Statistics Module Cheat File

08 November 2022 10:01

Descriptive stat:- Analyzing the data / finding the patterns
those patterns are useful to present data

MCT:-

(a) mean :- $\Sigma x/n = \bar{x}/\mu$
(b) median : middle point
(c) mode : most repeated value in data

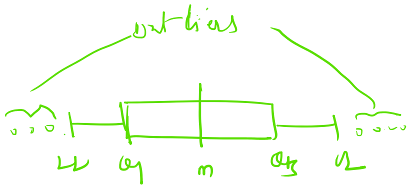
→ imputations of Null value
→ To show / identify the distribution of column

MD:-

(a) std : $\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$ (distance b/w sample point to mean)

(b) var : $\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n}$

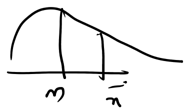
Outliers: inter Quartile Range (IQR)



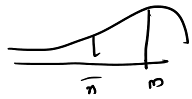
$LL = Q_1 - 1.5 * IQR$
$UL = Q_3 + 1.5 * IQR$
$IQR = Q_3 - Q_1$

measures of shape:

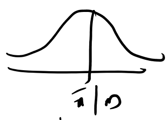
(1) mean > med → positive skewness
$sk > 0$



(2) mean < med → negative skewness
$sk < 0$

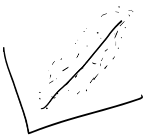


(3) mean ≅ med → skew symmetric
$sk \cong 0$



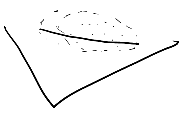
Note:- mean is affected by the outliers but not median

measures of relation: Correlation: Relation b/w two columns / variables



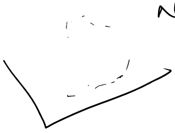
positive correlation = ↑ ↑ / ↓ ↓ → direction same

$r \geq 0.5$ $\qquad r = \frac{cov(x,y)}{\sigma_x * \sigma_y}$



Negative → ↑ ↓ / ↓ ↑ → direction is opposite

$r \leq -0.5$



No-Correlation → $-0.5 < r < 0.5$

Probability:- chance of occurrence / non-occurrence

$P(E) = \frac{\text{no of favourable outcome}}{\text{total no of outcomes}}$

A, B are two events

$P(A \text{ and } B) = P(A \cap B)$ → multiplication theorem

$P(A \text{ or } B) = P(A \cup B)$ → Addition theorem

Types of probability:-

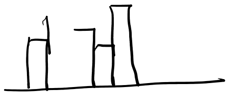
1. marginal probability : $P(A)$, $P(B)$
2. Joint probability : $P(A \text{ and } B)$
3. Conditional Probability : $P(A|B)$, $P(B|A)$

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

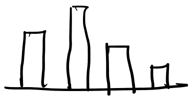
Extensions of conditional is Bayes's theorem

Measurement of statistics :-

→ Quantitative data
↓
Numerical variable
Ratio   interval
Continuous RV
↓
which can measurable
Eg:- height, weight, C°, salary, age



Qualitative Data
↓
categorical data
Nominal   ordinal
Discrete RV
↓
which can countable
no of student, cities
Gender, grading, colombs, ratings



Distributions :-

Binomial:
1. n independent trails | Experiments
2. each trail are having same outcomes

$$P(\text{success}\ (x)) = {}^{n}C_{x} * p^{x} * q^{n-x}$$

n = no of trails
p = P(success)
q = P(failure)
x = no of success

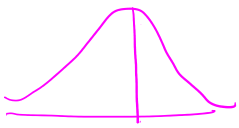
mean = np
var = $n * p * q$

Poisson:
1. Avg value in a certain time interval

$$P(x) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}, \quad \lambda = \text{Average}$$

mean = λ
var = λ

Normal Distn:
1. mean and median both are coincide / equal
2. It has symmetric curve / Bell
3. Gaussian - distn
4. ND is following Empirical Rule
   - 68.45% data lies in between $\mu \pm 1\sigma$



$\bar{x}/\mu$

Z-score
$\mu \pm 2\sigma$, and

Z-score
$\mu \pm 2\sigma$
$\mu - 2\sigma <$ out
$\mu + 2\sigma >$ out

(4) ND is following ...

- 68.45% data lies in the $\mu \pm 1\sigma$
- 95% data lies in the $\mu \pm 2\sigma$
- 99.73% data . . $\mu \pm 3\sigma$

$$Z = \frac{x - \mu}{\sigma}$$ (Standard scalar)

-3 to $+3$, mean $= 0$
$\sigma = 1$

CLT:-

① mean of sample mean is always approximately equal to population mean

$$E(\bar{x}) = \mu$$

② Histogram of all the sample mean is follows Normal distribution

Sampling Techniques:-

1. Simple Random Sampling
2. Stratified "
3. Systematic "
4. Cluster Sampling

Hypothesis Testing, statement about population

Step 1: Define statement
① Null (Ho) — Homogeneous state.
② Alternative (H1) — Heterogeneous state.

Step 2: Define los (or) α%.

$\alpha = P$ (Rejecting Ho | Ho True)
Type-I error)

β = Type-II error
P (Accepting Ho | Ho False)

Step 3: Statistical tests t | z | χ^2-test

Step 4: P value = evidence about to Reject Ho

$P \leq \alpha \rightarrow$ Reject Ho
$P > \alpha \rightarrow$ Failed to Reject Ho