Introduction to Statistics and Probability

WiMLDS Accra and AIMS Mathematics Bootcamp

AIMS Ghana

African Institute for Mathematical Sciences
Ghana

May 19, 2024

Overview

1. Introduction
2. Purpose and Learning Outcomes
3. Mesures of Central Tendencies
4. Discrete & Continuous Probability Distributions
5. Hypothesis Testing (Test of Significance)
6. Exploratory Data Analysis(EDA)

Introduction

- We introduce two important concepts: Statistics and Probability.
- There are two major types of variables used in almost every field,these are: **non-stochastic** or deterministic and **stochastic** or random variables.
- Stochastic variables have an associated probability structure
- for example, tossing a coin- we can't tell with certainty which side of the coin will sure up
- Non-stochastic variables are deterministic in nature without a probability attachment
- E.g, interest and annuity calculations based on fixed time periods.

Purpose and Learning Outcomes

- The purpose of this presentation is to equip students with basic ideas of statistics and probability, their use and applications.
- By the end of this lesson, students should be able to understand;
  1. Measures of Central Tendencies
  2. Discrete and Continuous Probability Distributions
  3. Inferential Statistics(Hypothesis Testing)
  4. How the above 3 are applied in Machine Learning

Mesures of Central Tendencies

AIMS Ghana

- Before that, let us talk about Data Reduction Techniques.
  - The process of putting data in such a way such that meaning is made is known as **Data Reduction**.
  - To determine the significance use of data, it must first be organize into some form so that at a mere glance, one can visualize the data and draw reasonable conclusions.
  - Statistical tools or techniques that are useful for organizing data include:
    1. Frequency Tables
    2. Cross tabulations
    3. Stem and leaf plot
    4. Pie Charts
    5. Bar Charts
    6. Histograms, etc
- Exploratory analysis such as graphs, were done using R Statistical package.
- The variables considered were 938 research questions spanning across
  - Crop
  - Livestock
  - Livelihood

- Questions such as: “how many calories do I eat per day?” or “how much time do I spend talking per day?” can be hard to answer because the answer will vary from day to day. It’s sometimes more sensible to ask “how many calories do I consume on a typical day?” or “on average, how much time do I spend talking per day?”.
- In this section we will study three ways of measuring central tendency in data, **the mean**, **the median** and **the mode**. Each measure has its own strengths and weaknesses.
- A **population** is the collection of all persons, places, or things of interest in a particular study.
- A **sample** is a subset of the population.
- **parameter**: value computed from the population.
- **statistic**: value computed from the sample.

Mean

- The **population mean** of m numbers $x_1, x_2, ..., x_m$(the data for every member of a population of size m) is denoted by μ and is computed as follows:

$$\mu = \frac{x_1 + x_2 + ... + x_m}{m}$$

- The **sample mean** of n numbers $x_1, x_2, ..., x_n$(the data for every member of a population of size n) is denoted by $\bar{x}$ and is computed as follows:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

- **Example:**Consider the following set of data, showing the number of times a sample of 5 students check their e-mail per day: 1, 3, 5, 5, 3.
- Calculate the sample mean $\bar{x}$.

Mean Cont:

$$
\begin{aligned}
\bar{x} &= \frac{1+3+5+5+3}{5} \\
&= \frac{17}{5}
\end{aligned}
$$

- The mean depending on the data can be calculated using frequencies also

The Median

- **The Median** of a set of quantitative data is the middle number when the measurements are arranged in ascending order.
- **To Calculate the Median:** Arrange the n measurements in ascending (or descending) order. We denote the median of the data by M.
  1. If n is odd, M is the middle number.
  2. If n is even, M is the average of the two middle numbers.
- **Example**The number of goals scored by the 32 teams in the 2014 world cup are shown below:
- 18, 15, 12, 11, 10, 8, 7, 7, 6, 6, 6, 5, 5, 5, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 2, 2, 2, 2, 1, 1, 1
- Find the median of the above set of data.Answer: 4

The Mode

- The **mode** of a set of measurements is the most frequently occurring value; it is the value having the highest frequency among the measurements.
- **Example:** What is the mode of the data on the number of goals scored by each team in the world cup of 2006?
- 18, 15, 12, 11, 10, 8, 7, 7, 6, 6, 6, 5, 5, 5, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 2, 2, 2, 2, 1, 1, 1 Answer:4

Measure of Position

$$S_k = \frac{3(mean - median)}{standarddeviation} \\ = \frac{3(\bar{x} - m)}{s},$$

where S_k denotes the skewness, $\bar{x}$ denotes the mean and m denotes the median.

