Introduction To Statistics

M T W T F S S
Page No.:
Date:
YOUVA

* What is Stats ?

→ Stats is a science of Coulecting, analyzing and organizing data.

* What is Data ?

→ Facts or Pieces of information that can be measured.

* Types of Stats :

→ Descriptive Stats :-
It Consist of organization & organization of data.

→ Inferential Stats :-
Techniques where used data that we have measured to form the condisions.

* Sampling Techniques :

→ Simple Random Sampling :-
Every membes of Population (N) have an equal chance of getting selected in sample (n).

→ Stratified Sampling :-
Where we split the Population [N] into non-overlapping groups [strata].

→ Systematic Sampling :-
Surveying on N^{th} individual

→ Convenience Sampling :-
Surveying only those people who have knowledge of that particular domain.

* Variables :

→ A variable is a property that can take any values.

* Types of Variables :

→ Qualitative / Categorical Variables :

Nominal :- Categories with no order. [e.g., colors, gender]

Ordinal :- Categories with natural order. [e.g, education level]

→ Quantitative / Numerical Variables :

Discrete :- Countable values [e.g., number of children]

Continuous :- Any values within a range. [e.g., height, weight]

* Measure of Central tendency :

→ Mean [Average] :-

Formula :- $\bar{x} = \Sigma x / n$

Sum of values divided by count

Affected by outliers

→ Median :-

Middle value when data is ordered

Formula for position :- $[n+1]/2$

Not affected by outliers

Best for skewed data

→ Mode :-

Most frequent value

Can have multiple modes

only measure for categorical data

* Variance Definition :-

→ The average squared deviation of values from their arithmetic mean.

Formulas :

I. Population Variance [σ^2] :-

$$\sigma^2 = \frac{\Sigma[x - \mu]^2}{n}$$

x = individual values

μ = Population mean

N = Population size

II. Sample variance [s^2] :-

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

$\bar{x}$ = Sample mean

n = Sample size

$(n-1)$ = degree of freedom

Properties :-

Always non-negative

Unit is squared

Sensitive to outliers

* Standard Deviation :-

→ Square root of variance; measure average deviation from mean.

⇝ Formulas :-

I. Population :-

$$\sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{N}}$$

II. Sample :-

$$s = \sqrt{\frac{\Sigma (x-\bar{x})^2}{(n-1)}}$$

⇝ Key Point Variance & Standard

→ Variance gives you the average squared distance from the mean.

→ Standard Deviation gives you a measure of squared in the same unit.

* Five number summary :

→ Minimum : Smallest value

→ Q1 [First Quartile] : 25^{th} Percentile

→ Median : 50^{th} Percentile

→ Q3 [Third Quartile] : 75^{th} Percentile

→ Maximum : Largest values

⇝ Calculations :-

IQR = Q3 - Q1

Lower fence = Q1 - 1.5 [IQR]

Higher fence = Q3 + 1.5 [IQR]

* Data Distribution :-

→ Pattern of how data values are spred and their frequency of occurrence.

~> Key Point / Component :-

→ Center

→ Spread

→ Shape

→ outliers

* Normal distribution :-

→ The normal distribution is a Probability distribution that occurs naturally in many Phenomena.

~> Key characteristics :-

→ Bell-Shaped, Symmetric around the mean

→ 68% of data within 1 standard deviation

→ 95% of data within 2 standard deviation

→ 99.7% of data within 3 standard deviation

~> Applications :

→ Natural Phenomena [height, weight, IQ]

→ Financial markets

→ Quality control

→ Statistical inference

* Z-score :-

→ Measures distance from mean in standard deviations.

→ Formula:

$$z = \frac{x - \mu}{\sigma}$$

→ Used for comparing values from different distributions

→ Common in statistics, education [test score], Finance.

