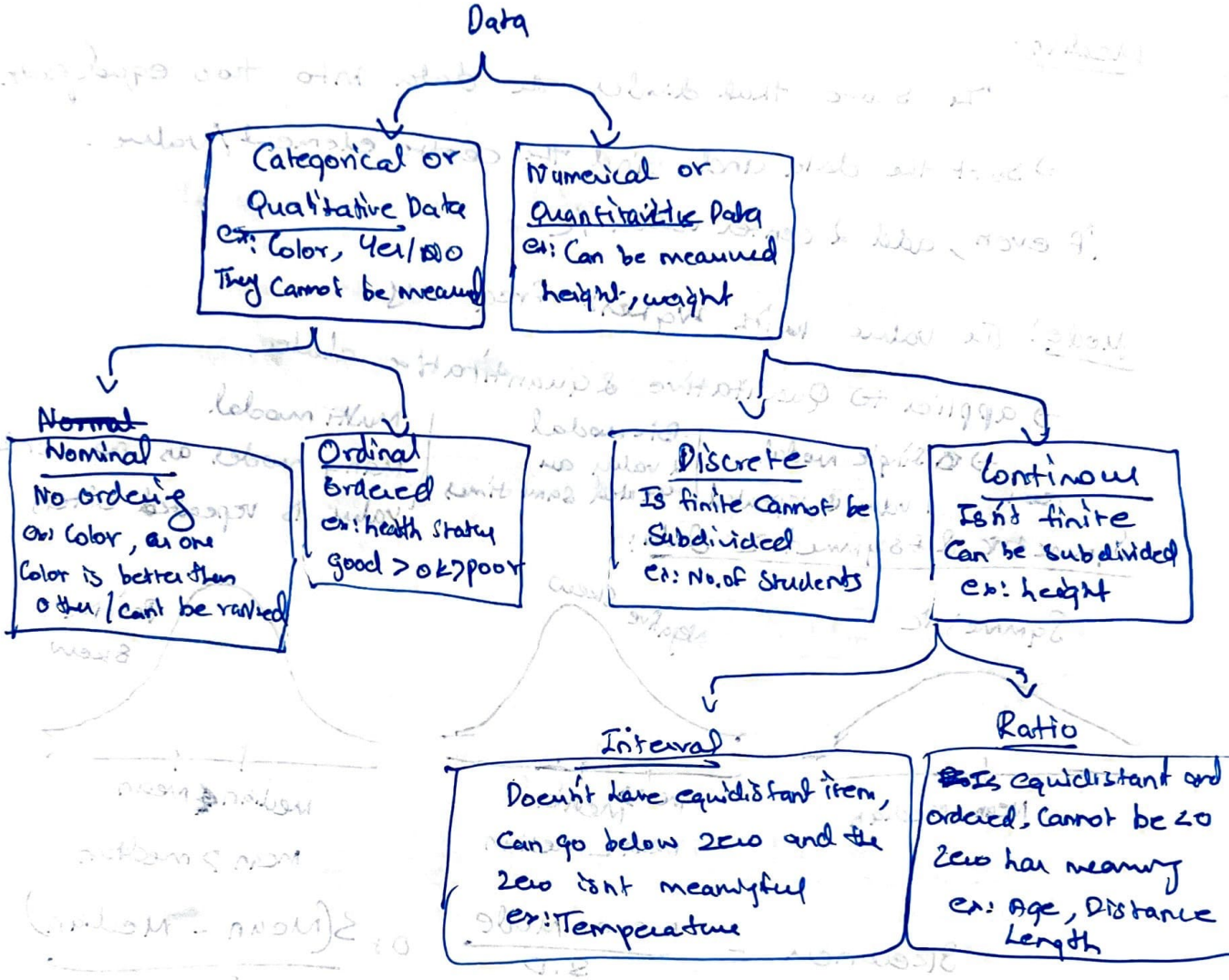
→ Statistics is the study of collection, analysis, interpretation, presentation and organization of data.