A **Skewed** Data creates an uneven curve distribution.

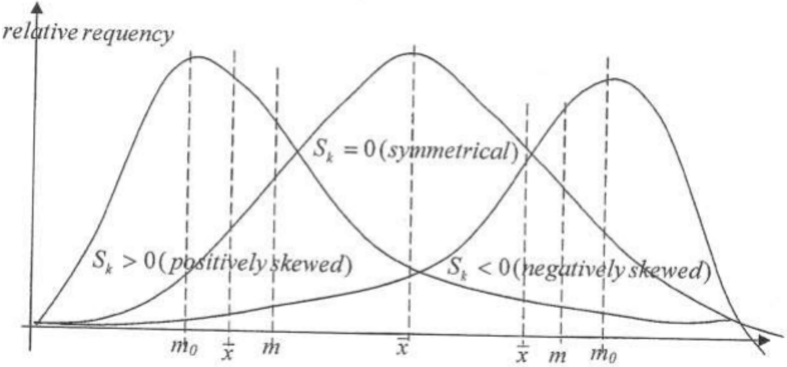
Rightly Skewed: Mean < Median<Mode

Left Skewed: Mean > Median>Mode

If $S_k = 0$, $(\bar{x} = m)$, then the distribution is said to be *symmetric*.

If $S_k > 0$, $(\bar{x} > m)$, then the distribution is said to be *skewed to the right* or *positively skewed*

If $S_k < 0$, $(\bar{x} < m)$, then the distribution is said to be *skewed to the left* or *negatively skewed*



Graph of Symmetrical and Non-Symmetrical Distribution

Discrete & Continuous Probability Distributions

AIMS Ghana

- **A Random variable** is a real-valued function that assigns values to each possible outcome of an experiment.
- The two types are: **Discrete** & **Continuous** random variables.
- **Discrete** random variables take on finite or countably infinite number of values. E.g, the number of Heads obtain from the toss of a coin twice.
- **Continuous** random variables take on infinite or uncountable finite number of values.E.g interval of time an accident occurs.
- **Probability Distribution** is the tabular or functional representation of a random variable and its respective probabilities.

Discrete Probability Distributions

1. Bernoulli Distribution
2. Binomial Distribution
3. Geometric Distribution
4. Negative Binomial(Pascal)Distribution
5. Poisson Distribution
6. Hypergeometric Distribution

Discrete Probability Distributions Cont:

- **The Bernoulli Distribution**
  1. has two possible outcomes(success or failure).
  2. single performance of the experiment.

$$P_X(X=x) = \begin{cases} p^x q^{1-x},\ if\ x = 0, 1 \\ 0, else \end{cases}$$

- **The Binomial Distribution**.
  1. n fixed number of trial
  2. n independent Bernoulli Trials
  3. two possible outcomes
  4. constant success and failure probabilities.

$$P_X(X=x) = \begin{cases} \binom{n}{x} p^x q^{n-x},\ if\ x = 0, 1, ..., n \\ o, else \end{cases}$$

Discrete Probability Distributions Cont:

- **Geometric Distribution**

$$P_X(X = x) = \begin{cases} pq^{x-1}, \; if \quad x = 1, 2, ... \\ 0, else \end{cases}$$

1. E[X] = $\frac{1}{p}$
2. V[X] = $\frac{q}{p^2}$

- **Negative Binomial Distribution**

$$P_X(X = x) = \begin{cases} \binom{x-1}{k-1} p^k q^{x-k}, if \quad x = k, k+1, ... \\ 0, else \end{cases}$$

- E[X] = $\frac{k}{p}$
- V[X] = $\frac{kq}{p^2}$

Discrete Probability Distributions Cont:

