Gaussian Distribution (or) Normal distribution

continuous

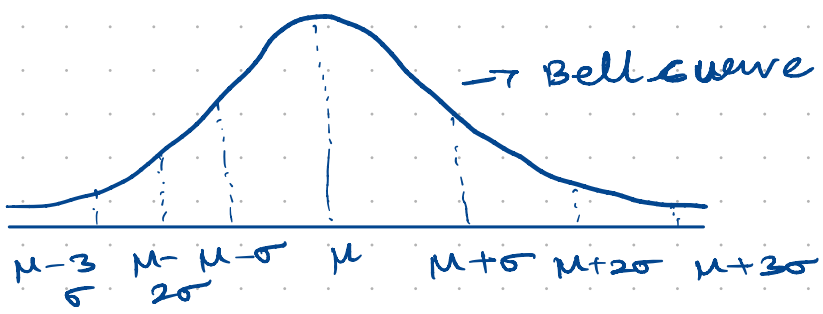
x is a R.V ; $x \in N(\mu, \sigma)$

mean, s.D

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

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

$$\sigma = \sqrt{var}$$



$P(\mu - \sigma \leq x \leq \mu + \sigma) = 68\%$

$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \simeq 95\%$

$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \simeq 99.7\%$

Log Normal Distribution:

$X \sim$ Log Normal Distribution

if $\ln(x)$ is normally distributed

$X = \{x_1, x_2, \ldots, x_n\}$

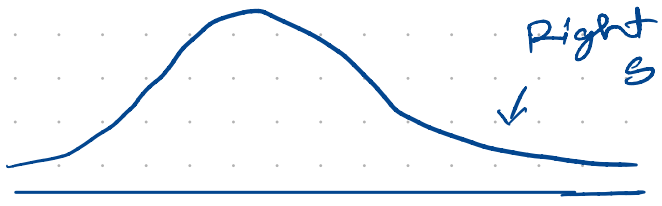
$\{\ln(x_1), \ln(x_2) \ldots \ln(x_n)\}$

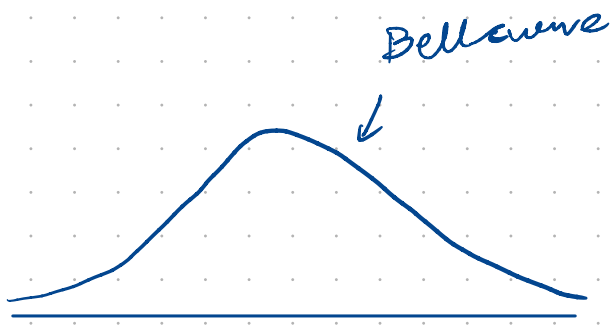
if this is normally distributed (formed a bell curve)

then X belongs to a Log Normal distribution.

Log Normal Distribution

Gaussian distribution





E.g: Income of the people.

Product comments (length of description)

If you have a data, which is not normally distributed & it's distb looks like it is skewed, then apply log to the data & now your new data will follow normal distribution

Covariance :-

Consider an example in which you have =

size of the house	price of the house
1500 sqm	100K $
1800 sqm	200K $
2000 sqm	300K $

Both are directly proportional -

$$cov(\underset{\downarrow}{size}, \underset{\downarrow}{price}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) * (y_i - \mu_y)$$

$x \quad y$

$$var(size) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)(x_i - \mu_x)$$

Relationship b/w variance & covariance is

$$cov(x, y) = var(x)$$ ← by the formula.

Covariance gives you an insight about

if $x \uparrow$ then $y \uparrow$ or $\downarrow$ by what %?