→ Types of data



Measure of Central Tendency:

↳ used to describe a set of ~~Scores~~ Scores with a single number that describes the performance of the group → center of data

- 3 used measures for central tendency
  - Mean
  - Median
  - Mode

- Mean: The arithmetic average $\bar{y} = \frac{\Sigma y}{N}$ y is the values, N is the count of values

ex: 1, 2, 3, 4, 5 $\quad \bar{y} = \frac{1+2+3+4+5}{5} = 3$

If presented in group for

x	1	2	3
f	2	1	3

→ Raw form: 1, 1, 3, 3, 3, 3

$\bar{y} = \frac{\Sigma fy}{N} = \frac{1\times2 + 2\times1 + 3\times3}{6}$

$= 2.166$

↳ A subset of population

$\bar{y}$ for Sample, μ for population

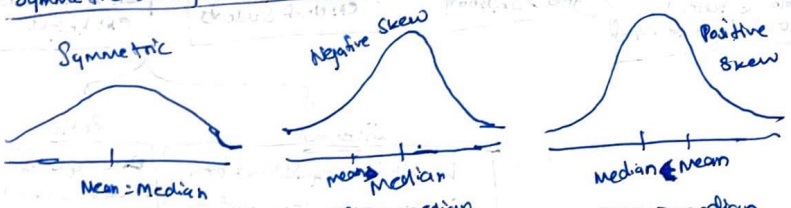
↳ ordinal data/group

→ Mean is the most stable and popular as it includes all data
→ Affected by extreme scores
→ Sum of each score's distance from mean is zero
→ Use when we want to measure standard deviation, coefficient of variation and skewness, and stability is desired.

Median:
The score that divides the data into two equal parts.
→ Sort the data and find the centre element / value.
if even, add 2 centre values /2

Mode: The value with highest frequency.
→ applies to Qualitative & quantitative data.
→ Single modal | Bi-modal: 2 values are repeated same times | Multi modal: Many modes as no 1 or 2 value is repeated often

Symmetric & Asymmetric data:



$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{S.D}} \quad \text{or} \quad \frac{3(\text{Mean} - \text{Median})}{\text{S.D}}$$

SD = Standard deviation

Measure of Variability:

- Range
- Variance
- Standard deviation

- Variability is the distance measure on the difference bet scores
- It describes the spread of scores or dist. of score from mean
- It captures how well the data is distributed and how well an individual scores represents the distribution

Range: Distance covered from smallest value to largest
= Max - Min value = Range

Standard Deviation: It represents how distributed the data is from mean
Most common & imp measure of variability
- Measures the standard or avg distance from mean

ex: take the data 1, 3, 7, 8, 9
the mean is 6 $\mu = 6$

~~Deviation = $X - \mu = (x_i - \mu)$~~

Standard Deviation $\sigma = \sqrt{\text{variance}} = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$ (value, mean, count)

$= \sqrt{\frac{(6-6)^2 + (5-6)^2 + (7-6)^2 + (8-6)^2 + (9-6)^2}{5}}$

$= \sqrt{\frac{25+61+1+4+9}{5}} = \sqrt{\frac{60}{5}} = \sqrt{8}$

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

$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$ for sample | In a sample σ becomes s, μ becomes $\bar{x}$, N is $n-1$

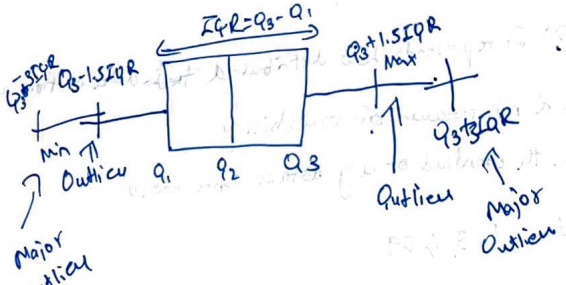
$\sigma^2 = \frac{SS}{n-1}$ → Sum of squared. Sum as $\Sigma(x_i - \mu)^2$

Mean and Std. Distribution help to describe the data

five point Summary: Repeat data in 5 values
- Minimum
- Q1
- Median
- Q3
- Maximum

Interquartile range = $Q_3 - Q_1$
IQR

Box and Whisker plot



Lower limit = $Q_1 - 1.5\,IQR$

Upper limit $Q_1 + 1.5\,IQR$.

