

Probability and Statistics

Variables:

Random Variable: For a set $X = \{x_1, x_2, x_3, \dots, x_n\}$, if the probability of $P(x_i) = P(x_j)$ for all $i \neq j$, then the output is called a random variable.

A r.v. which has a finite set of outcomes or values is called a discrete random variable.
eg: Rolling of a dice

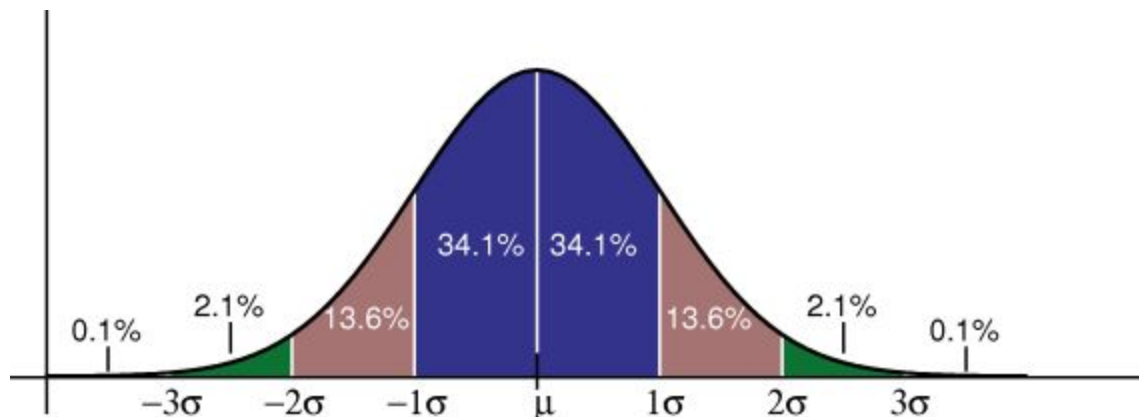
A r.v. which can take any real value is called a continuous random variable.
eg: The probability of height of a group of students

An observation point that is distant from other observations is called an outlier.

Note: Mean and variance gets corrupted by an outlier, hence we use median and median absolute deviation.

(Median Absolute Deviation is the summation of all the absolute differences of the median from the individual points)

Gaussian/Normal Distribution:




If X is a continuous r.v. that has a PDF like that of a bell shaped curve, then we say X has a distribution which is a Gaussian Distribution.

Mean and Variance are the parameters of Gaussian Distribution.

Variance is the spread of the curve. So, if it is small, the curve is going to be steeper.
 μ or the mean is generally the peak of the distribution.

CDF of the Gaussian Distribution is the integral of it's PDF which can also be written as:

$$\frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x - \mu}{\sigma \sqrt{2}} \right) \right]$$



The Normal Distribution: as mathematical function (pdf)

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2}$$

Note constants:

$\pi=3.14159$

$e=2.71828$

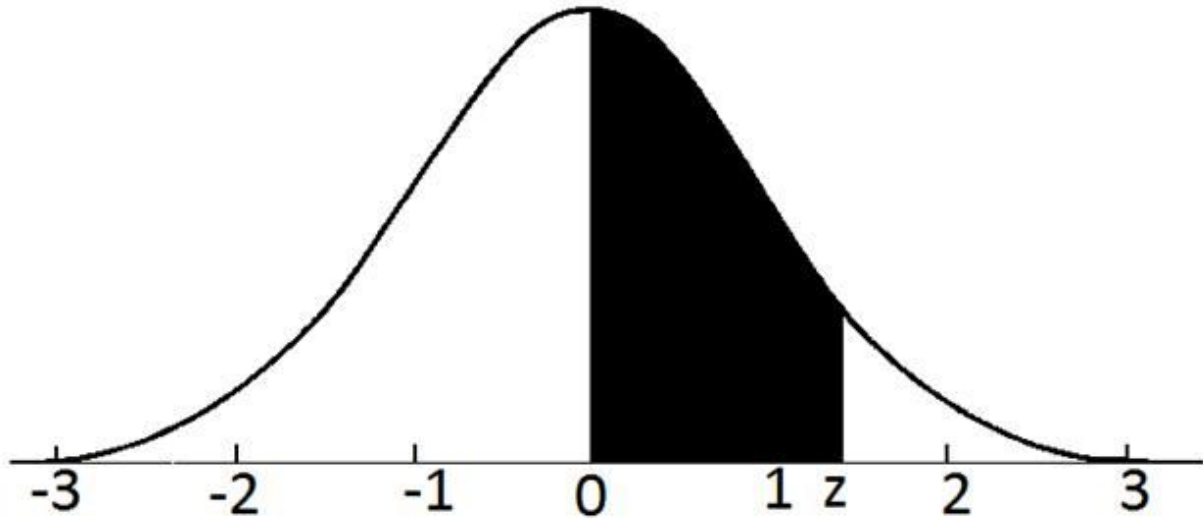
This is a bell shaped curve with different centers and spreads depending on μ and σ

Standard Normal Variate (Z):

A **standard normal variate** is a normal variate with mean $\mu=0$ and standard deviation $\sigma=1$ with a probability density function is:

$$f(z) = \left(\frac{1}{\sqrt{2\pi}} \right) e^{-z^2/2} \quad -\infty < z < \infty$$

It is denoted by Z.