- **Poison Distribution**
  1. The occurrence of an event in an interval of time is independent of the occurrence of another event in the same or different interval of time
  2. The probability of the occurrence of an event in an interval is proportional to the length of the interval
  3. For an infinitesimal small portion of an interval the chances of finding more than one event is negligible, in fact 0.
  4. If $X \sim \text{Poi}(\mu)$, then

$$P_X(X = x) = \begin{cases} \frac{e^{\mu}\mu^{x}}{x!}, \mathit{if} \quad x = 0, 1, 2, \ldots \\ 0, \mathit{else} \end{cases}$$

  5. E[X] = μ
  6. V[X] = μ

Continuous Probability Distributions

- **Uniform Distribution**(X∼U(a,b))

$$f_X(x) = \begin{cases} \frac{1}{b-a}, if \quad a \leq x \leq b, \\ 0, else \end{cases}$$

- E[X] = $\frac{a+b}{2}$
- V[X]= $\frac{(b-a)^2}{12}$
- **Exponential Distribution(X$\sim$ exp(λ))**

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, if \quad x \geq 0 \\ 0, \qquad else \end{cases}$$

- E[X] = $\frac{1}{\lambda}$
- V[X] = $\frac{1}{\lambda^2}$

Continuous Probability Distributions Cont:

- **Gamma Distribution(X$\sim \Gamma$(n,λ))**

$$f_X(x) = \begin{cases} \frac{\lambda^n x^{n-1} e^{-\lambda x}}{\Gamma(n)}, if \quad x \geq 0 \\ 0, \qquad else \end{cases}$$

- E[X] = $\frac{n}{\lambda}$
- V[X] = $\frac{n}{\lambda^2}$

Continuous Probability Distributions Cont:

- **Normal Distribution(X $\sim$N(μ,σ))**

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < X < \infty, -\infty < \mu < \infty, \sigma > 0$$

Hypothesis Testing (Test of Significance)

AIMS Ghana

- Standard procedure, based on sample evidence and probability, used to test claims regarding a characteristic of one or more populations.

	True state of affairs (population)	
Decision	The null hypothesis is true H_0 is TRUE	The null hypothesis is false H_0 is not TRUE
We decide to reject the null hypothesis [reject H_0]	Type I error	Correct Decision
We fail to reject the null hypothesis [Do not Reject H_0]	Correct Decision	Type II error

Table: Caption for the table

Hypothesis Testing (Test of Significance Cont:)

- **Null Hypothesis(H_0):**It is the hypothesis we will actually test.
- **Alternative hypothesis(H_A)**: Is the conclusion we will accept if we decide that there is too much evidence against the null hypothesis.
- **Type I error(α)**: We reject the null hypothesis when the null is true.It occurs when you reject a true null hypothesis.

$$\alpha = P(reject \quad H_0|H_0 \quad is \quad true)$$

- **Type II error**: We accept the null hypothesis when it is not true.

$$\beta = P(Fail \quad to \quad reject \quad H_0|H_0 \quad is \quad false)$$

- The two probabilities are inversely related.

Steps In Hypothesis Testing

- Choose the parameter of interest
- Specify/state the Hypothesis
- Specify the level of significance of the test
- Specify the appropriate Test Statistic and its Sampling Distribution
- Determine Statistical Significance
- Compute the Test Statistic
- Make Decision Rules
- Make a Statistical

Hypothesis Testing For Single Population Mean

- **Large Sample ($n \geq 30$) Assumptions**
  1. Population is normally distributed
  2. A random sample of size n is selected
  3. If not normal, requires large samples. When n is large by C.L.T. the population has an approximate normal distribution.
- (a) $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ (Two-Tailed alternative hypothesis)
  1. If the population variance is known, we use z as the test statistic.
  2. **Test Statistic:** $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
  3. If the assumptions are correct and H_0 is true, the test statistic follows the standard normal distribution. Therefore, we calculate a z-score and use it to test the hypothesis.

