CLASICAL PROBABILITY

→ states the possible outcome of any event. It predicts the likeliness of an event happening.

$$P(E) = \frac{\text{no. of favourable outcomes}}{\text{tot. no. of exhaustive events.}}$$

→ if a random exp. or trial results of n exhaustive mutually exclusive & equally likely outcomes i out of which m are favourable to the occurance of an event E. $P(E) = m/n$

LIMITATIONS:
- various outcomes of the exp. are not likely. (coin toss, die roll)
- only works well in simple well defined experiments.
- doesn't use actual data or obs., thus less reliable.

STATISTICAL PROBABILITY

→ statistical represent^n of any random event involves data collectn, analysis and interpratatn
→ if a trial is repeated n no. of times under essentially, identical conditions then $P(E) = \lim_{n \to \infty} \frac{m}{n}$.

LIMITATIONS:
- requires large sample size, a no. of trials is done.
- no unique events, repeated trials may occur
- relies heavily on past data, which may not reflect future outcomes.

Additive law of probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$$

Multiplicative law of probability

$$P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$$

Total probability theorem

$$P(A) = \sum P(E_j) \cdot P(A|E_j)$$

Conditional probability

$$P(A|B) = P(A \cap B) / P(B)$$

Bayes' theorem

$$P(E_i|A) = \frac{P(E_i) \cdot P(A|E_i)}{\sum P(E_i) \cdot P(A|E_i)} = \frac{P(E_i) \cdot P(A|E_i)}{P(A)}$$

Measures of Central Tendency

MEAN — avg. of set of values

Sample mean: $\bar{X} = \frac{\sum x_i}{N}$ (subset of a population)

Population mean: $\mu = \frac{1}{N} \sum x_i$ (for complete population)

Grouped data: $\frac{\sum f_i x_i}{\sum f_i}$

MEDIAN — middle value when in sorted data

Ungrouped data:

$$Md \begin{cases} \frac{(N/2)^{th} \text{ obs.} + (N/2+1)^{th} \text{ obs}}{2}, & N \in \text{even} \\ \left(\frac{N+1}{2}\right)^{th} \text{ obs.}, & N \in \text{odd} \end{cases}$$

Grouped data:

$$Md = l + \left(\frac{N/2 - C_f}{f}\right) \times h$$

- Median class: class in which $N/2$ lies before cumulative freq.
- l: lower limit, f: median class freq., h: width of class
- C_f: cumulative freq. of preceding class.

MODE most frequently occ. value

Ungrouped: (counting)

Grouped: $M_0 = l + \frac{(f_1 - f_0)}{(2f_1 - f_0 - f_2)} \times h$

Modal class: class with highest frequency

l: lower limit — f_1: freq. of modal class

h: width of class — f_0: freq. of ~~modal~~ preceeding class

f_2: freq. of succeeding class.

MODE = 3 MEDIAN − 2 MEAN

COMBINED MEAN $\bar{\bar{X}} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$

Measures of dispersion

Absolute dispersion → expressed in terms of original unit

RANGE: diff. b/w max and minimum values

RANGE = (max value) − (min value)

VARIANCE dispersion of a set of values

$\sigma^2 = \frac{\sum (x_i - \bar{X})^2}{N}$: population variance

$S^2 = \frac{\sum (x_i - \bar{X})^2}{n-1}$: sample variance

$n-1$ → (estimate of the populatn) → BESSEL'S CORRECTION

STANDARD DEVIATION measure of spread of data

$S.D = \sqrt{\sigma^2}$

$S.D\ \sigma = \sqrt{\frac{\sum (x_i - \bar{X})^2}{N}}$: population variance

$\sigma = \sqrt{\frac{\sum (x_i - \bar{X})^2}{n-1}}$: sample variance

MEAN DEVIATION ∞ ~~PARTITION~~

avg. dist blw each data point and mean of dataset

$M.D = \frac{\sum f_i |x_i - A|}{N}$: grouped data

$M.D = \frac{\sum |x_i - \bar{x}|}{N}$: sample $\quad M.D = \frac{\sum |x_i - \mu|}{N}$: populatn