$x \uparrow \quad y \uparrow = +ve$

$x \uparrow \quad y \downarrow = -ve$

Central Limit theorem :-

Consider a R.V 'X' which may / may not belong to guassian distribution

$$X \neq N(\mu, \sigma^2)$$

Take a sample from the R.V X $\boxed{n \geq 30}$ Data is drawn randomly

Sample $S_1 = x_1, x_2, \ldots, x_{30} = \bar{X}_1$

Sample $S_2 = x_1, \ldots \ldots x_{30} = \overline{X_2}$

$\vdots$

$S_{50} = \ldots \ldots \ldots \ldots = \overline{X_{50}}$

Now take all the sample means & plot by using histogram (for ex).

$$\overline{X} = G.D\left(\mu, \frac{\sigma^2}{n}\right)$$

Proper defn of CLT:

If you have a population with μ & S.D σ & take sufficiently large random samples from the population with replacement, then the dist of sample means will be approx normally distributed.

Chebyshev's Inequality:

$X \sim G.D(\mu, \sigma)$, If a R.V follows a gaussian distribution, then you can make use of 68-95-99.7%.

If you have a R.V Y that does not follow a Gaussian distribution, then how can you find out:

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

Generalized formula is

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

If $k = 3$

$$P(\mu - 3\sigma < x < \mu + 3\sigma) > 1 - \frac{1}{9}$$

$> 8/9 \simeq 0.88 = 88\%$

Also be written in the form

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

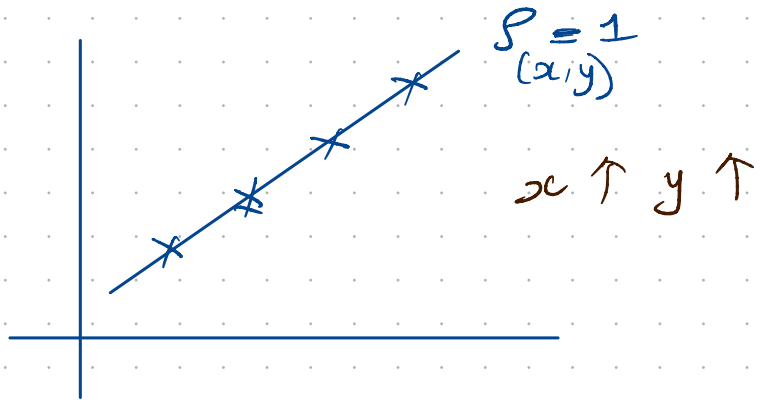
Pearson correlation Coefficient

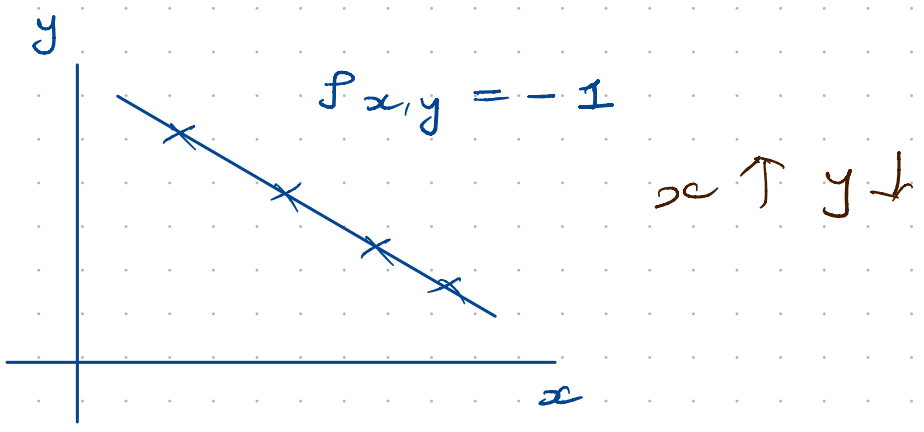
$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

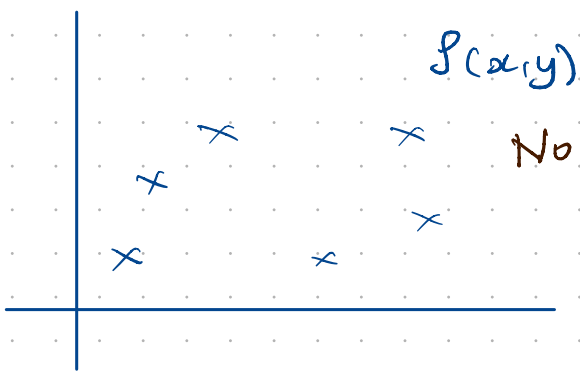
You know both in correlation.