For a given variable $X = [x_1, x_2, x_3, \dots, x_n]$ if the mean is m and the standard deviation is s and X is a Gaussian Distribution, then to make $X \sim N(0, 1)$ we compute,
 $z_i = (x_i - m)/s$

This is convenient for understanding the density of elements in a Gaussian curve with the 68-95-99 rule.

Central Limit Theorem:



Central Limit Theorem

3 important results for the distribution of \bar{X}

- Mean Stays the same

$$\mu_{\bar{X}} = \mu$$

- Standard Deviation Gets Smaller

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- If n is sufficiently large, \bar{X} has a Normal Distribution

34

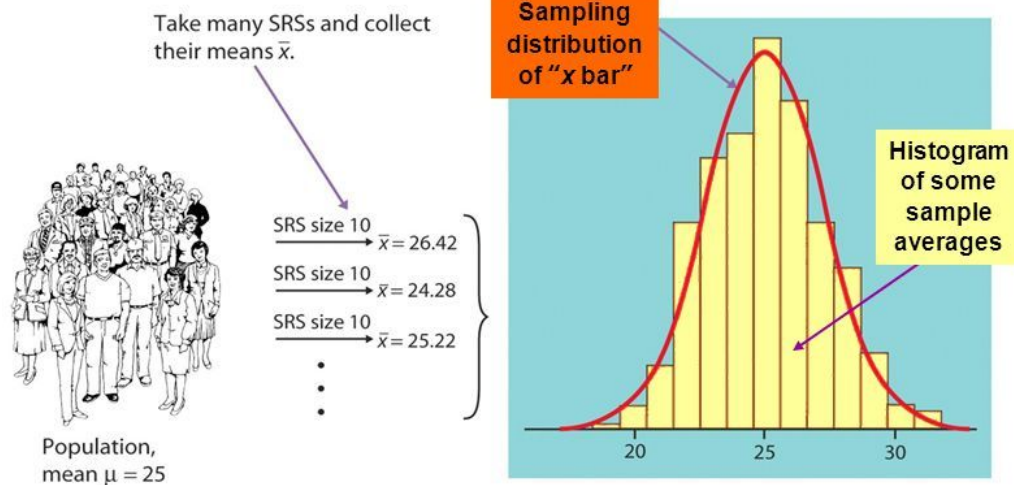
If we take m samples with n data points from a population distribution X which may not be Gaussian, and we calculate the mean of every samples, then, CLT states that, the distribution of the mean of the samples would be a Normal distribution with the population mean as it's mean (roughly) and the standard deviation equal to (population variance/the number of data points in each sample) which can be written as, $N(m, s^2/n)$ as $n \rightarrow \infty$

Note: As a rule of thumb, if the sample space $n \geq 30$, then the resulting distribution of the sample means is a Gaussian Distribution

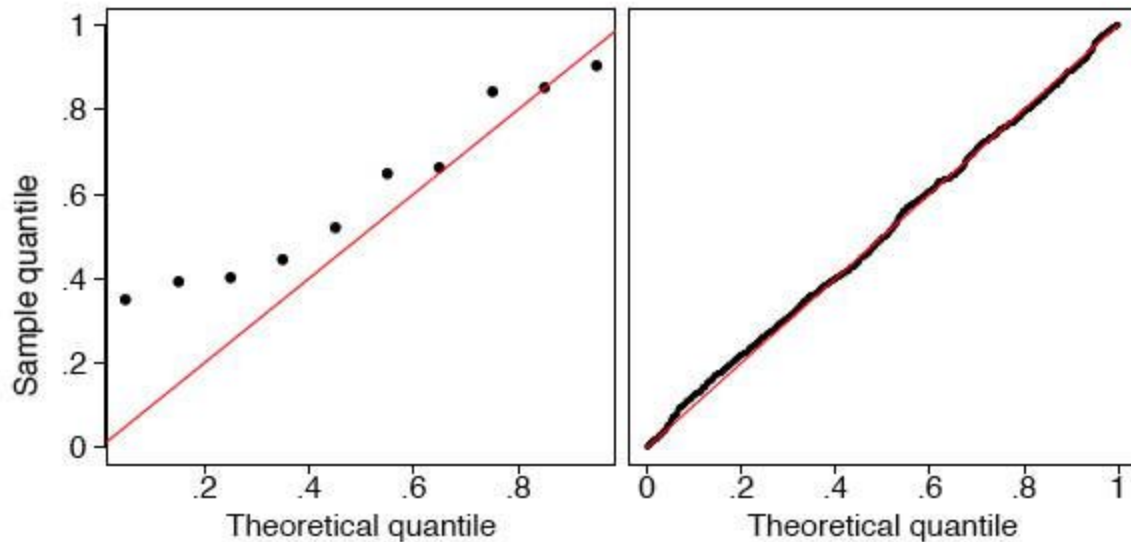
Sampling distribution of the sample mean

We take many random samples of a given size n from a population with mean μ and standard deviation σ .

Some sample means will be above the population mean μ and some will be below, making up the sampling distribution.



Q-Q Plot:



First, sort the given r.v. in ascending order.

Then, generate a theoretical Quantile of the test distribution and sort it.

Then, graph the respective elements matching each of the given r.v. and the theoretical distribution.