COMBINED VARIANCE $= \sigma_{12}^2 = \frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{(n_1 + n_2)}$

$d_i = \bar{x}_i - \bar{\bar{x}}$ → combined mean

↓ individual mean

COEFFICIENT OF VARIATN standr measure of dispersn relative to mean

$CV = \frac{\sigma}{\mu} \times 100$ (%) : populatn

$CV = \frac{S}{\bar{x}} \times 100$ (%) : sample

(finding variability)

- **PARTITION VALUES** : divide the series into no. of equal parts.

Quartile $Q_i = iN/4$ th

Decile $D_i = iN/10$ th

Percentile $P_i = iN/100$ th

for grouped data : $Q_i = l + \frac{iN/4 - C_f}{f} \times h$

$D_i = l + \frac{iN/10 - C_f}{f} \times h$

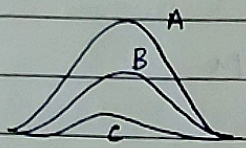
$P_i = l + \frac{iP_i/100 - C_f}{f} \times h$

ABSOLUTE DISPERSION

(disp. is exp. in terms of original units)

distance measure	spread of dispersion
- Range	- mean deviation
- Interquatile Range ($Q_3 - Q_1$)	- Standard deviation

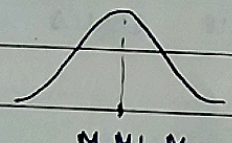
Kurtosis : measure of peakness of data distribution.



A: LEPTOKURTIC	$\beta_2 > 3$	$\gamma_2 > 0$
B: MESOKURTIC	$\beta_2 = 3$	$\gamma_2 = 0$
C: PLATYKURTIC	$\beta_2 < 3$	$\gamma_2 < 0$

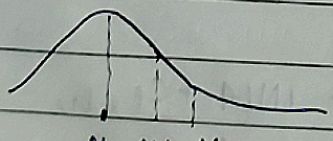
- $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, $\gamma_1 = \sqrt{\beta_1}$ $\quad \beta_2 = \frac{\mu_4}{\mu_2^2}$, $\gamma_2 = \beta_2 - 3$

Skewness : measure of symmetry $\quad S_k = \frac{3(M - M_d)}{\sigma}$ or $\frac{3(M - M_o)}{\sigma}$



Symmetric $\quad \beta_1 = 0 \quad \gamma_1 = 0$

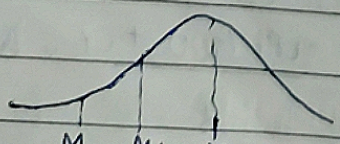
$M = M_d = M_o$



Right skewed $\quad \beta_1 > 0$

+ve skewed

$M > M_d > M_o$



Left skewed $\quad \beta_1 < 0$

-ve skewed

$M < M_d < M_o$

KARL PEARSON'S SKEWNESS COEFF = mean - mode or mean - median

BOULEY'S SKEWNES COEFF = $\frac{Q_3 + Q_1 - 2\text{median}}{Q_3 - Q_1}$

Page No.
Date

Moments

moments are quantitative measures related to the shape of dataset's distribution / spread and central tendency.

RAW MOMENTS (Moments about origin)

→ measures values of random variable relative to the origin. used to describe overall shape of distribution

$$M_r' = E[x-A]^r = 1/N \cdot \sum_{i=1}^{N} [x_i - A]^r$$ (A: arbitary point)

M_r' (about origin) $= E[X^r]$

M_1' (first row moment) $= \bar{X}$ (mean)

$M_2' = E(X^2) = E[(x-A)^2]$

$M_3' = E(X^3) = E[(x-A)^3]$ ....