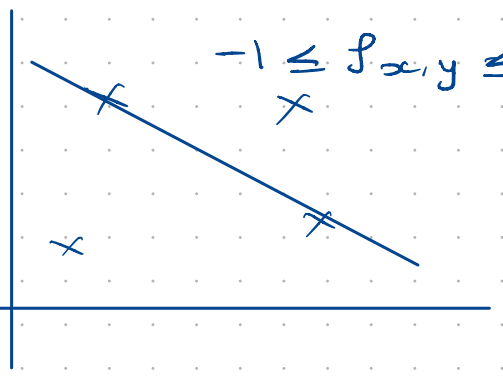
1. Strength (how much +ve or how much -ve).
2. Direction of Relationship (+ve or -ve)

$$-1 \leq \rho \leq 1$$









Spearman's rank correlation coefficient

$$r(rho) = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x}\, \sigma_{rg_y}}$$

$rg_x \rightarrow$ Rank of x.

$$r_s = \rho_{rg_x, rg_y}$$

Spearman's correlation coefficient (rho) is simply calculating Pearson's correlation coefficient but for rank of x & y instead of just x & y.

Consider an example, where X is an IQ value and Y is number of hours spent in front of T.V.

X	Y
103	29
97	20
113	12
112	6
110	17

First find then Rank

i) Sort the first column. Create new column x_i & assign ranked values.

ii) similarly do for Y.

iii) create column d_i for diff b/w x_i & y_i

iv) Final column d_i^2

X	Y	rank x_i	rank y_i	d_i	d_i^2
97	20	1	4	-3	9
101	29	2	5	-3	9
110	17	3	3	0	0
112	6	4	1	3	9
113	12	5	2	3	9

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 36}{20} = 1 - \frac{216}{5(24)}$$

$$= 1 - \frac{216}{120}$$

$$= \frac{-96}{120} = -0.8$$

$\boxed{\rho = -0.8}$ $x \uparrow y \downarrow$

Inter Quartile Range

4, 4, 6, 7, 10, 11, 12, 14, 15

Median = 10

IQR is finding the middle of the first & second half diff b/w

(25 & 75th percentile)

4, 4, 6, 7 $\Rightarrow$ Median = 5

11, 12, 14, 15 $\Rightarrow$ Median = 13

IQR = 13 − 5 = 8

Find outliers in dataset using IQR:

1) Arrange data in ascending order.

2) calculate First & third quantile

3) Find IQR range ($q_3 - q_1$)

4) Find lower bound = $q_1 - (1.5 * IQR)$

5) Find upper bound $q_3 + (1.5 * IQR)$

Anything that lies outside of lower & upper is an outlier.

Standardization Vs Normalization:

* Normalization helps you to scale down the feature between 0 to 1.

* Standardization helps you to scale down the feature based on standard normal distribution.

Min Max Normalization:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Min Max Scaler:

If you have two features whose magnitudes are huge away, so we should try to scale down between same scale.

Standardization (Z-score Normalization):-

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

In standardization, all the features will be transformed in such a way that it will have the propeerties of a standard normal dist with $\mu = 0$ & $\sigma = 1$.

What is PDF, PMF, CDF?

Probability distribution functions.