     Critical region is $\{Z : Z > z_{\frac{\alpha}{2}} \quad or \quad Z < -z_{\frac{\alpha}{2}}\}$

Hypothesis Testing For Single Population Mean Cont:

Decision:

- if observed value of the test statistic is in the critical region: ”$\mathbf{Reject \quad H_0}$”
- If observed value of the test statistic is not in the critical region: ”$\mathbf{Do \quad not \quad Reject \quad H_0}$”

Conclusion: At 100% significance level there is (in)sufficient statistical evidence to ”favor H_1

Exploratory Data Analysis(EDA)

AIMS Ghana

- An analysis approach that identifies general patterns in the data. The steps are;
- **Data Collection:** Gather the dataset you want to analyze. This can be from various sources such as databases, surveys, APIs, etc.
- **Data Cleaning:** Clean the dataset by handling missing values, outliers, and inconsistencies. This step ensures that the data is suitable for analysis.
- **Univariate Analysis:** Explore individual variables in the dataset to understand their distributions, central tendency, spread, and detect outliers. Common techniques include histograms, box plots, and summary statistics.
- **Bivariate Analysis:** Examine the relationship between pairs of variables in the dataset. This can involve scatter plots, correlation analysis, and contingency tables.

- **Multivariate Analysis:** Analyze relationships between multiple variables simultaneously. Techniques such as heatmaps, pair plots, and cluster analysis can be used to identify patterns and correlations among variables.
- **Visualization:** Create visualizations to summarize and present findings from the analysis effectively. This can include bar charts, pie charts, line plots, and more complex visualizations like heatmaps and network diagrams.
- **Statistical Testing:** Conduct hypothesis tests and statistical inference to validate findings and make predictions about the population based on the sample data.
- **Iteration:** EDA is often an iterative process. As you gain insights from the initial analysis, you may go back to earlier steps to refine your understanding or explore new questions.

Histogram

- A histogram summarizes the distribution of the data by placing observations into intervals (also called classes or bins) and counting the number of observations in each interval.

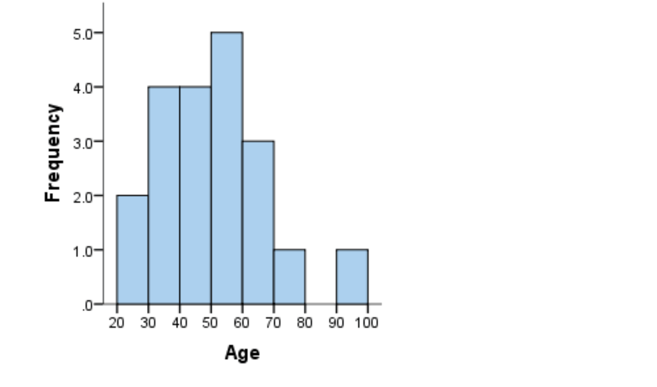


Figure: A histogram

Boxplot

- A box and whisker plot (also referred to as boxplot) provides a compact summary of the distribution of a variable.
- Used for numerical data

Boxplot Cont:

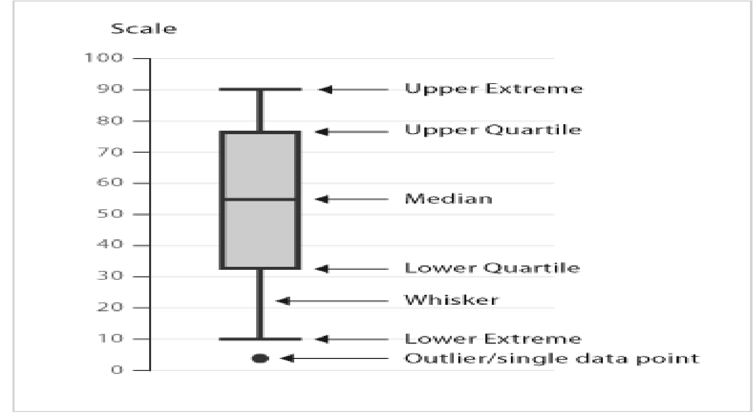


Figure: A Boxplot

Barplot

- A barplot is a type of data visualization that represents categorical data with rectangular bars.
- The length of each bar corresponds to the value it represents.
- Barplots are commonly used to compare the values of different categories or to show the distribution of a single categorical variable.