CENTRAL MOMENTS

→ calculated with respect to the mean.
used to understand the shape and variability of dataset

$$M_r = E[(X-\bar{X})^k] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^k$$

M_1 (first central moment) $= 0$ (zero)

M_2 (second " ") $= E[(X-\bar{X})^2] = \frac{1}{N}\sum(x_i - \bar{X})^2 = \sigma^2$ (variance)

- RELATION BTW RAW MOMENTS & CENTRAL MOMENTS

$$M_2 = M_2' - (M_1')^2$$

$$M_3 = M_3' - 3M_2'M_1' + 2M_1'^3$$

$$M_4 = M_4' - 4M_3'M_1' + 6M_2'(M_1')^2 - 3(M_1')^4$$

First (raw or central moments) : mean
Second (central moment : variance
Third central moment : skewness
Fourth central moment : kurtosis

SKEWNESS

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, $\gamma_1 = \sqrt{\beta_1}$

KURTOSIS

$\beta_2 = \frac{\mu_4}{\mu_2^2}$, $\gamma_2 = \beta_2 - 3$

Random Variable

- variable whose values depend on the outcome of a random phenomenon.

Discrete random variable : have finite distinct possible values. (PMF probability functn)

Continuous random variable : ~~have~~ can take any value within a certain range. (PDF probability functn)

Probability mass function : it is the prob. of a discrete random variable, it must satisfy the foll. condn : 1. $P(X) \geq 0$ 2. $\sum P(X) = 1$

Probability Density function : it is the prob. functn of a contin^n random variable, it must satisfy the foll. condn

1. $f(x_i) \geq 0$ 2. $\int_{-\infty}^{\infty} f(x) \cdot dx = 1$

Cumulative Distribution function : it describes the prob. that a random variable takes on a value less than or equal to a given point. $F(x) = P(X \leq x)$

→ both discrete and ~~random~~ contⁿ random vars.

1. $0 \leq F(x) \leq 1$ 2. $F(x) \leq F(y)$ if $x \leq y$

Expectation (expected value) of random var

measure of its avg or mean value, based on its probability distⁿ. It provides the weighted avg. of all possible values ~~of~~ random var can take

(DISCRETE RANDOM VAR) $E(X) = \sum_x x_i \cdot f(x_i)$

(CONTⁿ RANDOM VAR) $E(X) = \int_{-\infty}^{\infty} x \cdot f(x_i) \cdot dx$

Properties

1. $E(X+Y) = E(X) + E(Y)$
2. $E(XY) = E(X) \cdot E(Y)$ X, Y are indepen
3. $E[a\,f(x)] = a\,E[f(x)]$
4. $E[af(x)+b] = a\,E[f(x)] + b$

↳ $E(b) = b$

Variance

- measures the spread or dispersion of possible values around the expected value.

$$Var(X) = E(X^2) - (E(X))^2$$

$$= \sum x^2 \cdot f(x) - (\sum x \cdot f(x))^2 \quad : \text{discrete}$$

or

$$= \int x^2 \cdot f(x) - (\int x \cdot f(x))^2 \quad : \text{continuous}$$

Standard Deviatⁿ

- measure of the avg. devⁿ from the mean and is expressed in same units as random var.

$$\sigma = \sqrt{Var(X)}$$

Properties

1. $V(aX) = a^2 V(X)$
2. $V(X+b) = V(X) + 0$
3. $V(aX+b) = a^2 V(X) + 0$

Bivariate Random Variable

involves 2 random variable X and Y, defined on the same probability space.

~~Measures of Association~~

Joint Probability distribution: joint dist^n of 2 random var X and Y gives the probability of both variables taking specific values simultaneously.