1) PDF ⇒ Probability density function ⇒ Continuous R.V (obtained by measuring)

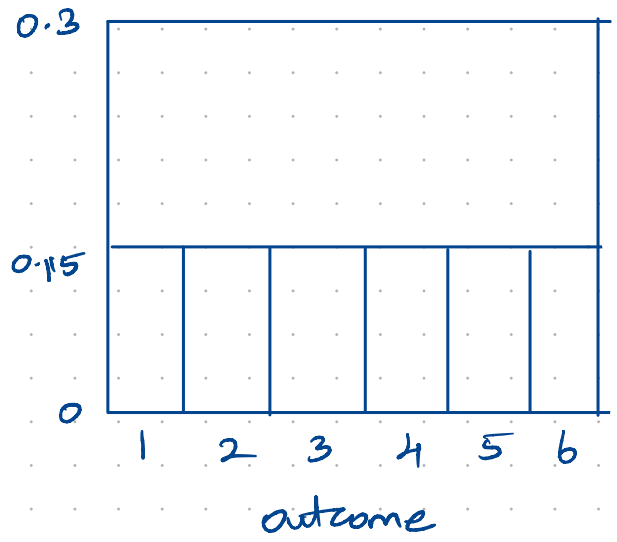
2) PMF ⇒ Probability Mass function ⇒ Discrete R.V (obtained by counting)

3) CDF ⇒ Cumulative Distribution function

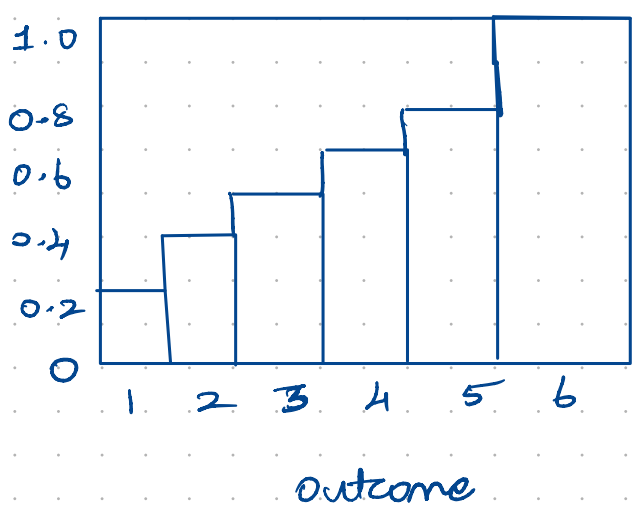
* PMF :-

Ex:- Dice Outcomes

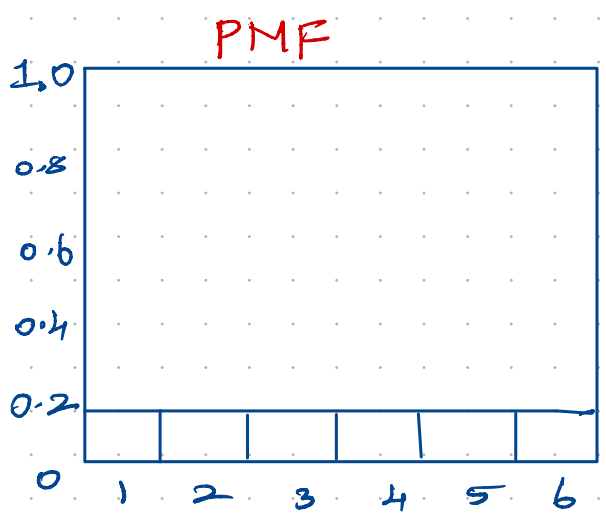
All values in a dice has the probablity of 1/6. = 0.167

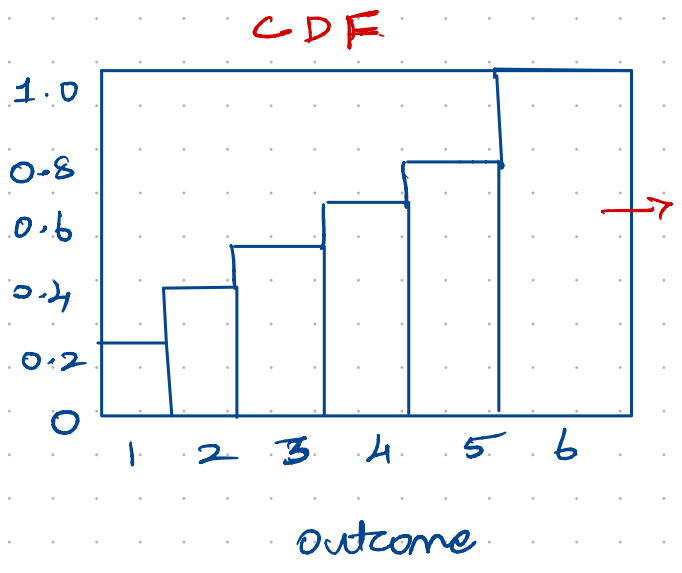


CDF for the same:



The diff b/w the above PMF & CDF diagram is its scale. PMF ranges from 0 to 0.3 & CDF from 0 to 1. Draw the PMF also in the scale of 0 to 1.





The final bar always needs to be 1 in CDF.

So in CDF, let's take $P(X \leq 3)$, then

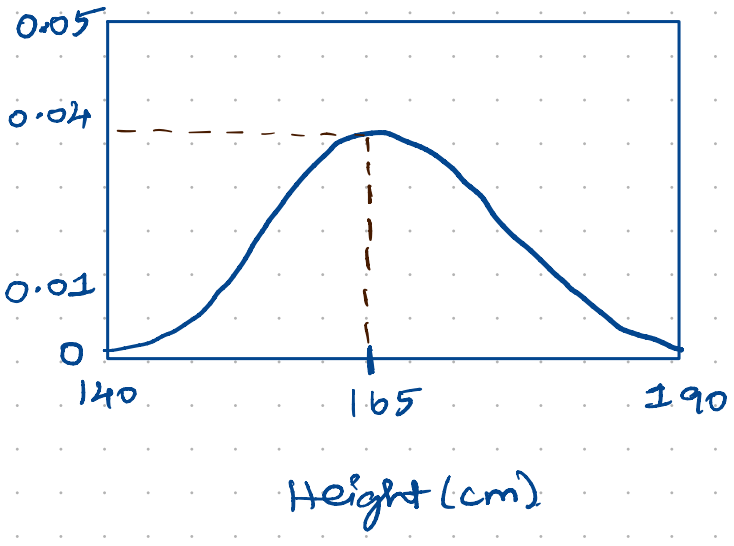
$$P(X \leq 3) = P(X=1) + P(X=2) + P(X=3)$$

$$= 0.167 + 0.167 + 0.167$$

$$\simeq 0.5$$

* PDF :

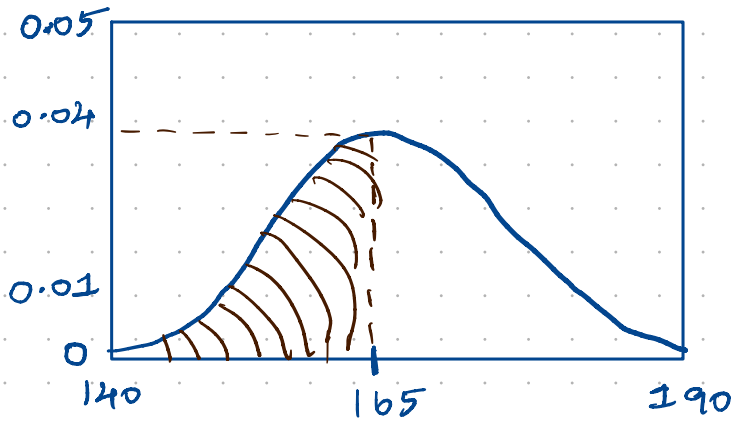
Ex:- Heights of Men



Does this plot means that $P(X = 165) = 0.04$?