If the graph generates a straight line, then the 2 distributions are statistically same, else not.

Power Law Distribution:

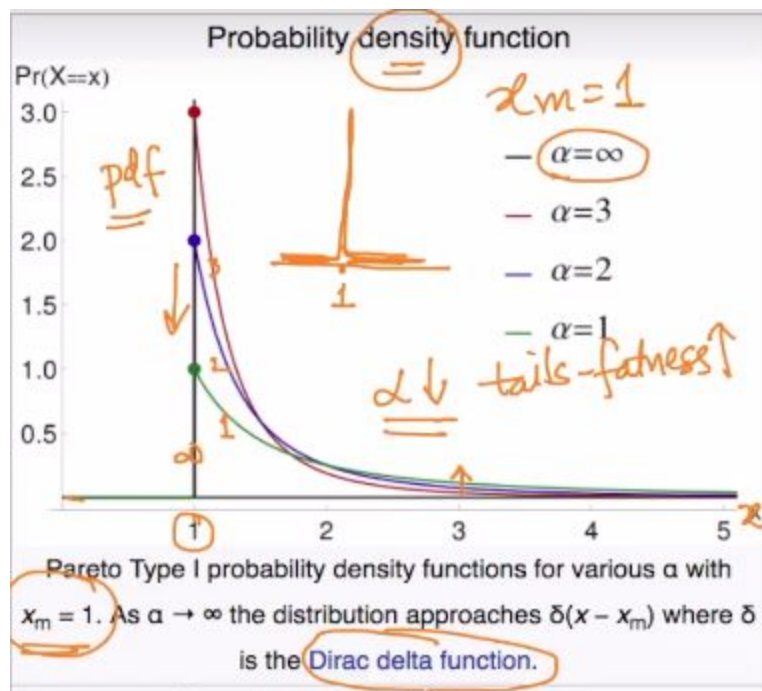
The power law (also called the scaling law) states that a relative change in one quantity results in a proportional relative change in another. The simplest example of the law in action is a square; if you double the length of a side (say, from 2 to 4 inches) then the area will quadruple (from 4 to 16 inches squared). A power law distribution has the form $Y = k (X^\alpha)$, where:

X and Y are variables of interest,

α is the law's exponent,

k is a constant.

(It roughly follows the 80-20 rule, i.e., 80% of the points lie in the 20% of the region of the distribution)



Other examples of phenomena with this type of distribution:

Distribution of income,

Magnitude of earthquakes,

Size of cities according to population,

Size of corporations,

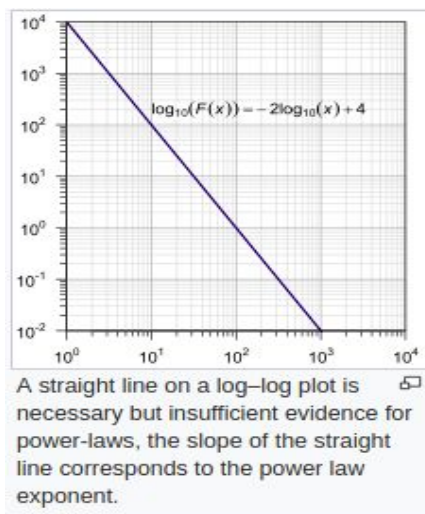
Trading volumes on the stock market,

word frequencies.

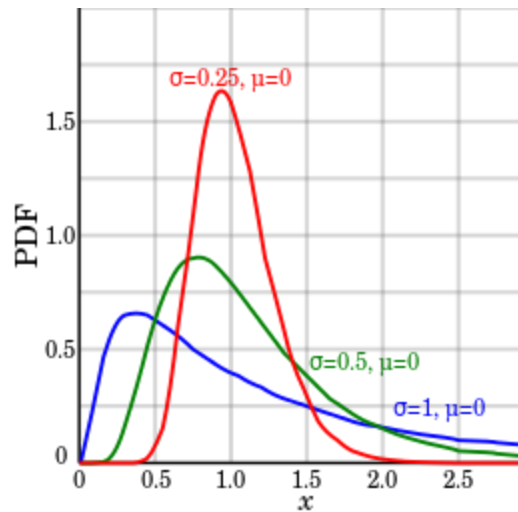
NOTE: To check whether a distribution is Pareto distr, use log-log plot i.e., plot $\log(X)$ vs $\log(Y)$.

(Refer Fig 2)

We can also use Q-Q plot.



Log-normal Distribution:



A random variable is log-normally distributed if its logarithm is normally distributed.

A lognormal (log-normal or Galton) distribution is a probability distribution with a normally distributed logarithm.

In other words, if $X \sim \text{lognormal}(\mu, \sigma)$
it implies that $Y = \log(X)$ is normally distributed, i.e., $Y \sim N(\mu, \sigma^2)$.

The following phenomenon can all be modeled with a lognormal distribution:

Milk production by cows.

Lives of industrial units with failure modes that are characterized by fatigue-stress.

Amounts of rainfall.

Size distributions of rainfall droplets.

The volume of gas in a petroleum reserve.

Covariance:

Covariance is the establishment of a relationship between two or more variables.