Joint probability mass function: it is the joint probability function of discrete random var X and Y, for their simultaneous occurance.

$$P(X = x, Y = y) = f(x, y)$$

it satisfies the cond^n: 1. $f(x, y) \geq 0$ 2. $\sum_x \sum_y f(x, y) = 1$

Joint probability density function: it is the joint probability function of continuous random var X & Y, for their simultaneous occurance.

$$P(X = x, Y = y) = f(x, y)$$

it satisfies the cond^n: 1. $f(x, y) \geq 0$ 2. $\int_x \int_y f(x, y) = 1$

Marginal distribution: gives the distribution of one var X or Y alone regardless of other.

Marginal PDF: for discrete random variable.

$$f_x(x) = \sum_y f(x, y) \qquad f_y(y) = \sum_x f(x, y)$$

Marginal PDF: for continuous random variable

$$f_x(x) = \int_y f(x, y) \cdot dy \qquad f_y(y) = \int_x f(x, y) \cdot dx$$

X\Y	y_1	y_2	y_3 ----	y_m	
x_1	$f(1,1)$	$f(1,2)$ ----		$f(1,m)$	$g_1(x)$
x_2	$f(2,1)$	$f(2,2)$ ----			$g_2(x)$
:	:	:		:	:
x_n	$f(n,1)$	$f(n,2)$ ----		$f(n,m)$	$g_n(x)$
	$h_1(y)$	$h_2(y)$ ----		$h_m(y)$	1

Conditional distribution.

PMF discrete case : $P(X=x \mid Y=y) = \frac{P(X=x, Y=y)}{P(X=x)}$

PDF continuous case : $f(x|y) = \frac{f(x,y)}{f(y)}$ $\quad f(y|x) = \frac{f(x,y)}{f(x)}$

Measure of Association

COVARIANCE: a measure of how 2 variables are linearly related.

$cov(X,Y) = E[X-E(X)].E[Y-E(Y)] = E(XY) - E(X).E(Y)$

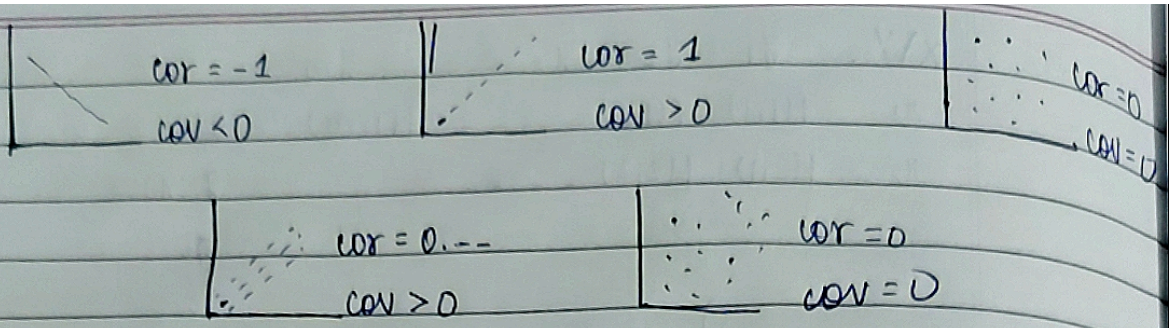
$cov(X,Y) = E[XY] - E(X).E(Y)$

if X and Y are independent
$cov(X,Y) = 0$

CORRELATN: measure of linear relationship b/w 2 variables.
(value b/w -1 and 1)

$corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$

VARIANCE: population : $\sigma^2 = \frac{1}{N} \sum_i (x_i - M)^2$ Sample : $S^2 = \frac{1}{n-1} \sum_i (x_i - \bar{X})^2$



$$-1 \leq \overset{\nearrow r}{cor} \leq 1$$

$r = -1$ ∴ strongly negatively related

$r = 0$: no relation

$r = 1$: strongly positively related.