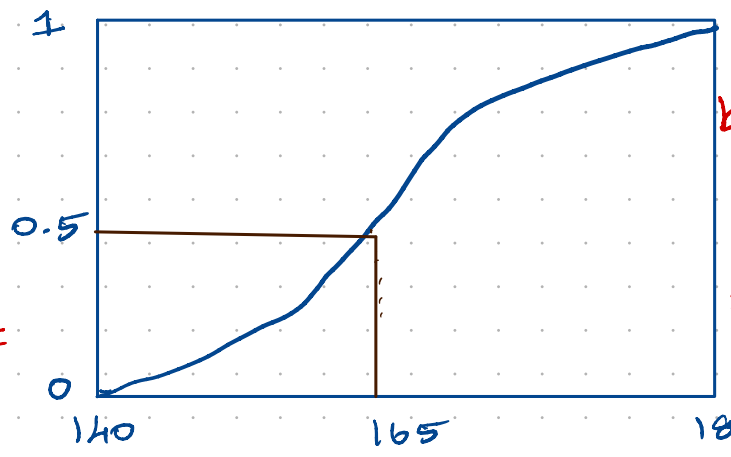
NO

PDF



Link b/w PDF & CDF

CDF



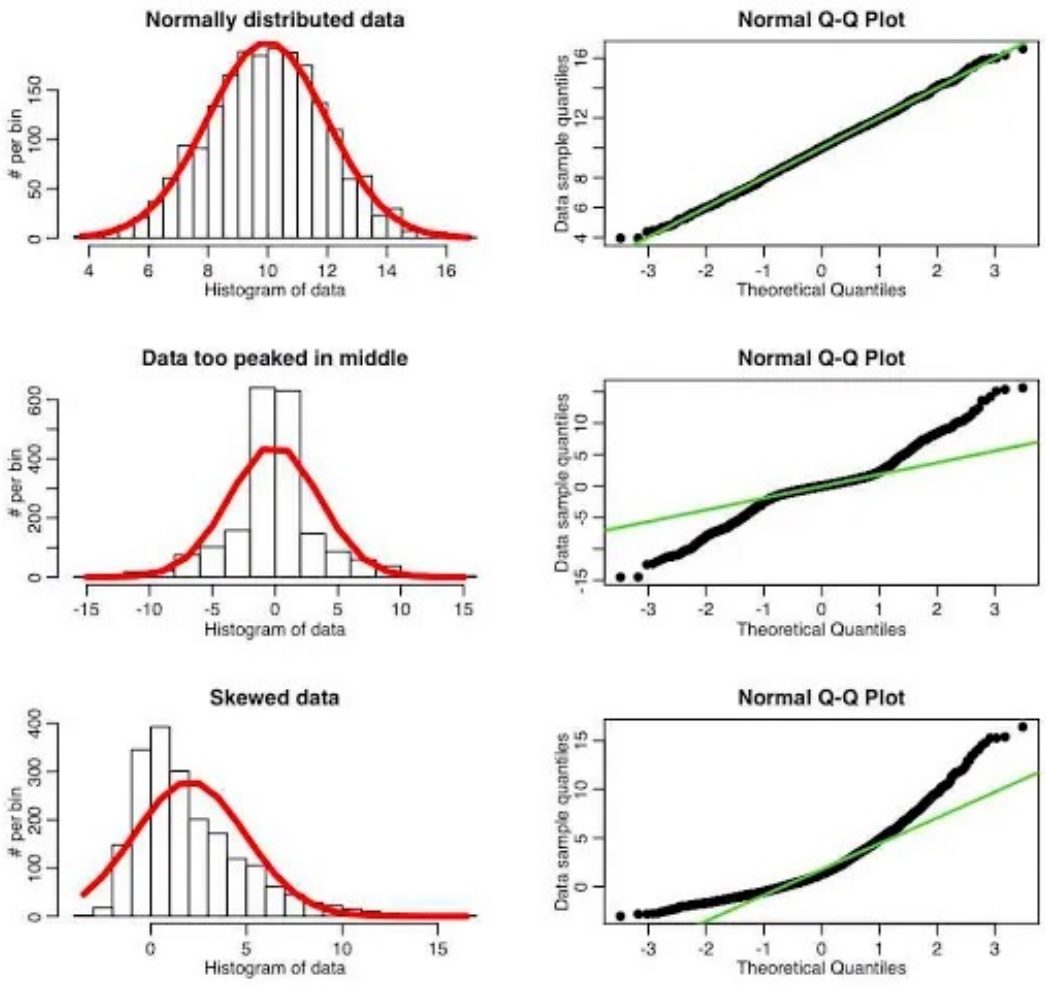
CDF is supposed to be in shape of "S". But in this diagram it is not. Bear with it