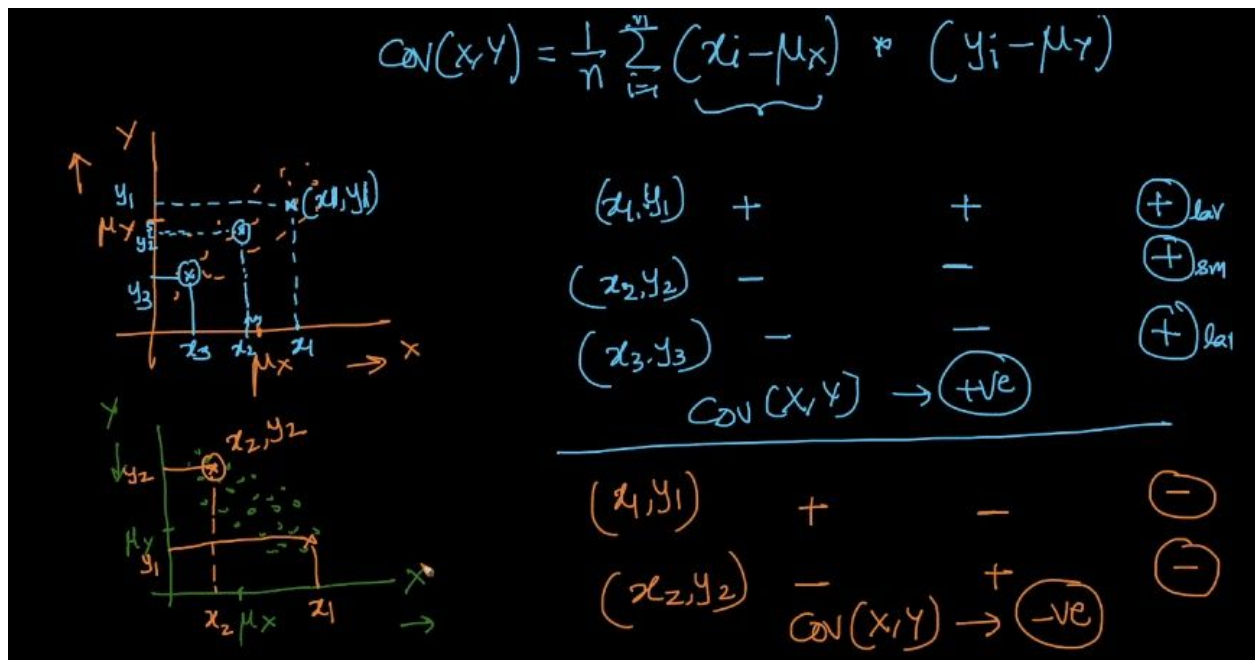
$$\text{COV}(X, Y) = \frac{1}{n} \sum_{i=1}^n \{x_i - \underline{\mu_x}\} * (\underline{y_i - \mu_y})$$

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x) * \underline{(x_i - \mu_x)}$$

$$\checkmark \text{COV}(X, X) = \text{Var}(X)$$

$$\begin{aligned} \text{COV}(X, Y) &= +ve & x \uparrow, y \uparrow \\ \text{COV}(X, Y) &= -ve & x \uparrow, y \downarrow \end{aligned}$$

NOTE: By changing the units of measure, co-variance may differ, which is a huge drawback.



Box-cox transformation:

Handwritten notes on a blackboard:

Pareto $\sim X: [x_1, x_2, \dots, x_n]$
Gaussian $\sim Y: y_1, y_2, \dots, y_n$

Conversion

① $\text{box-cox}(X) = \lambda$
 λ is labeled as $\lambda(\lambda)$ and λ is also written above the arrow pointing to the result.

② $y_i = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \text{if } \lambda = 0 \end{cases}$

It is used to convert a pareto distribution to a normal distribution.

Basically, for a distribution X , if we apply the function $\text{boxcox}(X)$ we get a result λ . Then we generate a distribution Y in accordance to the image above.

Pearson Correlation coefficient:

pearson correlation coeff: (p.c.c)

$$\rho_{x,y} = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$$
$$\sigma_x = \sqrt{\text{Var}(X)}$$

$x \uparrow, y \uparrow$, $\text{Cov}(X,Y)$ (+ve)
 $x \uparrow, y \downarrow$, $\text{Cov}(X,Y)$ (-ve)

Note: PCC is bound between +1 and -1

If $\rho=0$, then there is no relation between X and Y

The slope of the line doesn't matter

PCC cannot capture complex non-linear relationships

Spearman's rank-correlation coefficient:

Spearman rank-corr. coeff (γ)

$\rho_{x,y} \rightarrow$ linear relationship

$\gamma = \rho_{\underline{x}, \underline{y}}$

$\rho = 1 \leftarrow$ linear $x \uparrow y \uparrow$

$\rho = -1 \leftarrow$ linear $x \uparrow y \downarrow$

linear or not

linear or not

$\gamma = 1$

$\gamma = -1$

| | $\checkmark x$ | $\checkmark y$ | $\textcircled{\gamma_x}$ | γ_y |
|-------|----------------|----------------|--------------------------|------------|
| s_1 | 160 | 52 | 4 | 3 |
| s_2 | 150 | 66 | 2 | 4 |
| s_3 | 170 | 68 | 6 | 5 |
| s_4 | 140 | 46 | 1 | 1 |
| s_5 | 158 | 51 | 3 | 2 |