Moment generating function.

it is used to generate moments of the distribution.

$$M_X(t) = E[e^{tx}]$$

$$M_X(t) = E[e^{tx}] = \begin{cases} \sum_x e^{xt} \cdot f(x) & : \text{discrete} \\ \int_x e^{tx} \cdot f(x) & : \text{continuous} \end{cases}$$

$$M_X(t) = E[e^{tx}] = E\left[1 + \frac{(xt)}{1!} + \frac{(xt)^2}{2!} + \dots \frac{(xt)^n}{n!} \dots\right] \rightarrow e^x \text{ series}$$

$$= E(1) + t E(x) + \frac{E(x^2) \cdot t^2}{2!} + \dots \frac{t^r E(x^r)}{r!} \dots \quad (1)$$

$$= 1 + \frac{t}{1!} M_1' + \frac{t^2}{2!} M_2' + \dots \frac{t^r}{r!} M_r'$$

$$M_X(t) = \sum_{r=0} \frac{t^r}{r!} M_r \quad \text{where } M_r' = E[x^r]$$

(Alternate method to find RAW MOMENTS)

diff^{n} ①

$$M'_x(t) = \frac{dM_x(t)}{dt} = 0 + E(x) + \frac{2t}{2!}(x^2) + \ldots \Big|_{t=0}$$

$$M_1' = E(x)$$

$$M_2' = M_x''(t)$$

$$M_3' = M_x'''(t) \text{ and so on.}$$

$$M_r' = \frac{d^r[M_x(t)]}{dt^r}\Big|_{t=0}$$

Properties

1. if 2 random var have the same MGF, they have the same distribut^{n}.
2. MGF of the sum of a no. of independent random var is equal to the product of their respective MGF.

$$M_{x_1+x_2+\ldots x_n}(t) = M_{x_1}(t) \times M_{x_2}(t) \times \ldots M_{x_n}(t)$$

3. $M_{cx}(t) = M_x(ct)$
4. effect of change of origin and scale.

$$U = \frac{x-a}{h}$$ → shifting of origin

→ shifting scale.

$$M_U(t) = E[e^{Ut}] = E\left[e^{\left(\frac{x-a}{h}\right)t}\right]$$

$$M_U(t) = e^{-\frac{at}{h}} M_x(t/h)$$

DISCRETE DISTRIBUTIONS

BINOMIAL DISTRIBUTN

A random experiment whose outcome is either 'success' or 'failure' with probability p and q resp.

PMF: $P(X=x) = {}^{n}C_x (p)^x (q)^{n-x}$, $x = 0, 1, \ldots n$

- PROPERTIES:

  Mean: $E(X) = np$

  Variance: $V(X) = npq$

  if n is large, it approx the normal distr for $p \simeq 0.5$

POISSON DISTRIBUTN

Models the no. of ~~trials~~ events occuring in a fixed interval of time or space, given a constant rate λ and independence of events.

→ no. of trials is infinitely large ($n \to \infty$)

→ probability of success is very small ($p \to 0$)

PMF: $P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}$ $x = 0, 1, 2 \ldots \infty$, $\lambda \geq 0$

- PROPERTIES

  Mean: $E(X) = \lambda$

  Variance: $V(X) = \lambda$

  it is often used as an approx^n for binomial distr^n where $n \to \infty$, $p \to 0$.

GEOMETRIC DISTRIBUTN

no. of trials required to get first success in series of independent Bernoulli trials.

PMF: $P(X=x) = (1-p)^{x-1}(p) = pq^{x-1}$ $x = 0, 1, 2 \ldots \infty$

- PROPERTIES

  Mean: $E(X) = q/p$ Variance: $V(X) = \frac{q}{p^2}$

CONTINUOUS DISTRIBUTIONS

NORMAL DISTRIBUTN

A bell shaped distribution characterized by its mean μ and sd σ.

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

- Mean = μ
- Variance = σ^2

→ total area under curve is 1.

UNIFORM DISTRIBUTN

All values within a specified interval $[a, b]$ are equally likely.