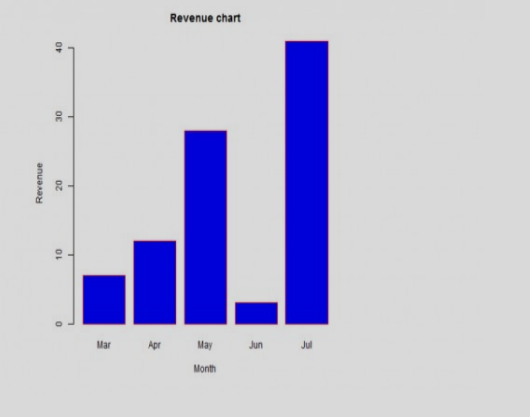


Figure: A Barplot

Pie Chart

- A pie chart is another type of data visualization used to represent categorical data.
- Unlike a barplot, which uses rectangular bars, a pie chart uses a circular shape to represent data.
- The entire circle represents the total sum of the data, and each category is represented by a slice of the pie.

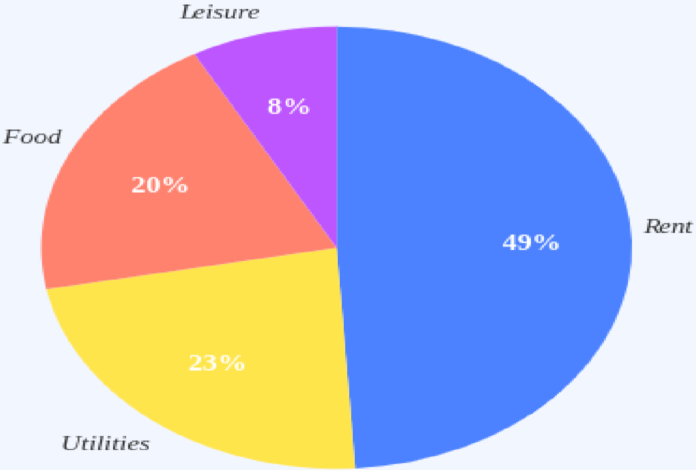
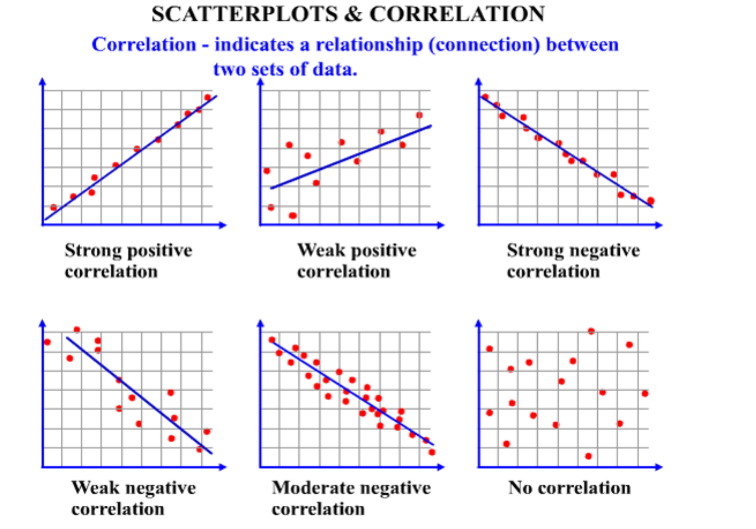


Figure: A Pie Chart

Scatter Plot and Correlation

- A scatter plot is a type of data visualization that is used to display the relationship between two continuous variables.
- Each point on the plot represents an observation in the dataset, with the x-coordinate representing one variable and the y-coordinate representing the other variable.



SCATTERPLOTS & CORRELATION

Correlation - indicates a relationship (connection) between two sets of data.

Strong positive correlation

Weak positive correlation

Strong negative correlation

Weak negative correlation

Moderate negative correlation

No correlation

Figure: Scatter plot and correlation

Thank you for your attention

AIMS Ghana

African Institute for Mathematical Sciences
Ghana

May 19, 2024