Q - Q Plots (Quantile - Quantile Plots):-

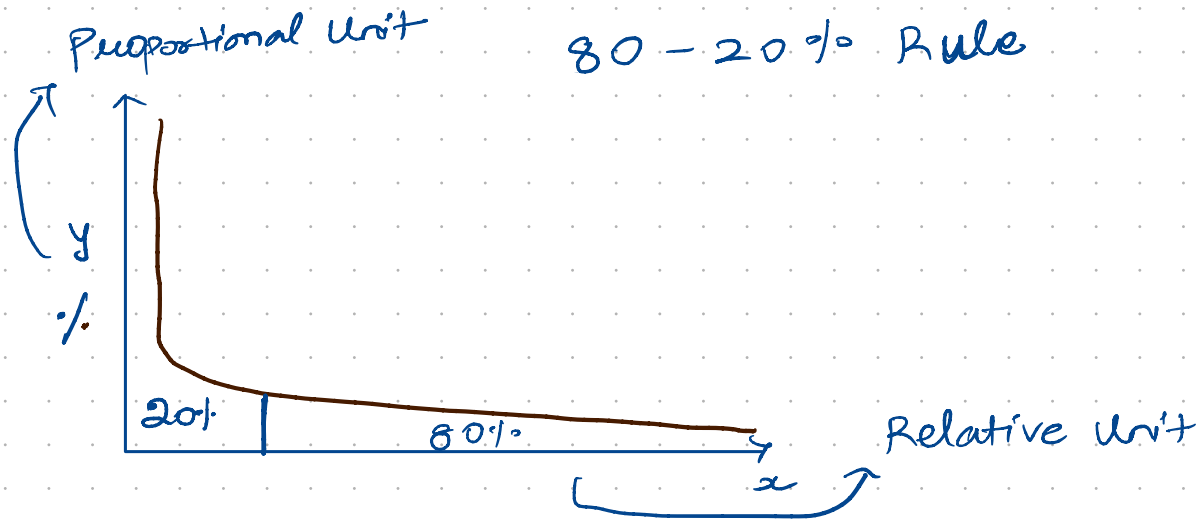
How to test if a R.V is normally distributed or not.

* Helps in graphically analyzing, and comparing two probablity distributions by plotting their quantiles against each other.

* As we are comparing two distributions, if they are equal, then the Q - Q plot will lie in the straight line ($y = x$)



Power law distribution:-



Ex:-

80 %- of sales are coming from 20 %. of overall products

It describes a particular pattern in which few extreme value occurs much more frequently than most other values. It follows a long tail distribution, where small number of events have high frequency & vast majority have a low frequency

Power-law:

Relationship between two quantities, wherein a relative change in one quantity results in a proportional change in other quantity

Confidence Intervals:

Point estimate:

In many cases, knowing population mean is difficult bcoz of the large input size, so we calculate sample mean ($\bar{x}$) with the help of $\bar{x}$, we will estimate the parameter (μ) of population mean.

* Confidence Interval is the mean of the estimate plus & minus the variation in that estimate.

Ex:-

Avg size of sharks in the Sea.

population S.D = 100

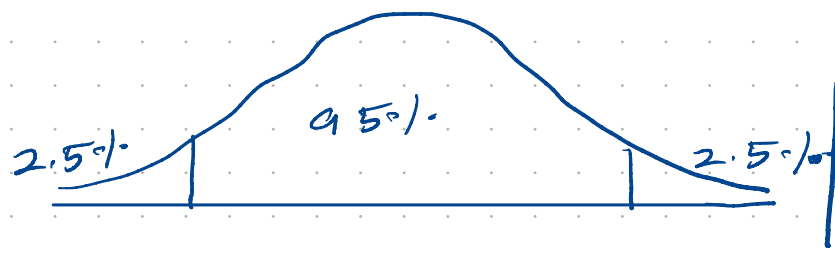
sample n = 30

$\bar{x}$ = 500

Confidence Interval = Point Estimate $\pm$ Margin Error.

