\*\*\*Assaignment :- 1 \*\*\*

*USE CASE:*

-summarize data (Avg, Mean…)

-Data assumption

-- To find out what will happen

-- To spot a pattern

--to take a decision with using data

DIKW pyramid:-

**DATA** -

-unorganized, raw level data

--no context, no meaning

**INFORMATION** -

-Organize my data

-to convert into meaningful context, report, tables

**KNOWLEDGE -**

-interpret knowledge & find out patterns and logic

**WISDOM-**

-actionable insights or apply knowledge

Types of statistics

1. Descriptive Stat. - organize and summarized of data (mean, median, charts…)
2. Inferential Stat. - use sample to generalize large scale data.

*Inferential statics -->*

QUANTITATIVE : (Numerical data) - countable

-->continues (decimal value)

-->Discrete (whole no.)

QUALITATIVE: (Categorical Data)

#scales of measurement

-it define how the data is classified, compared.

Properties:

* Identify (label) -
  + Nominal -Mode, frequency ,percentage, graph: bar, pie, chi-square

* Order (ranking) -
  + Ordinal - median, percentile graph: bar, u-TEST

* Distribute in Equal interval -
  + Interval Level- mean, median , Std. Deviation , Histogram, box plot, Z-test, T-Test, an nova

**(Ratio)** - Line plot, correlation

Population - whole dataset

* A population is the entire group

 Sample - subset of dataset

*RANDOM VARIABLE:-*

It is a variable whose values is depend on outcomes of a random experiment.

-it assign numerical values to outcomes of events.

-it connects real world process to statistics analysis.

Example:

->tossing a coin - event

Head - 0

Tail - 1

Events = X

-->TYPES OF RANDOM VARIABLE

|  |  |  |
| --- | --- | --- |
| Features | DISCRETE | CONTINUOUS |
| Value | Countable (1, 23, 553…) | Real value (interval), decimal (9.54, 63.20, 5.8….) |
| * 1. Representation | Mass Function | Density Function |
| Examples | No. of sales | Revenue of that sale |

Discrete RV- count, clicks, yes/no…

Continuous RV- revenue, response time, stock market, sensor reading….

MEAN, MEDIAN, MODE

MEAN

* ARITHMETIC AVERAGE OF ALL VALUES IN GIVEN DATASET
* IT TELLS BALANCE P[OINT OF DATA

MEAN = x1+xx2+3+……. + xn / n

Marks: 50, 64, 93, 81

N = 4

Mean = (50+64+93+81)/4=72

Median - middle value of dataset

(Rearrange in descending manner)

N is odd -> middle number

N is even -> average of two middle number

Median = 50, 64, 81, 93 = (64+81)/2=72.5

Mode

Most repeated value in dataset

Marks: 50, 64, 93, 81, 50

Mode = 50

Center of tendency

* It define "center" of dataset.
* They provide summary of the data with a single value.\

(Mean / median/ mode)

**Outlier: - A** data point(s) who does not follow patterns or trend of the data

**MEASURE OF DISPERSION:-**

[5, 1, 2, 1, 1, 2] = 12/6=2

[2, 2, 2, 2, 2, 2] = 12/6=2

IT DESCRIBES SPREAD OR VARIABILITY OF DATASET.

THEY INDICATE HOW MUCH INDIVIDIUAL DATA POINTS DIFFER FROM THE CENTRAL VALUE.

->VARIANCE -

IT MEASURES HOW FAR EACH DATA POINTS IN THE DATASET IS FROM MEAN VALUE.

HIGHER VARIANCE - MORE SPREAD

LESSER VARIANCE - LESS SPREAD

**Variance (σ²)**

For population:

σ2=∑ (X−Xˉ) 2N\sigma^2 = \frac{\sum (X - \bar{X}) ^2} {N} σ2=N∑ (X−Xˉ) 2​

For sample:

s2=∑ (X−Xˉ) 2n−1s^2 = \frac{\sum (X - \bar{X}) ^2} {n-1} s2=n−1∑ (X−Xˉ) 2​

[5,1,2,1,1,2] = 12/6=2

|  |  |
| --- | --- |
| MEAN(X) - Individual Dt. Pts.( | Ans. |
| 5-2=3 | 9 (Xi- MEW) |
| 1-2=-1 | 1 |
| 2-2=0 | 0 |
| 1-2=-1 | 1 |
| 1-2=-1 | 1 |
| 2-2=0 | 0 |

= 12 /6=2

RANGE => Max. - Min. =9-0=9

Marks = [56, 62, 78, 85, 38, 43]