Values above or below lower & upper limit are potential Outliers

Probability:

-) Random experiment is used to describe an action whose outcome is not known in advance
ex: tossing a coin

-) Sample space of a random exp. is a set S with all possible outcomes

-) Event is a subset of a sample space.

classical -> Prob of an ~~event~~ event = $\frac{\text{No. of fav outcomes}}{\text{Total}}$ ex: $P(\text{Head}) = 1/2$

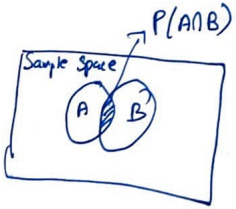
Empirical -> $\frac{\text{No. of times the event occurs}}{\text{Total}}$ ex: on 1000 coin toss 450 times head $= \frac{450}{1000}$

$P(S) = 1$

$0 \leq P(E) \leq 1$

$P(E_1)$ & $P(E_2)$ such that $P(E_1 \cap E_2) = \emptyset$ then

$P(E_1 \cup E_2) = P(E_1) + P(E_2)$



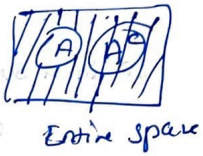
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2 Events are mutually exclusive if they dont have anything in common -> Dont intersect
↳ $P(A \cap B) = 0$ cannot happen at same time.

-> An independent event is one in which one event doesn't affect the other
-> If two events happen with replacement its independent

A^c Acomplement = $n(A^c) = n(S) - n(A)$
↓ sample space

$n(A) + n(A^c) = 1 = n(S)$



$n(A \cap B^c) = n(A) - n(A \cap B)$

$n(A^c \cap B) = n(B) - n(A \cap B)$

Probability of event A or B to happen is

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

If A & B are mutually exclusive

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

For independent events $P(A \cap B) = P(A) \cdot P(B)$

Conditional probability:

Prob of A given B has happened
↳ conditioning event

$P(A|B)$
↳ Happened $B \neq 0$

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

Multiplication Rule:

$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A|B) \cdot P(B) \quad B \neq 0$

$P(B|A) \cdot P(A) \quad A \neq 0$

$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot P(A_4|A_1 \cap A_2 \cap A_3) \dots P(A_n|A_1 \cap A_2 \dots A_n)$

Independent events

If B has no effect on A, $P(A|B) = P(A)$

$\therefore P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A) = \frac{P(A \cap B)}{P(B)}$

$P(A \cap B) = P(A) \cdot P(B)$ -> If this is true, the events are independent.

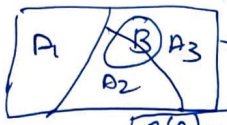
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

Law of total prob

If in a sample space with events A_1, A_2, A_3

$A_1 \cup A_2 \cup A_3 = S$ then it is called exhaustive events

if $A_1 \cap A_2 \cap A_3 = \phi$ It is mutually exclusive or the events don't intersect



$P(B) = P(A_1 \cap B) \cup P(A_2 \cap B) \cup P(A_3 \cap B) \Rightarrow P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$

$P(A_i \cap B) = P(B/A_i) \cdot P(A_i)$

$P(B) = P(B/A_1) \cdot P(A_1) + \dots + P(B/A_K) \cdot P(A_K)$

The prob of the event $= \sum_{i=1}^{k} P(B/A_i) P(A_i)$

B happens given A_1 has happened + prob B given A_2 has happened ... + given A_k has happened

Baye's theorem

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

Naive Bayes

-> Efficient ML algo for solving classification & regression problems

-> Assumes conditional independence, Bayes theorem does not.

i.e. this means the relationship bet. all input features are independent.

-) Bayes theorem for single variable input. $P(A/B)$

For more input with independence

-> It uses Bayes theorem to calculate the Posterior prob for events like Yes or NO

Ex: With given weather data find out if player should play on sunny day

frequency table

weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	4

To find out $P(Yes/Sunny)$ play on sunny day

$$P(Yes/Sunny) = \frac{P(Sunny/Yes) \cdot P(Yes)}{P(Sunny)}$$

$P(Sunny/Yes) = 3/10 = 0.3$ $\quad P(Yes) = 10/14 = 0.71$

$P(Sunny) = (3+2)/14 = 0.35$

$$P(Yes/Sunny) = \frac{0.3 \times 0.71}{0.35} = 0.6$$

$$P(NO/Sunny) = \frac{P(Sunny/NO) \times P(NO)}{P(Sunny)}$$

$P(Sunny/NO) = 2/4 = 0.5$ $\quad P(NO) = 4/14 = 0.29$

$$= \frac{0.5 \times 0.29}{0.35} = 0.41$$

$0.6 > 0.41$

$P(Sunny/Yes) > P(Sunny/NO)$

$\therefore$ Player can play on sunny day

→ A random variable assumes numerical values [illegible] random outcome of an experiment.

Its domain is the sample space and range is set of real numbers.

→ Discrete random variable: Takes finite or countably infinite no. of values

→ Continuous random variable: Takes numerical value in an interval or collection of intervals.

- Continuous random var are generated in exp. when things are 'measured' and not 'counted'

Ex: measure of time, weight, volume, length.

Discrete example: no of defective pieces, no. of steps.

~~between [0,1]~~ bet. [0,1]

If sum of all prob in a set is =1, It is pdf (each entry is bet. [0,1])

ex: x	1	2	3	4	Total	
$P_1(x)$	0.25	0.25	0.25	0.25	1	✓ PDF
$P_2(x)$	0.4	-0.1	0.3	0.4	1	✗ Not PDF as -ve value

$p(x)$ = likelihood that random variable takes the value x

PDF

Expected value: mean of a random variable.

$$E(x) = \mu = \Sigma x f(x)$$

Variance: Summarize the variability in the values of a random variable

$$\text{Var}(x) = \sigma^2 = \Sigma (x-\mu)^2 f(x)$$

Shortcut formula $= E(x^2) - [E(x)]^2 = \Sigma x^2 P(x) - \mu^2$

Standard deviation: positive square root of variance

Properties of ~~Random Variable~~ Expected value:

- $E(ax) = a E(x)$ a is constant
- $E(x+b) = E(x) + b$ b is const.
- $E(ax+b) = a \cdot E(x) + b$

Variance of discrete RV: $\sigma^2 = E[(x-\mu)^2] = \Sigma (x-\mu)^2 p(x)$.

Standard deviation of RV: $\sqrt{\sigma^2} = \sqrt{\Sigma (x-\mu)^2 p(x)}$.

Prop of variance:

$V(ax) = a^2 \cdot V(x)$ $\quad V(a) = 0$

$V(x+b) = V(x)$

$\sigma_{ax} = |a| \cdot \sigma_x \cdot \sigma_{x+b} = \sigma_x$

Random Variable

- Discrete: $P(x)$
- Continuous: $f(x)$

Validate: i) $0 \le P(x) \le 1$ $\quad 0 \le f(x) \le 1$

ii) $\Sigma P(x) = 1$ $\quad \int f(x)\,dx = 1$

} Should satisfy

Probability distribution function — Probability density function

example prob:

x	1	2	4	8	16
P(x)	0.05	0.10	0.35	0.40	0.10

Compute: a) $E(x)$

b) $V(x)$.

c) SD

d) $V(x)$ using shortcut formula.

x	P(x)	x P(x)	$(x-\mu)^2$	$(x-\mu)^2 P(x)$	$x^2 P(x)$
1	0.05	0.05	29.20	1.48	0.05
2	0.1	0.2	19.80	1.98	0.4
4	0.35	1.4	6.00	2.1	5.6
8	0.4	3.2	2.40	0.96	25.6
16	0.1	1.6	91.20	9.12	25.6
	1 ✓	↓		15.645	57.25

PDF $\quad E(x) = \mu = \Sigma x P(x) = 6.45$

a) $E(x) = \mu = \sum x p(x) = 6.45$

b) $V(x) = \sum (x-\mu)^2 p(x) = 15.645 = \sigma^2$

c) $SD = \sigma = \sqrt{V(x)} = \sqrt{15.645} = 3.96$

d) $V(x)$ using shortcut $= E(x^2) - [E(x)]^2 = \sum x^2 p(x) - \mu^2$

$E(x) = x p(x)$

$= 57.25 - (6.45)^2$

$\sigma^2 = 15.645$

If the problem or PDF is given as → Probability Density func.

$$f(x) = \begin{cases} Cx^2, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

it is a continuous distribution and the range is given $-1 \le x \le 1$

Imp formula:

① $\int x^n dx = \frac{x^{n+1}}{n+1}$ $n \neq -1$ } Learn this.

② $\int e^{ax} dx = \frac{e^{ax}}{a}$ }

③ $\int uv\,dx = u\int v\,dx - \int \left(\frac{du}{dx}\int v\,dx\right) dx$

④ $\int \frac{1}{1+x^2} dx = \tan^{-1} x$

problem:

x be a random variable with PDF

$$f(x) = \begin{cases} Cx^2, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

a) Find constant C

b) Find $E(x)$ and $V(x)$

c) Find $P(x \ge 1/2)$

a) as the limit is $-1 \le x \le 1$

$\int f(x)\,dx = 1$

$\int_{-1}^{1} Cx^2 dx = 1$

$C\left[\frac{x^3}{3}\right]_{-1}^{1} = 1$ → using formula ①

$C\left[\frac{1}{3} - \frac{(-1)}{3}\right] = 1$

$C\,2/3 = 1$

$\boxed{C = 3/2}$ ⇒ $f(x) = \begin{cases} 3/2\,x^2, & -1 \le x \le 1 \\ 0, & \text{others} \end{cases}$

b)

$E(x) = \int x f(x)\,dx$ → Not ≤ but $\int$

$= \int_{-1}^{1} x\left[\frac{3}{2}x^2\right] dx$

$\left[\frac{3}{2\times 4}x^4 - \frac{3}{2\times 4}x^4\right]_{-1}^{1}$

$= \left[\frac{3}{2\times 4} - \frac{3}{2\times 4}(-1)^4\right]$

$E(x) = 0$

$V(x) = E(x^2) - [E(x)]^2 = \int x^2 f(x)\,dx - 0$

$= \int_{-1}^{1} x^2\left[\frac{3}{2}x^2\right] dx$

$= \left[\frac{3}{2\times 5}x^5\right]_{-1}^{1}$

$= \left[\frac{3}{2\times 5} - \frac{3}{2\times 5}(-1)^5\right] = 3/5$

c) $P(x \ge 1/2)$ → change interval

$= \int_{1/2}^{1} f(x)\,dx = \int_{1/2}^{1} 3/2\,x^2 dx$

$= \left[\frac{3}{2}x^2\right]_{1/2}^{1}$ ⇒ $\left[\frac{3}{2\times 3}(1)^3 - \frac{3}{2\times 3}\left(\frac{1}{2}\right)^3\right]$

$= \left[1/2 - 1/16\right] = 7/16$

Properties of PDF → Probability Density function

1) $f(x) \geq 0$

2) $\int_{-\infty}^{\infty} f(x)dx = 1$

3) $P(a \leq X \leq b) = \int_a^b f(x)dx$ = area under $f(x)$ from a to b for any a and b

Cumulative Density function

$$CDF = F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

$$\text{PDF } f(x) = \frac{d\,F(x)}{dx} \quad \text{CDF}$$

Continous Random variable formulas:

Same as PDF but use $\int$ instead of Σ

$$\mu = E(x) = \int_{-\infty}^{\infty} f(x)dx$$

$$\sigma^2 = V(x) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$$

$$SD = \sigma$$

Joint probability distribution:

Discrete — now uses x, y 2 random variables instead of 1.

2 Condition if $X = \{x_1, x_2 .. x_n\}$ & $Y = \{y_1, y_2 .. y_m\}$

① $0 \leq P_{ij} \leq 1$ ~~$P_{ij} = 1$~~ ~~① $P_{ij} \geq 0$~~ 2 Condition.

$P_{ij} = P_{ij}$ ② $\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} = 1$ Sum of 2 random variable = 1

Ex:

X / Y	1	2	3	
0	0.1	0.1	0.2	0.4
1	0.2	0.1	0.3	0.6
Total	0.3	0.2	0.5	1

Marginal PDF x = 1 2 3, $P(x)$ = 0.3 0.2 0.5 — Marginal PDF of y = 0 1, $P(y)$ = 0.4 0.6

Continous:

~~① $f(x,y) \geq 0$~~

① $0 \leq f(x,y) \leq 1$

② $\iint_R f(x,y)\,dx\,dy = 1$

Bernoulli Distribution:

↳ Discrete random var

Probability mass function

$$P(x) = \begin{cases} p^x q^{1-x}, & x = 0,1 \\ 0, & \text{otherwise} \end{cases}$$

$\mu = \sum x P(x) = 0.P(0) + 1.P(1)$

$\mu = p$

$\sigma^2 = pq$ ⟶ Derivation: not needed $\sigma^2 = E[x^2] - (E[x])^2$

$= \sum x^2 P(x) - p^2$

$= 0^2.P(0) + 1^2.P(1) - p^2$

$= p - p^2$

$= p(1-p)$

$= pq$

Binomial Distribution:

To calculate prob of an event happening in n attempts Ex:

$$P(x) = \begin{cases} {}^nC_x p^x q^{n-x}, & x = 0,1,2 \ldots n \\ 0, & \text{otherwise} \end{cases}$$

Prob of getting 4 in dice roll = 1/6

" " not getting 4 = 5/6

Let no. of attempt $n = 10$

$P(x=10) = {}^{10}C_{10} p^{10} q^{10-10} = \left(\frac{1}{6}\right)^{10}$

→ It is discrete prob. dist

$\mu = np$ $\sigma^2 = npq$

- → n is finite
- → trails are independent of each other
- → prob of success p is constant for each trial
- → Each trail results in two mutually exclusive events success & failure.

Poisson

$$P(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!}, & x = 0, 1, 2 \dots \\ 0, & \text{elsewhere} \end{cases}$$

$\lambda = np$ for trials

-> It is used for rare events when p prob is very small and/or trials n is very large

-> Binomial dist. can be approximated by poisson dist. when $n \to \infty$, $p \to 0$ such that $\lambda = np$ = constant

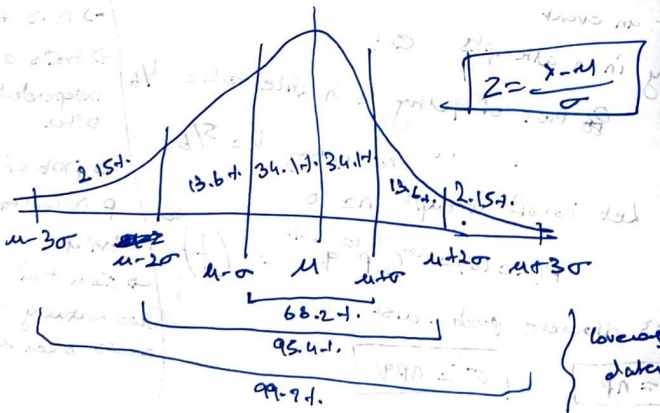
$$\boxed{\mu = \lambda} \quad \boxed{\sigma^2 = \lambda}$$

Normal Distribution:

Information such as heights, weights, length, size, profit, price or certain large data with mean & variance follows normal distribution.

-> A continuous random variable X which assumes all possible values in the entire real space $-\infty < x < \infty$ is said to follow normal dist. with 2 parameters μ & σ^2

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty, \; -\infty < \mu < \infty, \; \sigma > 0$$



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

properties:

1) Normal curve is bell shaped and symmetric about mean
2) Mean = Mode = Median
3) Total area under normal curve = 1
4) Normal curve approaches but never touches x axis as it extends farther away from the mean.

Normal Approximation to Binomial:

if $n > 30$ & $np, nq > 15$ then we normal approximation

$$Z = \frac{X-\mu}{\sigma} = \frac{X-np}{\sqrt{npq}} \quad \text{as } \mu = np \text{ & } \sigma = \sqrt{npq}$$

ex: for $n = 40$, $q = 0.5$, $p = 0.5$

$P(x = 12)$ = Creating a limit for approx. $P(11.5 < x < 12.5)$ = 2 Score $(-2.68 < Z < -2.37)$

find prob using Z table

$= P(-2.37) - P(-2.68) = 0.00889 - 0.00368 = 0.00521$

for $P(X \le 12) = P(X < 12.5)$

Sampling:

Selecting a group from a population

ex: selecting marks of 10 students from a class of 40

Random Sampling: Randomly select from population

Stratified sampling: Group the population and then select

ex: A C C A C B B B A C → Stratified: A A A A, B B B B, C C C C → Sample: A A, B B, C C

↑ population

Sampling error can occur due to a biased selection

Sampling variability is the variance in the sample data, larger the sample, lower the variability.

Sampling z formula = $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

When $n/N > 0.5$ $\frac{\text{sample size}}{\text{pop. size}} > 0.5$ use finite correction factor

No of samples of size x that can be drawn from a population of size n is
with replacement: n^x
without replacement: nCx

Central Limit theorem:

→ Regardless of the shape of population distribution, for large sample sizes ($n >= 30$), the sampling distribution of mean is normally distributed.

→ If $n < 30$ but the population is normally distributed, then the sampling dist. of mean $\bar{x}$ is also normally distributed.

→ The sample mean can be analyzed using z score

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

→ If the population is finite & $N/n > 0.05$ (ie) more the SD of population is used in sample, use finite correction

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}\sqrt{\frac{N-n}{n-1}}} \quad \text{if } N/n > 0.05$$

Sampling Proportion:

The group within a population

ex: 10 red balls in a pack of 50

$$\hat{p} = \frac{x}{n} = \frac{10}{50}$$

- Expected value of proportion $E(\hat{p}) = p$
- SD of sample proportion $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$

z formula for sample proportion

$$z = \frac{\hat{p}-p}{\sqrt{pq/n}}$$

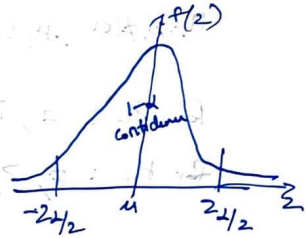
Confidence interval:

→ It is better compared to point estimate, used to estimate where the μ of a population lies.

→ ~~α is the confidence~~

the confidence interval is given by:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



α is the area under the normal curve outside the confidence interval area

ex: To find 95% confidence interval take α as 0.05 because

$1-\alpha = 95$; $\alpha = 5\% = \frac{5}{100} = 0.05$

Substitute $\alpha/2 = 0.025$ and get z by finding z value for 0.025

z score for 90% = 1.65 92% = 1.75
95% = 1.96
99% = 2.58

→ If a finite population is given use finite correction $\sqrt{\frac{N-n}{n-1}}$ if $N/n > 0.05$

Population proportion:

→ Confidence interval for population proportion

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}} \leq \mu \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$$

when np or $nq > 5$

Hypothesis testing:

-> Hypothesis is a statement or a claim

-> Hypothesis testing validates that claim based on data

- H_0 : Null hypothesis — H_0 or H_n
  ↳ no significant change or difference

H_1: Alternative hypothesis — H_1 or H_a
↳ opposite of null hypothesis

$H_0: \mu = 100$, $H_0: \mu \geq 100$ example of null hypothesis

$H_1: \mu \neq 100$, $H_1: \mu_1 > \mu_2$ ex of alternative hypothesis

for H_0 : $=$, $\geq$ or $\leq$ is accepted

for H_1 : $\neq$, $>$, $<$ is used

Test:

When $\mu_1 < \mu_2$ => one-tailed test or ~~right~~ left tailed test

$\mu_1 > \mu_2$ => one-tailed or ~~left~~ right tailed test

$\mu_1 \neq \mu_2$ => two tailed test

-> Critical region is the area when the hypothesis is rejected even if the hypothesis is true. It is ~~the~~ beyond the confidence interval area or extreme area.

α = P(Rejecting H_0 while H_0 is true)

β = P(Accepting H_0 while H_0 is false)

	H_0 is true	H_0 is false
Acc H_0	✓	β
Rej H_0	α	✓

α is type 1 error

β is type 2 error

2 test is used when

- Mean of single population (μ)
- Dif bet. 2 population means ($\mu_1 - \mu_2$)
- proportion of single population (p)
- Diff. bet. 2 proportions ($p_1 - p_2$)

for z test:

$n \geq 30$, σ should be known.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Hypothesis testing formula.

	Sample size	formula
One mean	Large	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
	Small	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Two mean	Large	$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
	Small	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
One proportion	Large	$z = \frac{p - p_0}{\sqrt{\frac{pq}{n}}}$
Two proportion	Large	$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$
Students paired t-test	Small	$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$

Critical vals	Level of significance α		
	1 %.	5 %.	10 %.
Two tailed test	$z_{\alpha/2} = 2.58$	$z_{\alpha/2} = 1.96$	$z_{\alpha/2} = 1.645$
One tailed test	$z_\alpha = 2.33$	$z_\alpha = 1.645$	$z_\alpha = 1.28$

P-value can be used to determine the evidence of rejection or ac

Give 2, find the value in 2 table

for ex: if $z = 1.65$, we get p value as $1 - 95.05 = 4.95$

if

p value < 1% ⇒ overwhelming evidence to support alt. hypoth

p value bet 1% & 5% ⇒ strong evidence to support alt. hypoth

p value bet. 5% & 10% ⇒ weak evidence

p value above 10% ⇒ no evidence to support alt. hypothesis

t-test

→ When sample is small < 30

Same formula, $\frac{\bar{x} - \mu}{s/\sqrt{n}}$

the obtained value is t value and should be checked in the t-table

Consider the degrees of freedom $n-1$ (sample size)

Chi-Square test:

- Independence test
- Goodness of fit

→ used only for frequencies

H_0: Independent ⇒ Observed = expected

H_1: Not - Independent ⇒ Observed ≠ expected

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \quad k = r \times c \text{ of cells}$$

Ex: Accident data by weekday

test for independence bet. weekday & accident

	O	E	O-E	$(O-E)^2/E$
Mon.	184	150	34	7.71
Tue	148	150	-2	0.03
Wed	145	150	-5	0.17
Thu	153	150	3	0.06
Fri	150	150	0	0
Sat	184	150	4	0.11
Sun	116	150	-34	7.71
	1050	1050		15.27

$E = \frac{1050}{7} = 150$

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

taking S.L. LOS at 1%

↳ degree of freedom $= 7 - 1 = 6$ $(n-1)$

we get critical $\chi^2_{critical} = 12.6$

as $15.77 > 12.6$ we reject H_0 and accept H_1

↓

they are dependent on day of week.

Ex 2:-

	Cancer	No cancer	Total
Smoker	400 (a) 326.67	800 (b) 373.33	200 r_1
Non Smoker	800 (c) 373.33	500 (d) 426.67	800 r_2
Total	700 c_1	800 c_2	1500 n

Calculate as $\frac{R_i \times C_j}{\text{Total}}$

= row total × column total / Total

$= \frac{700 \times 800}{1500}$

$$X^2_{cal} = \frac{(400 - 326.67)^2}{326.67} + \frac{(800 - 373.33)^2}{373.33} + \frac{(500 - 426.67)^2}{426.67} + \frac{(1300 - 373.33)^2}{373.33}$$

Shortcut $X^2 = n\frac{(ad-bc)^2}{r_1 r_2 c_1 c_2}$ where a, b, c, d are the entries

Fisher's exact probability $\Rightarrow p = \frac{1}{n!} \frac{a_1!}{a!} \frac{a_2!}{b!} \frac{c_1!}{c!} \frac{c_2!}{d!}$

Yate's correct $= \frac{(|O-E| - 0.5)^2}{E}$

Variance based test:

Fisher's F-test: ~~Ratio of two~~

Single variance: $H_0: \sigma^2 = \sigma_0^2$ or $H_1: \sigma^2 < \sigma_0^2$

$H_1: \sigma^2 > \sigma_0^2$ $\quad X^2 = \frac{(n-1)s^2}{\sigma_0^2}$

$H_1: \sigma^2 \neq \sigma_0^2$ $\quad \sim \chi^2_{(\alpha, n-1)}$

Ratio of 2 variance $H_0: \sigma_1^2 = \sigma_2^2$ or $H_1: \sigma_1^2 < \sigma_2^2$ $\quad F = \frac{s_1^2}{s_2^2} \sim F(\alpha, n_1, n_2)$

$H_1: \sigma_1^2 > \sigma_2^2$

$H_1: \sigma_1^2 \neq \sigma_2^2$

Ratio of 2 variance test formula:

$$f = \frac{s_1^2}{s_2^2} \sim f(\alpha, n_1 - 1, n_2 - 1)$$

~~Table~~ $f_{critical}$ is obtained from F table with degrees of freedom of 1st & 2nd sample.

Correlation analysis

→ Statistically checks for the existence & extent of relationship among 2 variables

Covariance:

$$\text{Covariance}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

if it is -ve, it means $x \propto \frac{1}{y}$, one increases the other decreases

if +ve, $x \propto y$

if no relation, $= 0$

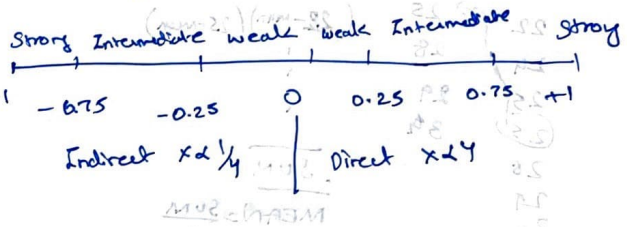
→ Covariance only shows how 2 variables change together not the dependency between them

Correlation coefficient:

→ Can find the strength of relationship by eliminating the unit of the value by dividing with SD.

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

for r value



Regression analysis:

Linear regression created a best fit line to predict value.

~~Linear~~ $y = \beta_0 + \beta_1 x$

Given values of y & x substitute in the below formula and solve the two eqn to get β_0 & β_1

$$\Sigma y = n\beta_0 + \beta_1 \Sigma x \quad \text{①}$$

$$\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2 \quad \text{②}$$

Maximum likelihood Estimation:

for $p = 1/n$

for binomial distribution $= \frac{\bar{x}}{n}$

for poisson distribution $= \bar{x}$

for normal distribution $\bar{x} = \frac{\sum x_i}{n} = \bar{x}$ and $\sigma^2 = \frac{\Sigma (x_i - \bar{x})^2}{n}$

for uniform distribution $= x_1$ & x_n when x_1 is min value & x_n is max value in the interval