$= \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Take 95 %. confidence Interval, then



α = significance value

As I have taken 95 %. CI, then 5 %. is the significance so 0.05

$$C.I = 500 \pm Z_{0.05/2} \frac{100}{\sqrt{n}}$$

$$= 500 \pm Z_{0.025} \frac{100}{\sqrt{n}}$$

$Z = 1 - 0.025$

$= 0.975$

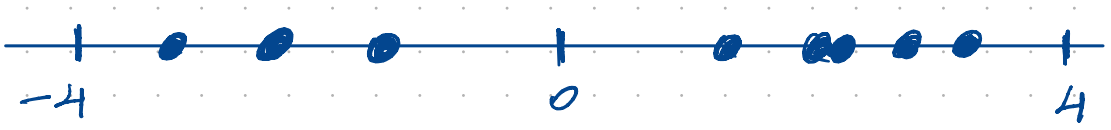
(Lookup in z-table for $Z_{0.975}$)

$$C.I = 500 \pm 1.96\left(\frac{100}{\sqrt{n}}\right)$$

$$500 - 1.96 * \frac{100}{\sqrt{n}} < C.I < 500 + 1.96 * \frac{100}{\sqrt{n}}$$

Confidence Interval Using Bootstrapping

Bootstrapping:



crete a new numberline and add 8 values randomly sampled with replacement.

One such example will be:

→ Bootstrapped dataset

As this is diff from original dataset, we will get diff mean.

Bootstrapping has 4 steps:

1) Make a Bootstrapped Dataset
2) Calculate mean (any statistics)
3) Keep track of mean (any statistics)
4) Repeat 1 - 3 for bunch of time

Bootstrap for 1000 times & plot it in the histogram.

We can find 95% confidence Interval using the Bootstrapping.

Hypothesis Testing & Null Hypothesis

The hypothesis that there is no difference b/w things is called the Null Hypothesis.

Hypothesis testing is used to assess the credibility of a hypothesis by using Sample Data.

P-value:

P-values are numbers b/w 0 and 1. It tells you how likely is that the data could have occured under null hypothesis. (Measure of evidence against null hypothesis)

The commonly used threshold in p-value is 0.05. It means we did the exact same experiment a bunch of times, then only 5% of the experiments results in wrong decision.

Ex:-

Test whether a new drug is effective in reducing blood pressure.

Null hypothesis (H_0) :- Drug has no effect

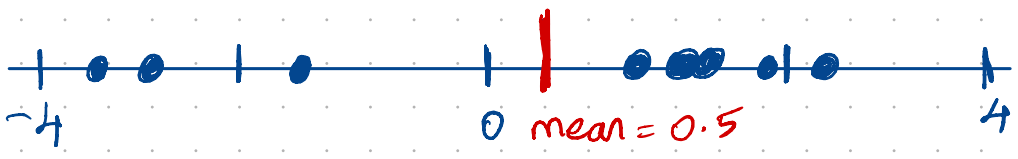
Alternate hypothesis :- Drug is effective.

We calculated a p-value of 0.02, but our significance level (threshold) is 0.05, then we would reject the null hypothesis

Smaller the p-value, stronger the evidence that you should reject null hypothesis.

Using Bootstrapping to calculate the p-value :-

Your original data is



Create a new line & shift the datapoints in such a way that mean will be 0.

* Then create a bootstrapped datasets from a collection of measurements with mean = 0, so the resulting histogram gives us a sense of what would happen if the Null hypothesis was true.

https://youtu.be/N4ZQQqylf6k ⇒ Refer for the example with diagrams.

As the mean on the original data is 0.5, on the bootstrapped data check the value for

pvalue = p(bootstrapped mean >= 0.5) + p(bootstrapped mean <= -0.5)

By the ex: p-value = 0.63

As this is greater than 0.05, we fail to reject the null hypothesis.