PDF: $f(x) = \begin{cases} 0, & x < a \text{ or } x > b \\ \frac{1}{b-a}, & a \le x \le b \end{cases}$

- Mean = $\frac{a+b}{2}$
- Variance = $\frac{(b-a)^2}{12}$

EXPONENTIAL DISTRIBUTN

models time until next event occurs in a Poisson Process.

PDF: $f(x) = \lambda e^{-\lambda x}$, $x \ge 0$

- Mean = $1/\lambda$
- Variance = $1/\lambda^2$

CORRELATION & REGRESSION

LINEAR CORRELATN

measures the strength and direction of a linear relationship b/w 2 variables X and Y.

correlation coefficient (r): $r = \frac{cov(X,Y)}{\sigma_x \sigma_y}$

RANK CORRELATION COEFFICIENT

measures ~~cor~~ correlation b/w ranks of 2 variables

Spearman's Rank correlation coefficient (ρ): $\rho = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$

REGRESSION

describes the relationship b/w a dependent variable Y and one more independent variable X.

simple linear regression : $Y = a + bX$

REGRESSN COEFFICIENT

R : of Y on X $\Rightarrow b_{yx} = r \frac{\sigma_y}{\sigma_x}$

R : of X on Y $\Rightarrow b_{xy} = r \frac{\sigma_x}{\sigma_y}$

} if $r = 0$, $b_{yx} = b_{xy} = 0$

- Properties

1. correlatn coeff $r = \sqrt{b_{yx} \, b_{xy}}$

   ~~if $b_{yx} < 1$, $b_{yx} > 1$~~

2. if $b_{yx} < 1$, $b_{xy} > 1$

3. $\frac{b_{yx} + b_{xy}}{2} > r$

4. $b_{yx} = r \frac{\sigma_y}{\sigma_x} = k \frac{\sigma_v}{h \sigma_u}$ $\left(u = \frac{x-a}{h} , v = \frac{y-b}{k} \right)$

→ independent of origin shifting, but depends on scale.

CENTRAL LIMIT THEOREM

Central limit theorem states that the sum (or average) of a large no. of independent and identically distributed random variables approaches a NORMAL DISTRIBUTION, regardless of original distribution.

APPLICABILITY:

- works for independent random variables with same distribution.
- Sample size should be large ($n \geq 30$)

STATEMENT:

if $X_1, X_2, X_3 \ldots X_n$ are (independent & identically distributed) random variables with mean μ & variance σ^2, the standardized sum:

$$Z = \frac{X - \mu}{\sigma} \qquad X \sim N(\mu, \sigma^2)$$

IMPLICATIONS

- allows approximation of probabilities using normal distribution.
- Justifies use of normality of large-sample hypothesis testing.

HYPOTHESIS TESTING

FORMATION OF HYPOTHESIS

A hypothesis is a claim or statement about a population parameter (eg mean proportion)

null hypothesis (H_0): assumes no effect or no difference. It is the statement being tested.

ex: $H_0: \mu = \mu_0$ (population mean equal to some value)

alternate hypothesis (H_1): Represents what we want to prove or detect.

ex: $H_1: \mu \neq \mu_0$ (mean is not equal to specific val)

LARGE SAMPLE TESTS

For large samples ($n \geq 30$), the Z-test is used. It assumes the sampling distribution of the test statistics follows a normal distribution.

a) test for single proportion: $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

b) test for difference of proportions: $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

c) test for single mean: $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

d) test for difference of means: $Z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

e) test for standard deviations: f-test

SMALL SAMPLE TESTS

For small samples ($n < 30$), the t-test is used, which accounts for additional uncertainty due to small sample size.

a) t-test for single mean: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

b) t-test for difference of means (equal variances):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

c) t-test for correlation coefficient

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

F-TEST FOR RATIO OF VARIANCES

F-test is used to compare variances of 2 samples

$$F = \frac{s_1^2}{s_2^2} \quad (F \geq 1)$$

CHI-SQUARE TEST

Chi-square test (χ^2) is used for categorical data.