* Standard Normal Distribution :

→ Mean = 0, Standard deviation = 1

→ Used as reference distribution

→ Enables Probability calculations using z-table

* Normalization :

→ Methods :- Min - max, z-score, decimal scaling

→ Min-Max formula :- $\frac{x - min}{x_{max} - x_{min}}$

→ Used in machine learning, data PreProcessing

→ Puts variables on same scale

* Skewed Distributions :

⇝ Positive skew :

→ Mean > median

→ Long tail right

→ Example :- income distribution

⇝ Negative skew :

→ Mean < median

→ long tail left

→ Example :- exam scores

* Continuous Distribution :

-> Infinite Possible values

-> Example :- height, time

-> Uses integration for Probability.

* Discrete Distribution :

-> Countable values

-> Example :- dice rolls, counts

-> Uses Summation for Probability

* Uniform Distribution :

-> Every outcome equally likely

-> Formula :- $f(x) = \frac{1}{b-a}$

-> Mean :- $\frac{a+b}{2}$

-> Variance :- $\frac{(b-a)^2}{12}$

→ Example :- dice rolls, random number generators

* Exponential Distribution :

→ Models time between events

→ Formula :- $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$

→ Mean : $1/\lambda$

→ Variance : $1/\lambda^2$

→ Memoryless Property

→ Example :- equipment failure times, customer arrival s.

* Bernoulli Distribution :

→ Single trial, success / failure

→ $P(X=1) = p$, $P(X=0) = 1-p$

→ Mean : p

→ Variance : $p(1-p)$

→ Example :- Single coin flip

* Binomial Distribution :

-> n independent Bernoulli trials

-> Formula :-

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

-> Mean :- ~~np~~ np

-> Variance :- np (1-p)

-> Example :- number of heads in n coin flips

* Poisson Distribution :

-> Rare events in fixed interval

-> Formula :-

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

-> Mean = Variance = λ

-> Example :- Website visits per hour

* Probability :

→ P(E) = favorable outcomes / total outcomes

→ $0 \leq P(E) \leq 1$

→ P(complement) = 1 - P(E)

* Hypothesis Testing :-

→ A statical hypothesis test is a method of statistical inference used to decide whether the data at hand is sufficient to support a particular hypothesis.

→ Hypothesis testing allows us to make probabilistic statement about population parameters.

* Null Hypothesis [H_0] :-

→ The null hypothesis assums that there is no significant relationship or effect b/w two variable.

→ It serves as a starting point for HT & represents 'static quo' or the assumption of no effect until proven otherwise.

* Alternate Hypothesis[H_0 or H_1] :-

→ It is a statement that contradicts the NH & claims there is significance effect or relation.

ship.

* Rejection Region method :-

1. Ho & Ha
2. α → value
3. assumptions
4. decide test → Z-test, T-test
5. value
6. Test conduct
7. Reject / Accept
8. State results

* T-test :-

→ It is determine significance changes between means of two groups.

→ 3 types :-

1. one sample t-test :-

→ Compares the mean of a single sample to a known μ.

2. Independent Two sample t-test :-

→ The samples from two independent groups are compared to determine if the means of the associated populations are significantly different.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

3. Paired t-Test :-

→ Compares the means of two related groups.

$$t = \frac{\bar{d}}{sd/\sqrt{n}}$$

$\bar{d}$ = 2 means difference

sd = Std difference.

* chi-square test :-

→ It is a Statical procedure for determing the difference between observed and expected data.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O = observed frequency

E = Expected frequency

$$E = \frac{\text{Row total} \times \text{column total}}{\text{Grand total}}$$

$$df = (r-1) \times (c-1)$$

r = no of rows

c = no of columns

* Kurtosis :-

→ Kurtosis is a measure of the tailedness of a distribution.

→ 3 Types of Kurtosis

1. Mesokurtic :-

-> Tails are similar to ND

-> distribution with kurtosis = 3

Eg. Standard Normal distribution

2. Leptokurtic :-

-> Heavy tails [with more / extreme outliers]

-> distribution with kurtosis > 3

Eg. t-distribution with very small df.

3. Platy kurtic :-

-> Light tails [fewer extreme outliers].

-> distribution with kurtosis < 3

Eg. Uniform dest