Unit II

Distribution and Summary Statistics

What is Distribution and Summary Statistics in Data Analysis?

- Distribution and summary statistics are important concepts in statistics and data analysis.
- They provide a way to describe and understand the characteristics of a dataset.
- They provide a way to quantify and communicate the essential features and characteristics of data

What is distribution in statistics?

- In statistics, a distribution refers to the pattern of the values that a variable takes in a dataset.
- It describes how frequently each value appears and provides insights into the spread and shape of the data.
- Different types of distributions may exhibit various characteristics, such as clustering, symmetry, skewness, or tails.

Types of Distribution

Common types of distributions include:

- **Normal Distribution:** Also known as the Gaussian distribution, it is characterized by a bell-shaped curve and is symmetric around its mean.
- **Uniform Distribution:** All values in the dataset have equal probabilities of occurring, resulting in a flat, rectangular distribution.
- **Exponential Distribution:** Often used to model the time between events in a process, it has a decreasing probability density function.

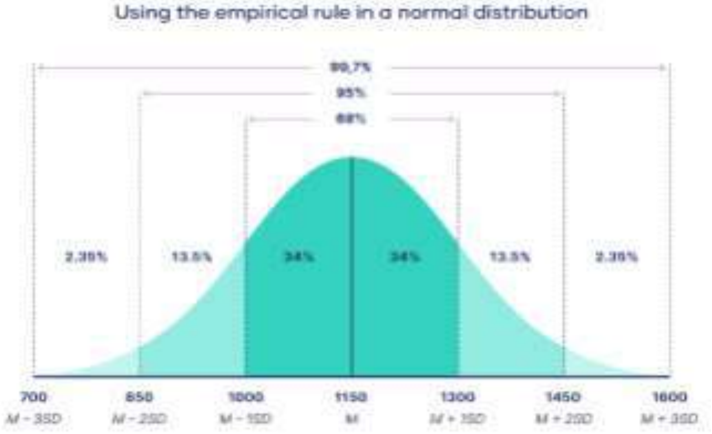
1. Normal Distribution

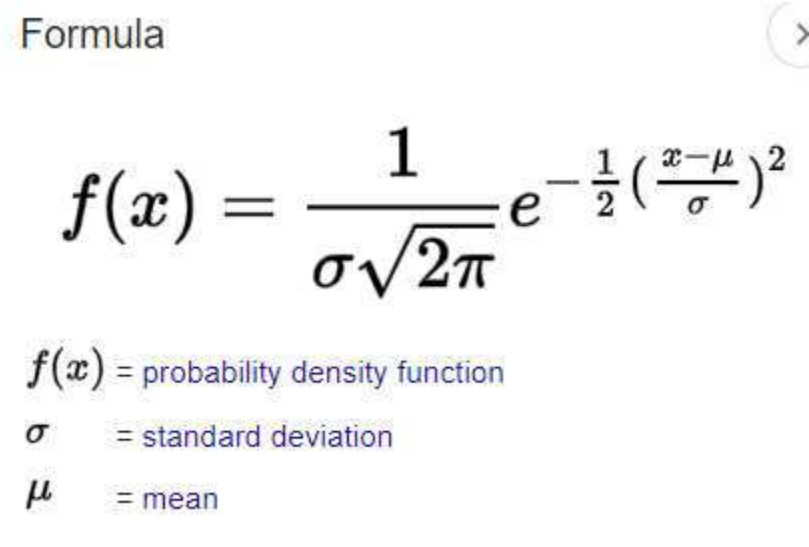
- A normal distribution, also known as a Gaussian distribution, is a fundamental statistical concept that describes the distribution of a continuous random variable.
- **Library:** Inbuilt Libraries
- **Function: rnorm(), dnorm(), pnorm(), qnorm()**

rnorm() - To generate random numbers from a norm distribution.

```
# Generate 100 random numbers from a normal
distribution with mean 0 and SD 1
random_numbers <- rnorm(100, mean = 0, sd = 1)
```

```
x <- 120
y <- dnorm(x, mean = 133, sd = 23.15)
print(y)
```


Using the empirical rule in a normal distribution


Formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$f(x)$ = probability density function