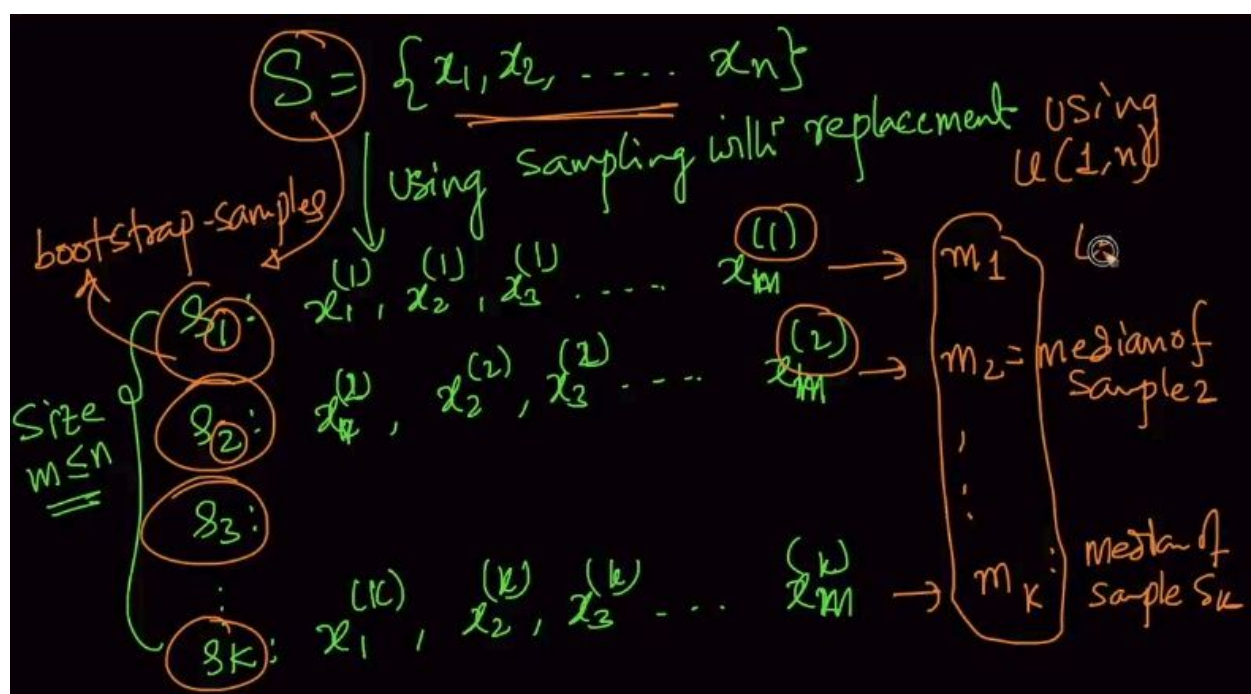
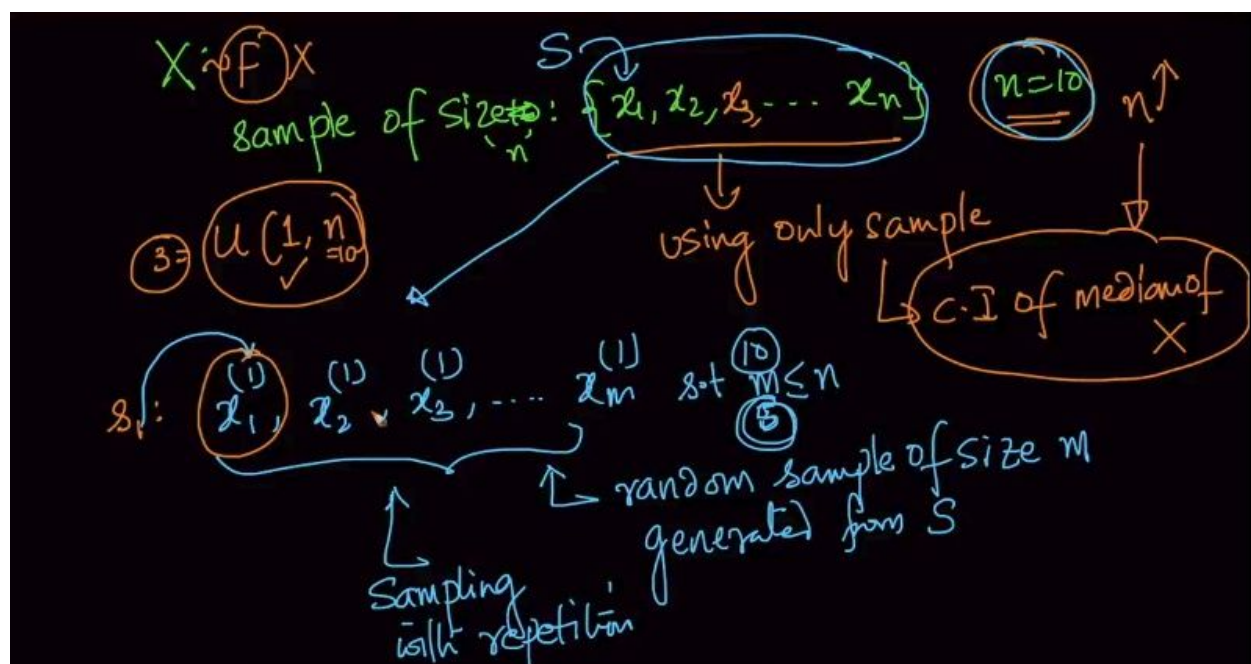
It is just the Rho of rank from Pearson correlation coefficient, which is the index of the sorted variables.