a) Goodness of Fit: tests whether observed frequencies (O_i) match expected frequencies (E_i)

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

b) Test for independence: tests whether 2 attributes are independent.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

TESTING FOR STATISTICAL HYPOTHESIS

Step 1: Null hypothesis (H_0)

Step 2: Alternate hypothesis (H_1)

- $H_1 : M \neq M_0$ (2-tailed)
- $H_1 : M > M_0$ (right-tailed)
- $H_1 : M < M_0$ (left-tailed)

Step 3: Level of significance

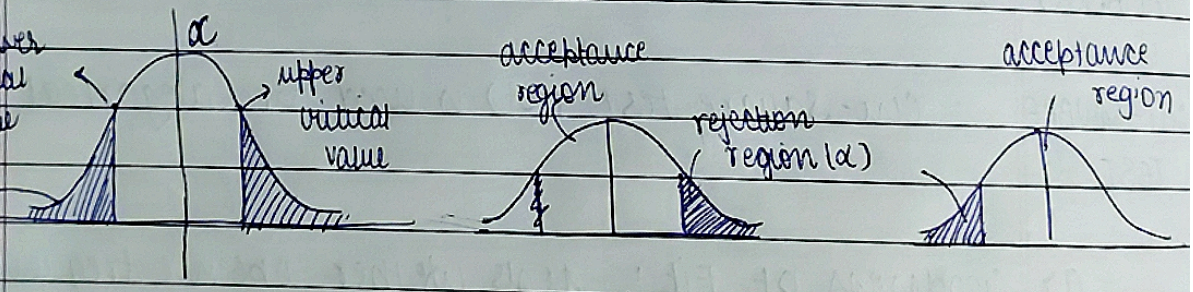
→ the probability of the variate falling in the critical region.

Step 4: Test statistic

compute test statistic z under null hypothesis.

$$Z = \frac{t - E(t)}{SE(t)}$$

Step 5: conclusion



2 tailed test | Right tailed test | Left tailed test

$|Z| > Z_\alpha$: Reject H_0 $|Z| < Z_\alpha$: accept H_0

	1% (0.01)	5% (0.05)	10% (0.1)
Two tailed	$\|Z_\alpha\| = 2.58$	$\|Z_\alpha\| = 1.96$	$\|Z_\alpha\| = 1.645$
Right tailed	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left tailed	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

Z-test

TEST OF SIGNIFICANCE FOR LARGE SAMPLES

$n > 30$

Single Proportion

$$S.E(p) = \frac{p - P}{\sqrt{PQ/n}} \sim N(0,1)$$

~~$E(p) = E\left(\frac{x}{n}\right) = \frac{np}{n} = P$~~

$p = x/n$: observed proportion of success
P : probability of success
Q : probability of failure

Difference of Proportion

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \qquad Q = 1 - P$$

Single Mean

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$\bar{X}$: mean sample μ : populatn mean
σ : S.D n : sample

limits of populatn mean $\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}$

Diff^n of means for 2 large Samples

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

σ : S.D $\bar{X}$: sample mean
n : sample size

$\sigma_1 = \sigma_2 = \sigma$; $\sigma^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$

Diff^n of Standard deviation

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}}$$

σ : populatn standard devn
S : sample standard devn
n : sample size

t-test

TEST OF SIGNIFICANCE FOR SMALL SAMPLE

$n < 30$

Mean of random sample

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$\bar{x}$: sample mean μ : populatn mean

s : sample S.D n : sample size

Mean of small sample from Normal populat

$$t = \frac{(\bar{x} - \bar{y})}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$\bar{x}, \bar{y}$: Sample mean

n : sample size

degree of freedom $= n_1 + n_2 - 2$

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \quad \text{or} \quad s^2 = \frac{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$