σ = standard deviation

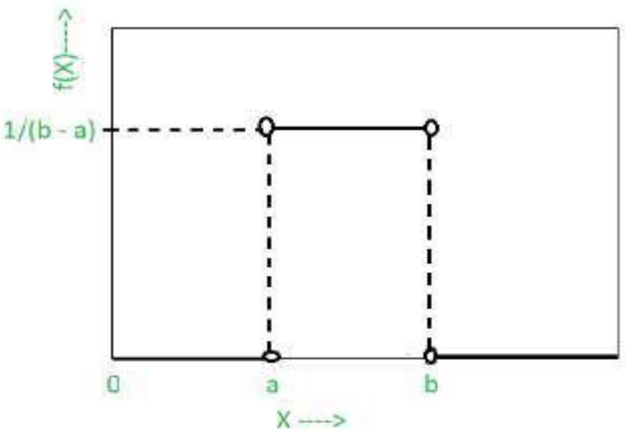
μ = mean

2. Uniform Distribution

- The uniform distribution is a fundamental concept in statistics and data analysis. It's a type of probability distribution that describes a situation where all outcomes in a given range are equally likely..
- **Library:** Inbuilt Libraries
- **Function: runif(), dunif(), punif()**
- **runif()** - To generates random numbers between a specified minimum and maximum value

```
# Generate 10 random numbers from a uniform
distribution between 0 and 1
random_numbers <- runif(10)
```

UNIFORM DISTRIBUTION GRAPH



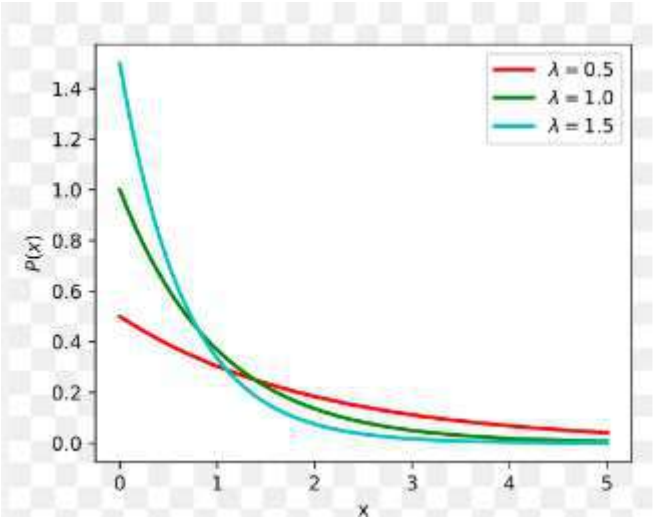
$$\begin{cases} 0 & \text{for } x < a \\ \dfrac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{for } x > b \end{cases}$$

```
pdf_values <- dunif(x, min = 0, max = 1)
pdf_values
```

$\mu = (a + b) / 2$ $\sigma^2 = (b – a)^2 / 12$

3. Exponential Distribution

- The exponential distribution is a probability distribution that models the time between events in a process that occurs randomly and independently at a constant average rate.
- **Library:** extraDistr
- **Function:** rexp(n, rate), dexp(x, rate)



```
# Generate random samples from exponential
distribution
samples <- rexp(n = 100, rate = 0.5)

# Calculate PDF and CDF
pdf <- dexp(samples, rate = 0.5)
print(pdf)
```

$$f(x)=\begin{cases}\lambda e^{-\lambda * x}, x \geq 0\\ 0, x < 0\end{cases}$$

$$\mu = \sigma = 1/\lambda$$

Statistics Summary

What is Statistics Summary in Data Analytics?

- A statistical summary in data analytics provides a concise and informative overview of key aspects of a dataset.
- It helps analysts and decision-makers quickly grasp the main characteristics of the data without having to examine every individual data point.
- A statistical summary typically includes various descriptive statistics that summarize central tendency, variability, distribution, relationships, missing values within the data.

What is Statistics Summary in Data Analytics?

- Here's what a typical statistical summary might include:

    1. Central Tendency
    2. Variability
    3. Percentiles
    4. Relationships
    5. Missing Data
    6. Data Type Information
    7. Outliers

1. Central Tendency

- Mean: The arithmetic average of the data values.
- Median: The middle value when the data is arranged in ascending order.
- Mode: The value that appears most frequently in the dataset.

- **Library:** inbuild

- **Function:** mean(), median(), mode()

```
# find mean
samples <- mean(100, 200)
samples

# mode
a <- c(10, 20, 30)
d <- mode(a)
print(d)

#median
a <- c(10,20,30)
x <- median(a)
print(x)
```

$$\textbf{Arithmetic Mean} = \sum x / N$$

$$\textbf{Median} = (n + 1) / 2$$

2. Variability

- Range: The difference between the maximum and minimum values.
- Standard Deviation: The square root of the variance, indicating the average deviation
- Variance: A measure of how spread out the values are from the mean.