= (55+65+75+85+35+45) /6 = 60

|  |  |
| --- | --- |
| 55-60=-5 | 25 |
| 65-60=5 | 25 |
| 75-60=15 | 225 |
| 85-60=25 | 625 |
| 35-60=-25 | 625 |
| 45-60=-15 | 225 |

= 25+25+225+625+625+225=1750/6= 291.67

Range = 625-25=600

-->Standard Deviation = SD

Equation:-

**Standard Deviation (σ)**

σ=∑ (X−Xˉ) 2N\sigma = \sqrt{\frac{\sum (X - \bar{X}) ^2} {N}} σ=N∑ (X−Xˉ) 2​​

**CONTINUOUS DISTRIBUTION:-**

-UNIFORM

-NORMAL

-Std. NORMAL

-EXPONENTIA-CHI-SQUARE

1. Uniform Distribution (a, b)

All values I [a, b] equally likely

Mean = a+ b/2

Variance= (b-a). (b-a)/12

Example:

Bus arrives uniformly between 0 and 20 after a hour, P (5<= X<=12):

P= 12-5 / 20-0

= 7/20

=0.35

Mean = 0+20 / 2

= 10 min

Variance = (20-0). (20-0) / 12

= 400/12

= 33.34

* 35% Chances bus let 5 to 12 min.
* Avg time = 10 min wait

1. NORMAL DISTRIBUTION

* AVG. HEIGHT (MEAN) = 170cm
* Probability height between (160-180)
* S.D.= 10cm

Z1 = 160-170 / 10

= - 1

Z2 = 180-170

= 1

Value:-

1 = 0.8413

-1 = 0.1587

(-1<Z<1)= 0.84-0.15

= 0.69

\*69% Probability

1. Standard normal distribution:- (Z-Score)

 Used in **hypothesis testing** and **confidence intervals**.

Ex: Test (100)

N = (15)

Score = 130

Z= 130-100 / 15

= 2

= (0.9772)

1. Exponential Distribution: - The **Exponential distribution** is a **continuous probability distribution** used to model the **time between events** in a Poisson process (events happening independently at a constant average rate).

F (λ) = {(e) −λ\*10}

What is the probability that the next bus arrives **within 5 minutes**?

P (X ≤ x) =1−e−λxP(X \ leq x) = 1 - e^ {-\lambda x} P(X ≤ x) =1−e−λx

P (X≤5) =1−e−0.1×5P(X \leq 5) = 1 - e^ {-0.1 \times 5} P (X≤5) =1−e−0.1×5 =1−e−0.5= 1 - e^ {-0.5} =1−e−0.5 =1−0.6065=0.3935= 1 - 0.6065 = 0.3935=1−0.6065=0.3935

**39.35% chance** that the bus arrives within 5 minutes.

\*\*\*Discrete Distribution\*\*\*

Type:-

* 1 Uniform Distribution :- Every Outcome Equally Likely

Ex: Roll no. 1/6= 0.1667

(1+2+3+4+5+6)/6= 3.5

* 2 Bernoulli: - Only Two Outcome Possible.

Ex: Success (1)

Failure (0) Probability (P) = Failure (0)

= (1-P)

Success = 7

Failure =? 1-0.7= 0.3 Variance (V) = P (1-P)

=0.7(1-0.7)

=0.7(0.3)

=0.21

* 3 Geometric Distribution: - Event not get first success until that repeat = 1F Success Probability=P.

P(x=k) = (1-P) ^ K-1\*P

X=variance

K=repeated value

Ex: P= 0.16, K=3

P(x=3) = (1-0.16) ^2\*(0.16)

= (0.84) ^2\*(0.16)

= (1.68) \* (0.16)

= 0.1129

Percentage = 11.2%

* 4 Poisson distribution: - -Given time interval.

-Given number of fixed interval.

-Event that happen independently at constant average (mean).

Equation: P(X=k) =k! e –λ λk​

Ex: - 3 customer arrives per hour find probability exactly 2 customer. Mean=3

K=2

E= 2.718

= (2.718). (2.718). (2.718). (3). (3) / 2\*1

= (20.07). (9)/2

= (180.71)/2

= 90.35

**\*\*\*TYPES OF SAMPLING\*\*\***

* 1 Random sampling: - **Random sampling** is a method of selecting a sample from a population **in such a way that every individual/item has an equal chance of being chosen**.
* 2 Stratified sampling: - Divided into group, take sample of each group.
* 3 Cluster sampling: - Divided by same properties and select one of them.
* 4 Systematic sampling: - **Systematic sampling** is a probability sampling method where you select every **kth element** from a population list, after choosing a random starting point.

\*Statistical inference:-

* Estimation: - Confidence interval sample mean.
* Hypothesis: - A **hypothesis** is a statement or assumption about a population parameter (like mean, proportion, and variance) that we want to test using sample data.