?)
Purpose: Relationship
Data: Continuous only

1. As alternative of *Pearson's correlation*, **Spearman's correlation** or **Kendall's Tau** can be used.
2. **Regression** can be used for both parametric & nonparametric tests.

Kinds of ANOVA

ANOVA (Analysis of Variance): It is used for **comparison** between *categorical and continuous* data when there are 3 or more levels or groups are involved. There are other types of ANOVA as well.

ANCOVA (Analysis of Co-Variance): It compares the means of 3 or more independent groups which can not be tested by ANOVA because the variables are affected by co-variance. For example, teacher wants to know if three different studying techniques have an impact on exam scores, but first the teacher wants to account for student's current grade in the class.

MANOVA (Multi-variate Analysis of Variance): It is used to analyze how one or more *factor variables* effects multiple *reponse variables*. For example, we might want to analyze how level of education affects both annual income and test score.

MANCOVA (Multi-variate Analysis of Co-Variance): It is identical to MANOVA, both feature two or more *response variables*, but the key difference between the two is the nature of independent variables. MANCOVA includes one or more *factor variables* as well as one or more *covariates*. For example, how level of eduction and number of study hours effects both annual income and test score.

Some Other Tests

- Reliability Tests
  - Kunder-Richardson's Formula 20 an 21 (KR20/21)
  - Cronbach's Alpha
- Inter-rater Reliability Tests
  - Krippendorf's Alplha (for categorical or continuous variables)
  - Fleis's Kappa (for categorical variables only)
- Validity Tests
  - Krippendorf's Alplha (for categorical or continuous variables)
  - Fleis's Kappa (for categorical variables only)
- Sample Size Computation - This makes sure how many samples are valid?
  - Cochran's Q Test
  - Yaman's Test
  - Many more...

> ***Final Note***: If the data is already normalized then next step should be performing **parametric tests**, otherwise, first normalize the data. But if the data can't be normalized then move towards the **nonparametric tests**.