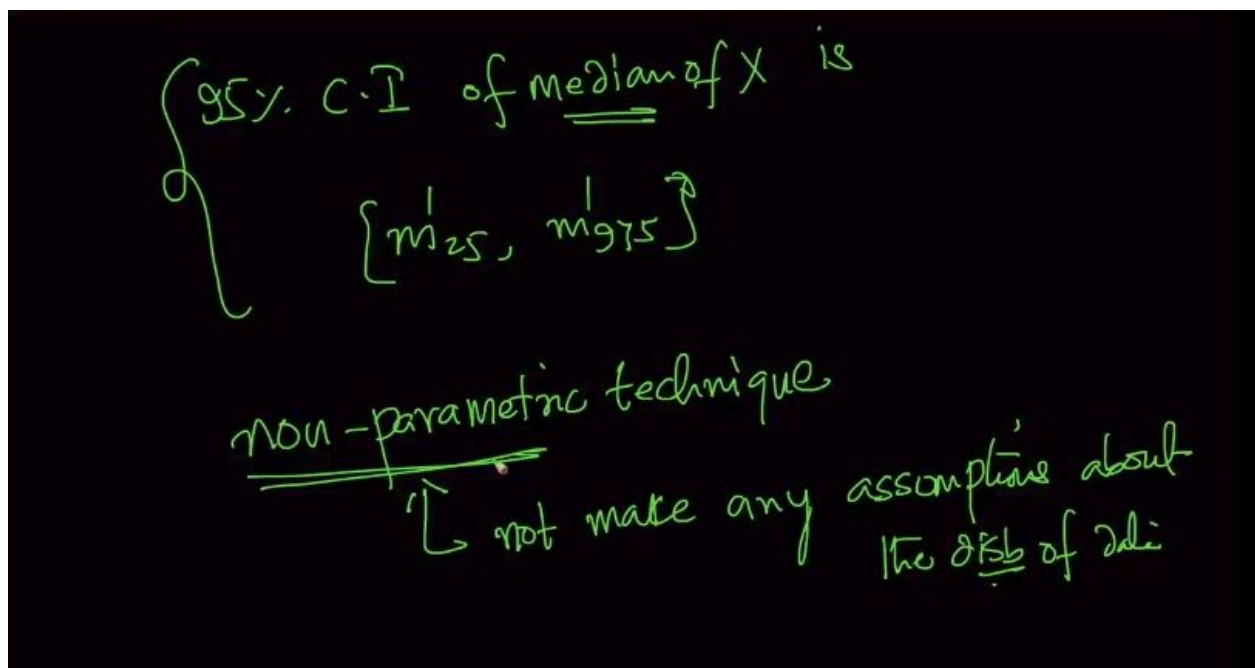
Confidence Interval using bootstrapping:

X is a random distribution of size n.

We take m elements from X into k Samples. (Repetition may occur in the same sample space but $m \leq n$)

Then, we sort and generate the objective from each sample.





How to Construct a Confidence Interval

There are four steps to constructing a confidence interval.

Identify a sample statistic. Choose the statistic (e.g, sample mean, sample proportion) that you will use to estimate a population parameter.

Select a confidence level. As we noted in the previous section, the confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.

Find the margin of error. If you are working on a homework problem or a test question, the margin of error may be given. Often, however, you will need to compute the margin of error, based on one of the following equations.

$$\text{Margin of error} = \text{Critical value} * \text{Standard deviation of statistic}$$

$$\text{Margin of error} = \text{Critical value} * \text{Standard error of statistic}$$