- **Library:** inbuild
- **Function:** range(), var(), sd()

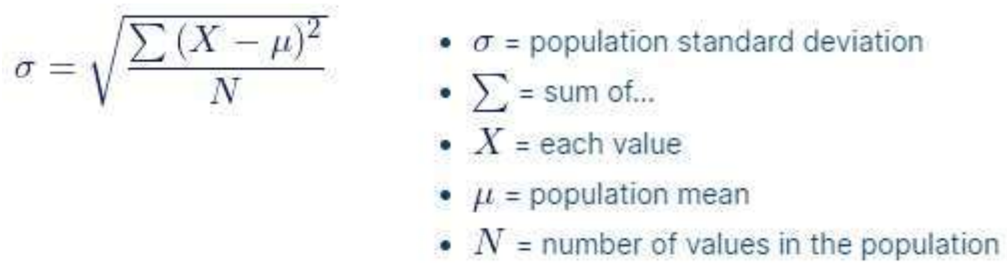
```
# Range
data <- c(12, 45, 67, 23, 56, 89, 34)
data_range <- range(data)
print(data_range)
O/P = 77
```

Range = Highest No. – Lowest No.

```
# Variance
data <- c(12, 45, 67, 23, 56, 89, 34)
data_variance <- var(data)
print(data_variance)
O/P = 702.924
```

Variance = (S.D)2

```
#Standard Deviation
data <- c(12, 45, 67, 23, 56, 89, 34)
data_sd <- sd(data)
print(data_sd)
O/P=26.513
```



$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

- σ = population standard deviation
- $\sum$ = sum of...
- X = each value
- μ = population mean
- N = number of values in the population

3. Percentaile

- In R, percentiles represent points in a dataset below which a given percentage of observations fall.
- They are often used to understand the distribution of data and to summarize the spread of values.
- **Library:** inbuild
- **Function: quantile()**

```
# Percentaile
data <- c(10, 15, 20, 25, 30, 35, 40, 45, 50, 55)
percentile_25 <- quantile(data, probs = 0.25)
cat("25th percentile:", percentile_25, "\n")
```

Percentile = (Number of Values Below “x” / Total Number of Values) × 100

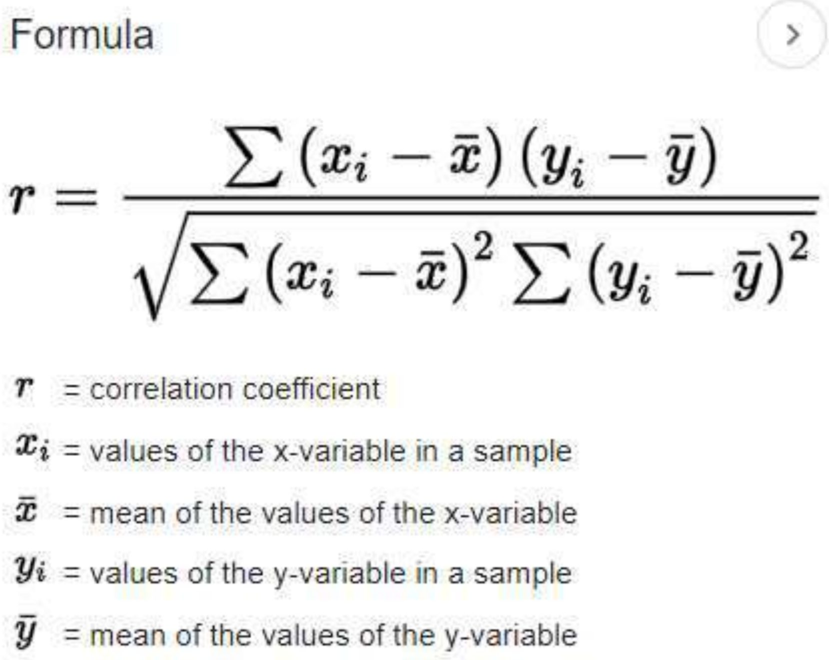
4. Relationship

- Relationship" typically refers to the connection or association between variables in a dataset.
- Understanding relationships between variables is fundamental to gaining insights, making predictions, and drawing conclusions from data.

Correlation:

- Correlation is the one of the type of relationship in data analytics
- Correlation measures the strength and direction of a linear relationship between two continuous variables.
- The correlation coefficient, often denoted as "r," ranges from -1 to 1.
- A positive value indicates a positive correlation and a negative value indicates a negative correlation
- **Library:** inbuild
- **Function: cor()**

```
# Correlation
data <- data.frame(x = c(10, 15, 20, 25, 30),
  y = c(20, 30, 25, 40, 35))
cor_matrix <- cor(data)
print(cor_matrix)
```



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable