* Bivariate Random Variable :-

Let S be a Sample Space associated with random experiment. Let $X(t)$, $Y(t)$ be 2 random variable function each assigning a real number to each outcome $t \in S$.

Then (X, Y) is called a Bivariate or 2-dimensional Random Variable.

→ PMF of (x, y) :-

If (x, y) is a 2-D discrete random variable, where $X = x_i$ & $Y = y_j$, $i \& j = 1, 2, \ldots$

Then $P(X = x_i, Y = y_j) = P_{ij}$ is called Probability Mass Function of (X, Y) Provided.

① $P_{ij} \geq 0$

② $\sum_{i=1}^{m} \sum_{j=1}^{n} P_{ij} = 1$

The set of triple (x_i, y_i, P_{ij}) is called Joint Probability distribution of (X, Y).

→ Marginal PMF of x

$P(X = x_i) = \sum_{j=1}^{n} P_{ij} = P_{i1} + P_{i2} + \ldots$

The collection of Pair (x_i, P_i) is called marginal PMF of x.

Similarly

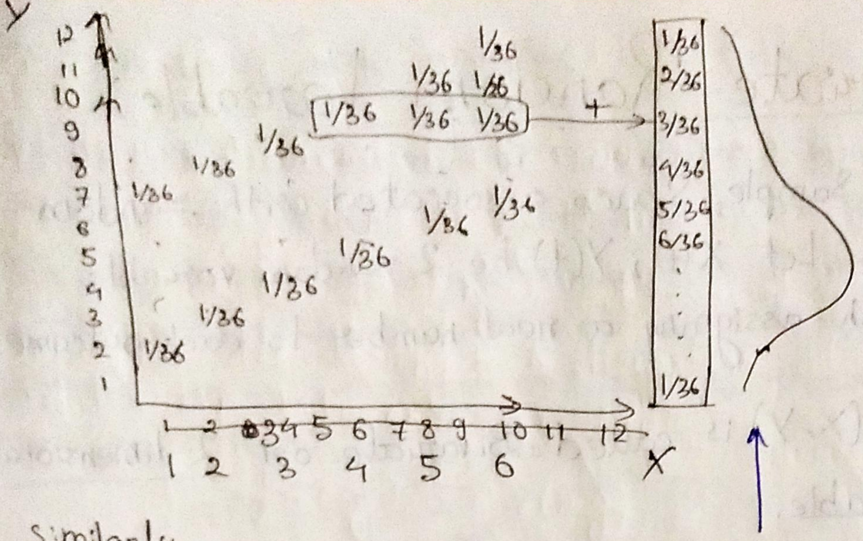
Marginal PMF of y :-

$P(Y = y_i) = \sum P_{ij} = P_{ij} + P_{2j} + \ldots$

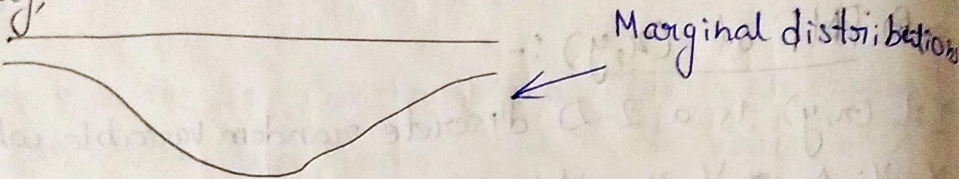
The pair (y_j, P_j) is called marginal PMF of y.

⇒ If X & Y are independent :-

$P_{XY}(x, y) = P_X(x) P_Y(y)$



similarly,



X : The no. rolled on the 1st dice
Y : Sum of the two dice

$P_X(x_i) = \sum_j P_{XY}(x_i, y_j)$ | $P_Y(y_j) = \sum_i P_{XY}(x_i, y_j)$

→ Conditional distribution:-

It is the probability distribution of a random variable, calculated according to the rules of conditional probability after observing the realization of another random variable.

$$P_{Y|X=x}(y) = P(Y=y|X=x) = \frac{P(X=9, Y=49)}{P(X=9)}$$

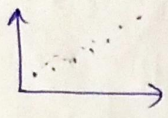
Taking, $x=9$ & $y=49$

* Covariance:-

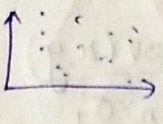
Covariance measures the direction of the relationship between two variables.

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

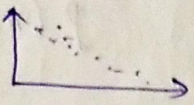
Eg:-



$Cov(x,y) > 0$
if $x \uparrow$ then $y \uparrow$
$x \propto y$



$Cov(x,y) \approx 0$



$Cov(x,y) < 0$
if $x \uparrow$ then $y \downarrow$
$x \propto \frac{1}{y}$

→ Covariance Matrix:-

Covariance Matrix is a measure of how much 2 R.V. gets change together. It is actually used for computing the covariance in between every column of data matrix.

(Also known as Dispersion matrix or Variance-covariance matrix)

$$\begin{bmatrix} Var(x_1) & \cdots & Cov(x_n, x_1) \\ \vdots & \ddots & \\ Cov(x_n, x_1) & \cdots & Var(x_n) \end{bmatrix}$$

Eg:-

$$\begin{bmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{bmatrix}$$

for only (x, y)

$$\begin{bmatrix} Var(x) & Cov(x,y) & Cov(x,z) \\ Cov(x,y) & Var(y) & Cov(y,z) \\ Cov(x,z) & Cov(y,z) & Var(z) \end{bmatrix}$$

for only (x, y, z)

* Pearson Correlation Coefficient :-

The Pearson Correlation Coefficient (r) is the most common way of measuring a linear correlation.

It ranges between -1 and 1 that measures the strength & direction of the relationship between 2 variables.

$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \times \sigma_y}$$

(Best in linear)

$$-1 \le \rho(x,y) \le 1$$

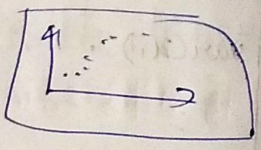
* Spearman Rank Correlation :-

While Pearson's correlation b/w the rank values of 2 R.V.s.

Spearman's correlation assesses monotonic relationships (whether linear or not).

It also ranges b/w -1 to 1.

$$r_s = \rho_{R(X),R(Y)} = \frac{cov(R(X),R(Y))}{\sigma_{R(X)}\ \sigma_{R(Y)}}$$



or

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

x_i	y_i	rank x_i	rank y_i	d_i	d_i^2
				$\sum$	$\sum d_i^2$

d_i = rank x_i - rank y_i

* Multivariate Gaussian Distribution :-

Multivariate Gaussian distribution is a generalization of the one-dimensional (univariate) to higher dimension.

$x = [x_1 x_2 \dots x_n]^T$ $\quad \mu = [\mu_1 \mu_2 \dots \mu_n]^T$ (Mean)

$$f_X(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$|\Sigma|$ → Determinant of covariance matrix

---

* Moments :-

Moments are popularly used to describe the characteristic of a distribution.

Let's say the random variable of our interest is X then, moments are defined as the X's expected values.

① $\frac{\sum x}{n} \Rightarrow$ Mean

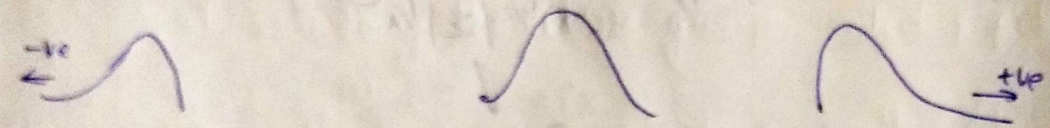
② $\frac{\sum x^2}{n} \rightarrow \frac{\sum (x-\mu)^2}{n} \Rightarrow$ Variance $\frac{\sum (x-\bar{x})^2}{n-1}$ (centralized) (population) (sample)

③ $\frac{\sum x^3}{n} \rightarrow \frac{\sum (x-\mu)^3}{n} \rightarrow \frac{1}{n^x} \frac{\sum (x-\mu)^3}{\sigma^3}$ [Skew] $\rightarrow \frac{n}{(n-1)(n-2)} \frac{\sum (x-\bar{x})^3}{s^3}$ (standardized) (population) (sample)

④ $\frac{\sum (x)^4}{n} \rightarrow \frac{1}{n} \left(\frac{\sum (x-\mu)^4}{\sigma^4} \right) \rightarrow \frac{n(n+1)}{(n-1)(n-2)(n-3)} \left(\frac{\sum (x-\bar{x})^4}{s^4} \right) - \frac{3(n-1)^2}{(n-2)(n-3)}$ (population) (sample)

→ Skewness :-

Skewness is a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data.



-ve skew 0 skew +ve skew

- Pearson's Method :-

① Mode skewness :-

$$\text{skew} = \frac{\text{mean} - \text{mode}}{\sigma}$$

② Median skewness :-

$$\text{skew} = \frac{3(\text{mean} - \text{median})}{\sigma}$$

$\because$ mode = 3median − 2mean
(In skewed data)

- With moments :-

$$\frac{n}{(n-1)(n-2)} \times \frac{\sum (x-\bar{x})^3}{s^3}$$

→ Kurtosis :-

Kurtosis is a measure of the tailedness of a distribution

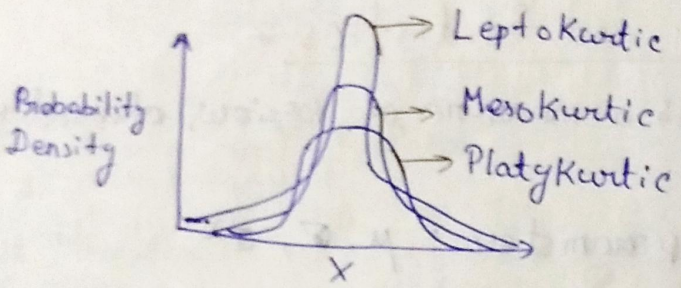
$$\text{Kurtosis}(K) = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \left(\sum \frac{(x-\bar{x})^4}{s^4} \right) - \frac{3(n-1)^2}{(n-2)(n-3)}$$

- A normal distribution has K = 3, (mesokurtic)
- Distribution with K > 3, (leptokurtic) (Excess tailedness)
- Distribution with K < 3, (platykurtic) (Thin tailed)

⇒ Types of Kurtosis:-

	Mesokurtic	Platykurtic	Leptokurtic
Tailedness	Medium-tailed	Thin-tailed	Fat-tailed
Outlier frequency	Medium	Low	High
Kurtosis	Moderate (3)	Low (<3)	High (>3)
Excess Kurtosis	0	Negative (-ve)	Positive (+ve)

Example



(Sampling Theory)

Sample & Sampling

* **Population** :- A population is a distinct group of individuals, whether that group comprises a nation or group of people with a common characteristic. Instead of examining the entire group, called population.

* **Sample** :- We may examine only a small part of this population which is called sample.

→ **Parameter and Statistics** :-

(I) Parameter is a measure of various characteristics of a population.

Population parameter : μ, σ, σ^2

(II) Statistics is a measure estimated through various sample of population.

Sample statistics :- $\bar{x}, S, S^2$

→ **Sampling & Non-Sampling Error** :-

(1) The difference b/w parameters and statistics is known as the sampling error.

(2) These can't be controlled by humans.

(3) These errors come into existence when various sample we taken.

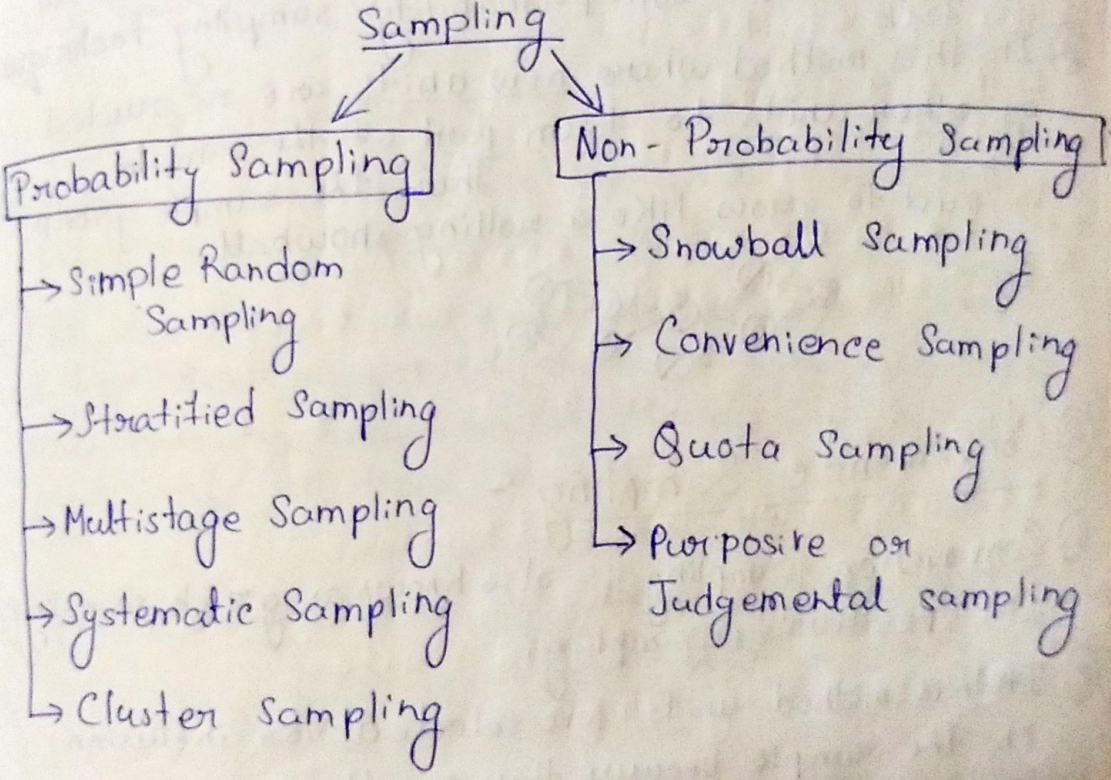
→ Non-Sampling Error :-

(i) These are various systematic errors and random error that are not due to sampling.

(ii) These can be controlled by humans.
Eg:- Errors in processing.

(iii) ~~These can control~~
Non-sampling errors are systematic error.

→ Sampling Techniques :-



* When we use non-probability Sampling?

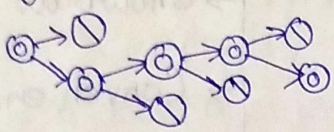
① Sample size should be small (less than 30)
② Biased (depend on research)
③ Qualitative research.
④ Also known as non-random sampling.

→ Non-Probability Sampling Types:-

① Snowball Sampling:-

① Snowball sampling / chain sampling / chain referral sampling is a non-probability sampling technique
② In this method where new units are recruited by other units to form part of the sample. Thus, the sample group is said to grow like a rolling snowball.

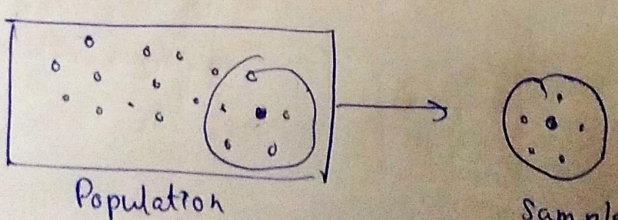
Eg:-



② Convenience Sampling:-

① Convenience sampling is also known as grab sampling or opportunity sampling.
② In this method units are selected for inclusion in the sample because they are the easiest for the researcher to access.
③ This type of sampling is most useful for pilot testing.

Eg:-

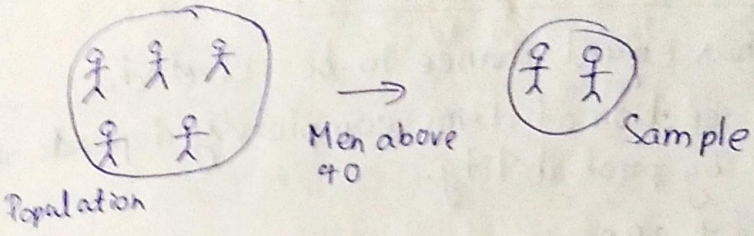


③ Quota Sampling :-

① Quota Sampling is the non-probability version of stratified sampling.

② In this assembled sample has the same proportions of individuals as the entire population with respect to known characteristics, traits or focused phenomenon.

Eg:-

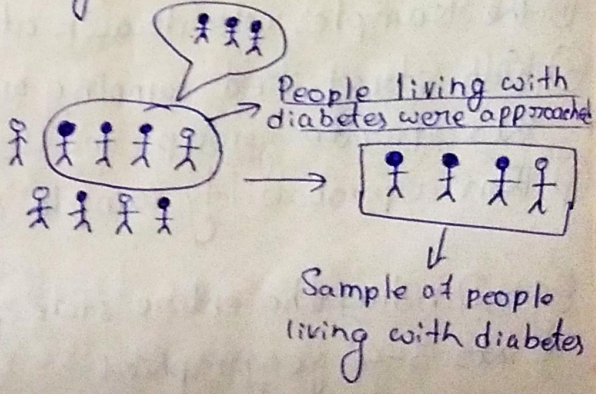


④ Purposive Sampling :-

① Purposive sampling, also known as judgement, selective or subjective sampling.

② It is a form of non-probability sampling in which researchers rely on their own judgement when choosing members of the population to participate in their study.

Eg:-



* When we use probability sampling? (Big dataset)

1. Almost no Biased → Bias ≈ 0
2. Quantitative Data
3. Sample size is large.

→ Types of probability Sampling:-

1. Simple Random Sampling (SRS):-

(i) Each item has equal chance to be selected.
(ii) Here the selection of items completely depends on chance or by probability.

⇒ Two Types of SRS:-

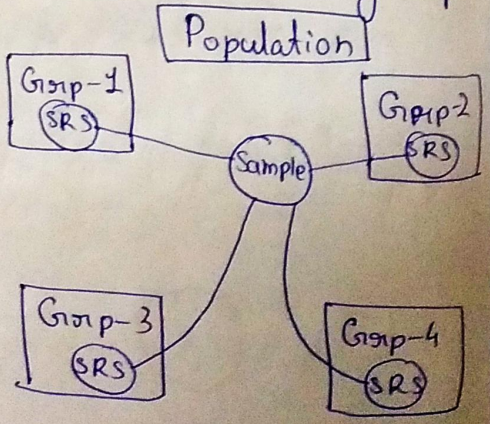
(i) With Replacement
(ii) Without Replacement

Eg:- Choosing 25 employee names from a hat containing 250 employee names. (Every name has equal chance of being chosen)

2. Stratified Sampling:-

1. The samples are drawn at random.
2. With stratified sampling researchers divides the population into separate groups called strata.
3. Then a probability sample is drawn from each group.

Eg:- Dividing the entire state into 3-4 geographical regions and then selecting required sample of districts from each of these regions.



③ Systematic Sampling :-

① In this sample members from a larger population are selected according to random starting point but with a fixed, periodic interval.

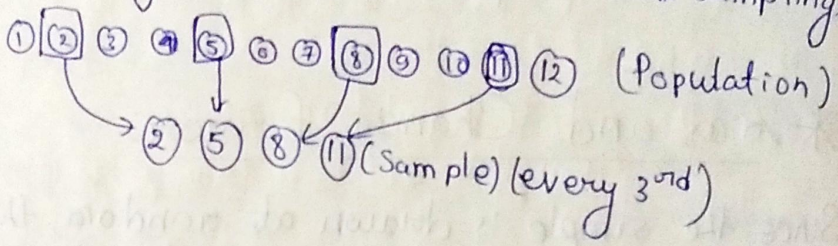
② This interval is called the sample interval (K).

③ Sampling interval is calculated by dividing the population size (N) by the desired sample size. (n)

$$K = \frac{N}{n}$$

④ It is functionally similar to simple random sampling.

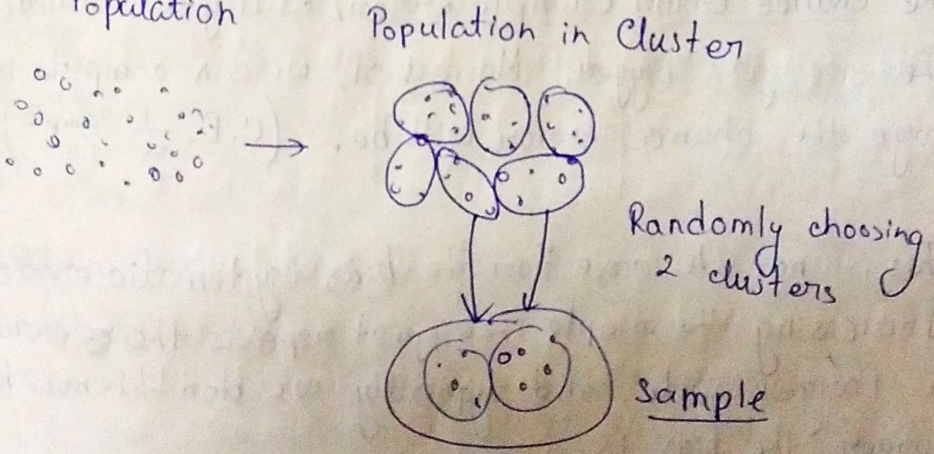
Eg:-



④ Cluster Sampling :-

A Probability sampling method in which we divide a population into cluster such as districts or schools and then randomly select some of these clusters as our sample.

Eg:-



(12) Multistage Sampling:

It is the taking of samples in stages using smaller and smaller sampling units in each stage.

Eg: Selection of households for the entire country.

1st stage unit → states
2nd stage Units → districts
3rd stage Units → Villages
: etc.

* Bias and Chance Error:

Since the sample is drawn at random, the estimate will be different from the parameter due to chance error.

Drawing another sample will result in a different chance error.

estimate = parameter + bias + chance error

- The chance error (sample error) will get smaller as the size gets bigger. Moreover, we can compute how large the chance error will be. (C.F. $\frac{1}{\sqrt{\text{Sample Size}}}$)

- This is not the case for the bias (systematic error): Increasing the sample size just repeats the error on a larger scale, and typically we don't know how larger the bias is.

* Observational Studies:-

- ~~This is~~ An observational study: I
- It measures outcomes of interest and this can be used to establish association.
- But association is not causation, because there may be compounding factors ~~such~~ to ~~alter~~ cause the association indirectly.

* Randomized Control Experiments / Trials (RCT):-

RCT are prospective studies that measure the effective-ness of a new ~~in the~~ intervention or treatment.

Although no study is likely on its own to prove casuality, randomization reduces bias and provides a rigorous tool to examine ~~cause~~ cause-effect relations b/w an intervention and outcome.

⇒ The Logic of RCT :-

① It makes the treatment group similar to the control group. Therefore influences other than the treatment operate equally on both groups, apart from differences due to chance.

② It allows to assess how relevant the treatment effect is, by calculating the size of chance effects when comparing the outcomes in the 2 groups.

* Standard Error

Note:-

There are 3 common error bars:-
1. Standard deviation
2. Standard error
3. Confidence Interval

The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation.

(# The Standard Error of the mean is the standard deviation of all the sample means from the actual population mean
(How far off from μ will $\bar{x}_n$ be?))

The square root Law:-

The accuracy of $\bar{x}_n$ as an estimator of μ (S.E.) is inversely proportional to the square root of the sample size n.

$$SE(\bar{x}_n) \propto \frac{1}{\sqrt{n}}$$

- The square root law is key for statistical inference:

$$SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$$

→ The importance of the square root law is twofold:-

1. It shows that the SE becomes smaller if we use a larger sample size n, We can use the formula to determine what sample size is required for a desired accuracy.
2. The formula for the SE does not depend on the size of the population but only depends on the size of the sample

→ Relation with Summation:-

The sum and the average are related by, $S_n = n\bar{x}_n$

Similarly,

$$E(S_n) = n\mu \quad ; \quad SE(S_n) = \sqrt{n}\,\sigma \qquad \left| \because SE = \frac{\sigma}{\sqrt{n}} \right.$$

(Expected Value)

So, the variability of the sum of n draws increases at the rate of $\sqrt{n}$.

→ Relation with Percentages:-

Then, ~~SE ∝ SE ∝ 1/∝~~

$$S.E. \propto \sqrt{n}$$

$$\therefore \quad S.E.(S_n) \propto \sqrt{n}$$

Percentages:- $E(\text{percentages of 1s}) = \mu \times 100\%$

$$SE(\text{percentages of 1s}) = \frac{\sigma}{\sqrt{n}} \times 100\%,$$

where we label people as 1 and 0.

Eg:- 1 1 0 0 1 then, sum = 3 = S_n

* The Law of Large Numbers:-

The Law of ~~Leg~~ Large numbers states that an observed sample avg. from a large sample will be close the true population avg. and that it will get closer the larger the sample.

Can be justified as according to the square root law:-
$SE. \frac{1}{\alpha} \sqrt{n}$

The Law of Large Numbers applies:-
~~for:~~

① for avg. and therefore percentages, but not for sums as their $SE(S_n)$ increases. $[SE(S_n) \propto \sqrt{n}]$

② for sampling with replacement from a population, or for simulating data from a probability histogram.

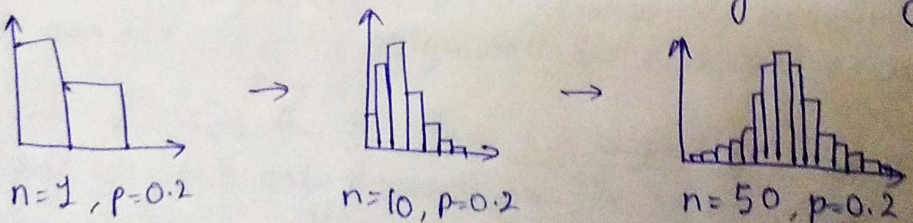
* The Central Limit Theorem:-

The Central Limit Theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, Regardless of the population's distribution.

~~The sample size o~~ The more the skewed the population distribution (histogram) is, the larger the required sample size n.

If there is no strong skewness then $n \geq 15$ is sufficient
otherwise $n \geq 30$ or $n \geq 40$ is sufficient in most cases

Eg:- Binomial distribution (with winning probability, p=0.2)



* Normal Approximation:-

Finding ~~percentages~~ areas under the ~~normal curve~~ normal curve is called normal approximation.

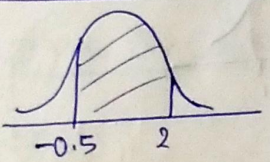
Eg:- Percentage of people with height b/w 67.4 in & 71.9 in?
$\mu_n = 68.3$ & $\sigma_n = 1.8$ in

→ Steps of Normal Approximation:-

① Standardize:

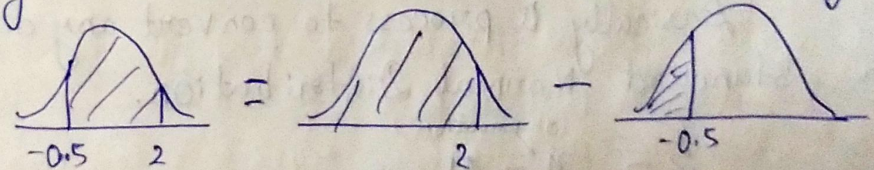
$$\frac{67.4\text{ in} - 68.3\text{ in}}{1.8\text{ in}} = -0.5 \quad ; \quad \frac{71.9\text{ in} - 68.3\text{ in}}{1.8\text{ in}} = 2$$

② Mark the area under the curve:-



③ Write the desired area in a form that can be computed by software:-

· Typically we can look up the area to the left of a given value.



④ Use software or a table to find these values:

$97.7\% - 30.9\% = 66.8\%$

→ Normal approximation to the binomial :-

We can approximate binomial probabilities using normal approximation.

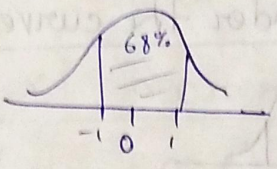
To standarize, subtract off np and then divide by $\sqrt{np(1-p)}$

Eg :- To find P(at most 12 small prizes)

We 1st standarize: $\frac{12 - np}{\sqrt{np(1-p)}}$ → P(win a small prize)=0.1

* Standard Normal Distribution :-

It is the Normal Distribution with $\mu = 0$ & $\sigma = 1$



* Standardization :-

Standardization is a scaling method where the values are centered around mean with a unit S.D (σ).

Basically a process to convert any distribution to a Standard Normal Distribution.

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

(z-score) → z

x ← (observation)

μ ← (Mean)

σ ← (S.D.)

* Z-Score:-

Z-score is a statistical measurement that describes a value's relationship to the mean of a group of values.

Z-score is measured in terms of standard deviations from the mean.

Eg:- $Z = 2.5$, means 2.5σ above μ

$Z = -1$, means 1σ below μ

→ Z-score Applications:-

(1) Standarization.

(2) Compare scores between different distributions. (After standarization)

(3) Eg:- Indian Cricket Team

2020	2021
$\mu = 181$	$\mu = 182$
$\sigma = 12$	$\sigma = 5$
Final Score = 187 runs	Final Score = 185 runs

$$Z_{2020} = \frac{187 - 181}{12} = 0.5$$

$$Z_{2021} = \frac{185 - 182}{5} = 0.6$$

∴ $Z_{2021} > Z_{2020}$, Cricket Team did better in 2021.

* Confidence Interval:-

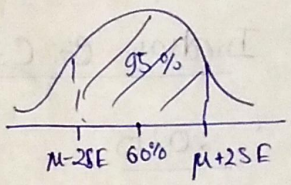
Confidence Interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

Eg:- If a point estimation is generated from a statistical model of 10 samples. with 95% confidence interval of 9.5 - 10.5, it can be inferred that there is a 95% probability that true value falls within that range.

- Confidence interval measure the degree of uncertainty or certainty in a sampling method.

Eg:- If we take a normal curve with $\mu = 60\%$ &

$$SE = \frac{\sigma}{\sqrt{1000}} = 1.6\%$$



Then, 95% chance that the sample percentage is no more than 2SEs way from μ, [54.8%, 61.2%]

→ Confidence Interval via C.L.T.

A confidence interval is centered at of plausible values for a population parameter μ.

Usually confidence interval is centered at an estimated average.

Since the central limit theorem applies for averages, the confidence interval has a simple form:

estimate ± Z x S.E.

⇒ Estimating the SE with the bootstrap:

SE is the standard error of the estimate.

$$SE = \frac{\sigma}{\sqrt{n}}$$

(But we don't Know σ)

The bootstrap principle states that we can estimate σ by its sample version 's' and still get an approximately correct confidence interval.

$$SE = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

⇒ Margin of error:-

The width of the confidence interval is determined by ~~z*SE~~. z^*SE, which is called margin of error.

It represents the degree of uncertainty that a survey result might have.

→ ~~We~~ We can decrease the margin of error by:

- Increasing Sample size (n). [We need $4n$ to cut the MOE to $\frac{MOE}{2}$]
- Decrease Confidence Interval.

(Trade off b/w Precision (MOE) and the Confidence Interval)

* Monte Carlo Method or (Simulation):-

The Monte Carlo Method is a probabilistic method than can include an element of uncertainty or randomness in its prediction.

Monte Carlo simulation provides multiple possible outcomes and the probability of each from the large pool of random data samples.

⇒ In most cases the general ~~problem where to~~ interest is to find a unknown parameter θ of a population.

→ We estimate θ with a statistic (estimator) $\hat{\theta}$ which is based on a sample of n. observations $X_1, \ldots, X_n$ drawn at random from the population:

$$\hat{\theta} = \text{avg of sample} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

And $\hat{\theta}$ tends to be close to the uncomputable population mean θ, even for ~~so~~ $n=100$.

OR.

• Monte Carlo Method:

→ We approximate a fixed quantity θ by the avg of independent random variables that have expected value θ.

→ By the Law of large numbers, the approximation error can be made arbitrarily small by using large sample size (n).

Monte Carlo Method for S.E.:

The Monte Carlo Method can also be used for more involved quantities, like S.E.

- The process ↓
  - → Get many (say 1000) samples of 100 observations each.
  - → Compute $\hat{\theta}$ for each sample, resulting in 1,000 estimates $\hat{\theta}_1, \ldots, \hat{\theta}_{1000}$
  - → Compute the standard deviation of these 1000 estimates.

$$s(\hat{\theta}_1, \ldots, \hat{\theta}_{1000}) = \sqrt{\frac{1}{999} \sum_{i=1}^{1000} (\hat{\theta}_i - \text{avg}(\hat{\theta}_i))^2}$$

The Law of Large numbers is also applied here, making $s(\hat{\theta}_1, \ldots, \hat{\theta}_{1000}) \approx SE(\hat{\theta})$

*Note:-

Monte Carlo Method or Simulation only works when we can sample as much as we want.

→ Plug-in Principle:-

The plug-in principle is a technique used to approximately compute or to estimate a feature of probability distribution (Eg. S.E., E.V., σ) that cannot be computed exactly. (Based on a sample distribution)

* The Bootstrap principle :-

- The basic idea of Bootstrapping is that inference about a population from sample data (sample → population) can be modeled by resampling the sample data and performing inference about a sample data (resampled → sample)

- The bootstrap uses the plug-in principle and the Monte Carlo Method to approximate quantities such as $SE(\hat{\theta})$.

- Simulating a bootstrap sample $X_1^*, \ldots X_n^*$ means that we draw n times with replacement from $X_1, \ldots, X_n$.

  Eg :- from ① ② ③ ④,

  → ① ② ③ ③
  → ① ② ① ③
  → ④ ④ ③ ①
  etc.

* The Bootstrap consists of 2 steps :-

① Draw N bootstrap samples and compute $\hat{\theta}^*$ for each bootstrap sample :

$X_1^{*1}, \ldots X_n^{*1} \rightarrow \hat{\theta}_1^*$

$\vdots$

$X_1^{*N}, \ldots X_n^{*N} \rightarrow \hat{\theta}_N^*$

② Use $\hat{\theta}_1^*, \ldots, \hat{\theta}_N^*$ to approximate the quantity of interest.

Eg :- We approximate $SE(\hat{\theta})$ by the standard deviation of $\hat{\theta}_1^*, \ldots, \hat{\theta}_N^*$

→ Parametric Bootstrap:-

In this method we assume that the observed data is drawn from a specific parametric distribution (eg:- Normal Distribution, exponential distribution, etc). Then, we estimate the parameters of this distribution from our sample data, and generate new ~~data~~ samples from the assumed distributions.

→ Non Parametric Bootstrap:-

Non Parametric Bootstrap makes no specific assumptions about the underlying distribution of the data. It directly resamples observed data.

* Statistical Inferences :-

Statistical inference is a fundamental concept in statistics that involves drawing conclusions or making predictions about a population based on a sample of data.

It is a way of using sample data to make inferences or generalizations about a larger population from which the sample is drawn.

→ Statistical Inference can be approached from a Bayesian or Frequentist ~~pre~~ perspective.

⇒ Bayesian Inference uses prior knowledge and probability distributions to update beliefs,

while, Frequentist Inference relies solely on sample data.

The choice b/w these approaches depends on the context and avaiable information.

→ There are mainly 2 types Statistical Inferences:

(1) ~~Point~~ Estimation
- (ii) Point Estimation
- (ii) Interval Estimation

(2) Hypothesis Testing

* Estimation:-

A process in which we obtain the value of unknown population parameters with the help of sample data.

→ Point Estimation:-

Point estimation involves providing a single value, often denoted as a point estimate, as the best guess for the population parameter.

(Eg: $\bar{x}$, s^2, & s)

→ Interval Estimation:- (Confidence Interval)

Interval estimation provides a range of values within which the population parameter (Eg: μ) is likely to fall.

It is expressed as a Confidence Interval.

* Maximum Likelihood Estimator (MLE):-

[Watch StatQuest for more clarity]

(Part of Point Estimation)

The MLE is a statistical method used to estimate the parameters of a statistical model. It is a powerful and widely used technique in statistics and Machine Learning.

The key idea of MLE is to find the set of parameter values that maximizes the likelihood function, measures how well the observed data fits the model, or to find the parameter values that make the observed data the most probable under the assumed statistical model.

$$MLE(\hat{\theta}) = \arg\max_{\theta} \mathcal{L}(\theta/\text{data}), \text{ where,}$$

- $\hat{\theta}$ is Maximum Likelihood estimate of the parameter θ.
- $\mathcal{L}(\theta/\text{data})$ is the likelihood function.

⇒ General steps to find MLE:-

① Define the Likelihood function.

② Take the natural logarithm for the equation.
(Makes it easier to differentiate)

③ Differentiating the log likelihood.

④ Set the score to zero and ⑤ Solve for parameter.

Additionaly,

⑥ We can check for 2nd derivative of log-likelihood to determine the nature of the critical point.
(Maxima, Minima & Saddle Point)

⇒ MLE for Bernoulli Distribution:- $x_1 x_2 x_3 x_4 x_5 \dots x_n$
00000...0

$X = (x_1, \dots, x_n)$

$X_i \sim \text{Bernoulli}(p)$

① $l(p, x) = p^{\sum_{i=1}^{n} x_i} \times (1-p)^{n - \sum_{i=1}^{n} x_i}$

② $\log(l(p,x)) = \log\left(p^{\sum_{i=1}^{n} x_i} \times (1-p)^{n - \sum_{i=1}^{n} x_i}\right)$

$= \log(p)\left(\sum_{i=1}^{n} x_i\right) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p)$

③ $\frac{d}{dp}(\log(l(p,x))) = \frac{d}{dp}\left[\left(\sum_{i=1}^{n} x_i\right)\log(p) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p)\right]$

④ & ⑤ $= \frac{\sum_{i=1}^{n} x_i}{p} + \frac{n - \sum_{i=1}^{n} x_i}{(1-p)}(-1) = 0$

④ & ⑤ $\Rightarrow \sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i - np + p\sum_{i=1}^{n} x_i = 0$

$\Rightarrow np = \sum_{i=1}^{n} x_i$

$\Rightarrow \boxed{p = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}}$

General form of finding MLE:

Let θ be the population Parameter:- PDF $\downarrow$ $f(x_i)$

(1) $L(\theta) = f(x_1)\cdot f(x_2)\cdots(f(x_n))$

$= \prod_{i=1}^{n} f(x_i)$

(2) $\log(L(\theta)) = \sum_{i=1}^{n} \log(f(x_i))$

$\because \log(f(x_1)\cdot f(x_2)\cdots f(x_n))$
$= \log(f(x_1)) + \log(f(x_2)) + \ldots + \log(f(x_n))$

(3) $\frac{d}{dp}\log(L(\theta)) = 0$

and ~~sdor~~ solve for θ.

→ MLE for Normal Distribution:

$X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$

Note:- $\sigma \neq 0$

(1) $L(\mu, \sigma, X) = \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$

$= \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}}$

(2) $\log(L(\mu,\sigma,x)) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}$

$\Leftarrow \boxed{l(\mu,\sigma)}$

for $\frac{\partial}{\partial\mu}$

(3) $\frac{\partial l(\mu,\sigma)}{\partial\mu} = -\frac{1}{2}\frac{\sum_{i=1}^{n} 2(x_i-\mu)(-1)}{\sigma^2}$

(4)&(5) $= \frac{1}{\sigma^2}(\sum_{i=1}^{n} x_i - n\mu) = 0$

$(\because \sigma \neq 0)$

$\Rightarrow \sum_{i=1}^{n} x_i = n\mu$

$\Rightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

Sample mean

for $\frac{\partial}{\partial\sigma}$

(3) $\frac{\partial}{\partial\sigma} l(\mu,\sigma) = -\frac{n}{\sigma} - \frac{1}{2}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)\frac{1}{\sigma^3}(-2)$

(4) & (5) $= \frac{-n}{\sigma} + \left(\sum_{i=1}^{n}(x_i-\mu)^2\right)\frac{1}{\sigma^3} = 0$

$\Rightarrow -n + \left(\sum_{i=1}^{n}(x_i-\mu)^2\right)\frac{1}{\sigma^2} = 0$

$\Rightarrow \left(\sum_{i=1}^{n}(x_i-\mu)^2\right)\frac{1}{\sigma^2} = n$

$\Rightarrow \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} = \sigma^2$

$\therefore \sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{n}}$

or $\sigma = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$

$\because$ By estimate $\mu = \bar{x}$

*Types of approaches for Statistical Inference:

Frequentist

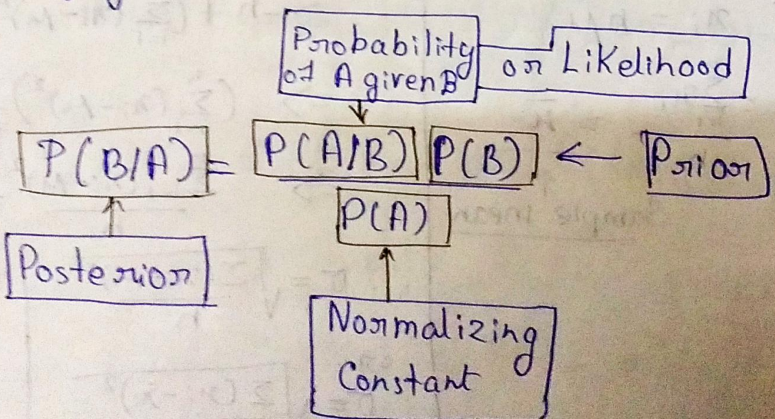
1. Probabilities represent long term frequencies
2. Parameters are fixed (but unknown) constants, so we can make probability statement about them.
3. Find the model that better explains the observed data.
4. Statistical procedures have well-defined long run frequency properties

Bayesian

1. Probabilities represent a degree of belief, or certainty.
2. We make probability statements about parameters, even if they are constant (beliefs)
3. Update beliefs on the model based on the observed data.
4. We make inferences about a parameter θ by producing a probability distribution of θ.

Basically, Frequentists imply that probabilities are objective properties of ~~appearance of a certain event in~~ the real world and that the parameters of the distributions are fixed constants. So, Frequentist interpret observed data as the sample of an unknown distribution.

While, Bayesians interpret probabilities as a degree of belief. The goal to Bayesians is to update this belief as we gather new data. Our result will be an updated probability distribution for the parameter we are trying to infer.



* Maximum a Posteriori (MAP):

Maximum a Posteriori (MAP) is an inference technique used to estimate the most likely value of an unknown variable or set of variables on the observed data and prior probability distribution.

It is used in the context of Bayesian Inference and estimates the value that maximizes the posterior probability distribution.

$$MAP(\theta) = \arg\max_{\theta} P(\theta|D), \text{ where,}$$

θ = Parameter of interest

$P(\theta|D)$ = Posterior probability of parameter given D.

$P(\theta)$ = Prior probability of the parameter

$P(D|\theta)$ = likelihood of observing data given the parameter

OR

MAP: $\hat{\theta} = \arg\max_{\theta} f_{\theta|X=x}$

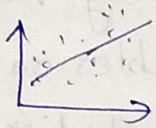
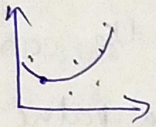
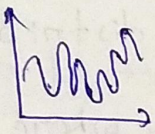
Since, MAP considers both the data and prior information to provide more data, while MLE only considers the data.

When there is no informative prior, MAP is same as MLE

* Regularization and MAP:-

Using L_2 Regularization
$L_2 = a^2 + b^2$ ... log
$y = ax^n + bx^{n-1} + c$

Scenario:-

		Loss	Equation	Penalty
Underfit		10	$y = 4x + 3$	$L_2 = 4^2 = 16$
Good fit		2	$y = 2x^2 - 4x + 5$	$L_2 = 2^2 + (-4)^2 = 20$
Overfit		0.1	$y = 4x^{10} - 9x^8 - 2x^6 + 3x^5 - 6x^4 - 10x + 4$	$L_2 = 4^2 + (-9)^2 + (-2)^2 + 3^2 + (-6)^2 + (-10)^2 = 246$

New Logs (Loss + Penalty)	
26	
22	← Best fit
246.1	

→ Regularization Term:-

Model: $y = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$

Log-loss : ll

L_2 Regularization Error : $a_n^2 + a_{n-1}^2 + \ldots + a_1^2$

Regularization parameter: λ

Regularization Error : $ll + \lambda (a_n^2 + a_{n-1}^2 + \ldots + a_1^2)$

- Take MAP into account, we ~~use~~ can use the prior information like simplicity of the model assuming simpler models are more likely to get selected.

Then, Likelihood of the Model,

→ P(Model 1) $= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_1^2}$

$a_1 x + b$

→ P(Model 2) $= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_2^2}$

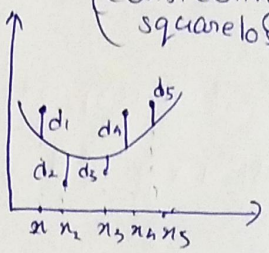
$a_1 x^2 + a_2 x + b$

→ P(Model 3) $= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_i^2}$

$a_1 x^{10} + \ldots + a_{10} x + b$

$y = a_2 x^2 + a_1 x + b$

(Considering square loss)



Then,

$P = P(\text{Data} \mid \text{Model}) \times P(\text{Model})$

$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2)} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(a_1^2 + a_2^2)}$

~~Then, to maximize P~~ (Now finding the MAP)

$\Rightarrow \log P = \frac{-1}{\sqrt{2\pi}} (d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2) - \frac{1}{\sqrt{2\pi}} (a_1^2 + a_2^2) = 0$

$\Rightarrow \underbrace{d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + a_1^2 + a_2^2}_{\downarrow \text{Minimize}} = 0$

- So, to maximize P, we need to minimize

$\underbrace{d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2}_{\downarrow}$ + $\underbrace{a_1^2 + a_2^2}_{\downarrow}$

Square ~~error~~ loss

Complexity of model (Lesser coefficients of x)