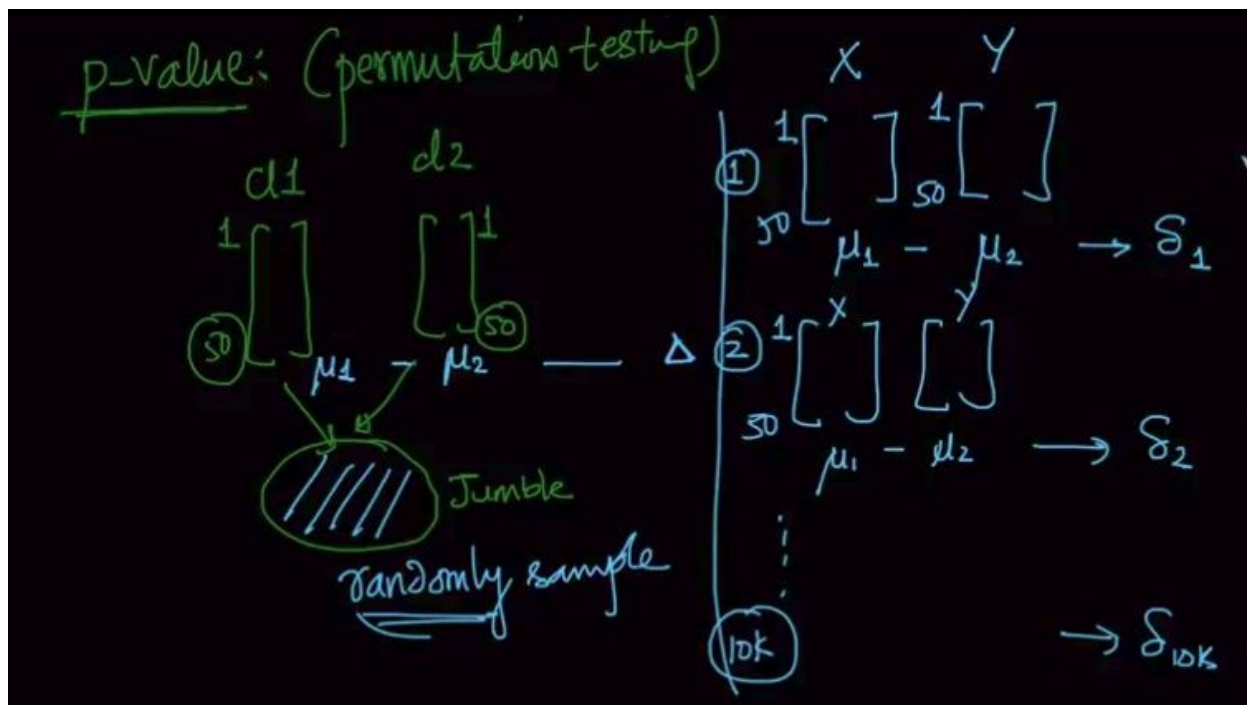
For guidance, see [how to compute the margin of error](#).

Specify the confidence interval. The uncertainty is denoted by the confidence level. And the range of the confidence interval is defined by the following equation.

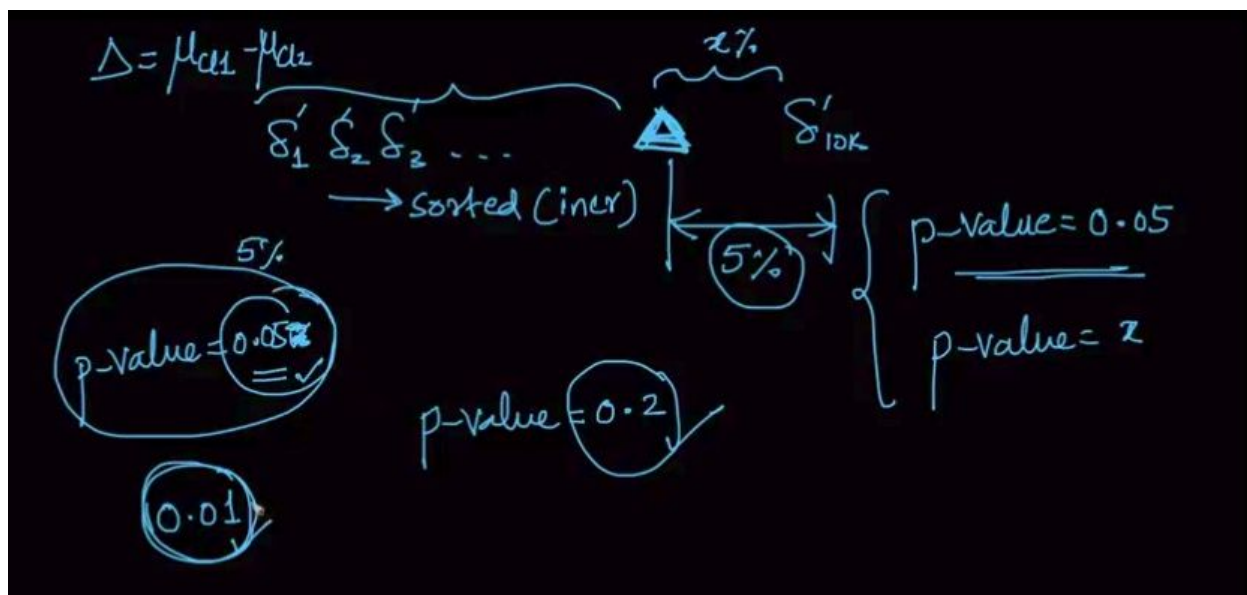
$$\text{Confidence interval} = \text{sample statistic} + \text{Margin of error}$$

Resampling and permutation test:

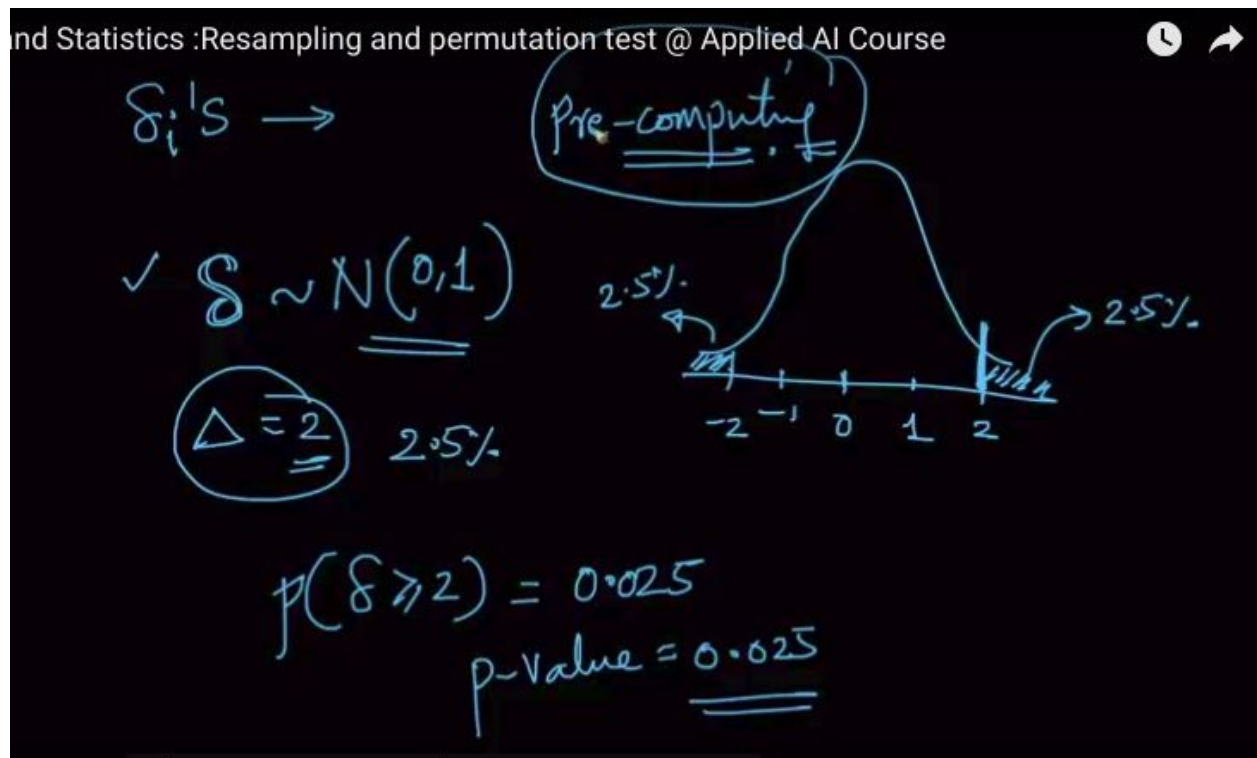
In permutation testing, we shuffle the two datasets of size n , or here, classes, into one pool. Then, we randomly select n samples into data sets (two here) and evaluate our statistics test into k deltas, with the initial statistics test being Δ .



Then, we sort the derived deltas and evaluate the position of Δ from the highest delta in the array. Here, the Δ is 5% below the max, hence, the p-value here is 0.05



In early days, when simulation was a thing of the future, a generalised distribution of deltas was established from where, the probability from the max furthestmost point was evaluated.



Hypothesis Testing:

In probability and statistics, this is a concept by which a statement is validated through a Proof by Contradiction.

We first start by choosing a test statistics.

Then we provide a Null hypothesis rejecting the test statistics.

We choose an Alternate hypothesis which is the complement of Null hypothesis.

Then we check the p-value or the probability of the test statistics, given the Null hypothesis is true.

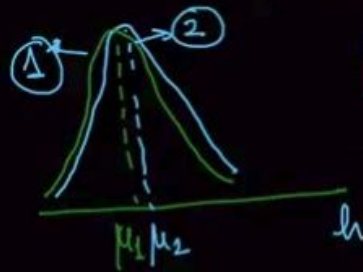
If,

p-value is closer to 1, we accept the Null Hypothesis or if p-value is closer to 0, we accept the Alternate Hypothesis.

p-value = Probability of observation given the assumption is true $[P(\text{obs}|\text{assumption})]$

Hypothesis testing:

(Q) Is there a diff in heights of students in cl 1 & cl 2?



| cl 1 | cl 2 |
|--------|--------|
| 1 160 | 1 162 |
| 2 152 | 2 156 |
| ... | ... |
| 50 148 | 50 182 |

① Choosing a test-statistic
 $(\mu_2 - \mu_1)$

μ_2 = mean hght of cl 2 students

μ_1 = " " cl 1

② Null hypothesis (H_0)

✓ H_0 : no-difference in μ_1 & μ_2

Alternative hyp (H_1): diff in μ_1 & μ_2

(PROOF BY CONTRADICTION)

③ p-value: prob. of obs $(\mu_2 - \mu_1)$ if null hyp is true.

assume H_0 is true.

accept H_0 ← if p-value = 0.9 ✓

⇒ prob of 10CM is 0.9 if H_0 is true

reject H_0 ← if p-value = 0.05 → 5% chance that 10CM if H_0 is true

cl1 cl2

✓50 ✓50

Chebyshev's Inequality:

If we don't know the distribution but we have a finite mean and non-zero and finite standard deviation, to know the percentage of point lying within the required standard deviations, we use the Chebyshev's Inequality.

Chebyshev's inequality:

finite mean $\neq \mu$

non-zero & finite std-dev $\neq \sigma$

don't know the dist

$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$

$X \geq \mu + k\sigma$
 $X \leq \mu - k\sigma$

$$P\left(\begin{array}{l} X \geq \mu + k\sigma \\ X \leq \mu - k\sigma \end{array}\right) \leq \frac{1}{k^2}$$

$$(n) \quad P(\mu - k\sigma < X < \mu + k\sigma) > 1 - \frac{1}{k^2}$$