22.1 CDF of a Gaussian/Normal Distribution

The CDF value for any distribution always ranges in between 0 and 1. The left bottom tail of the CDF curve has a value of 0 and the right top tail has a value of 1.
Let us assume we have a random variable 'X' and the CDF at a point 'x' indicates the probability of 'X' taking the values that are less than or equal to 'x'.

CDF(X=x) = P(X<=x)

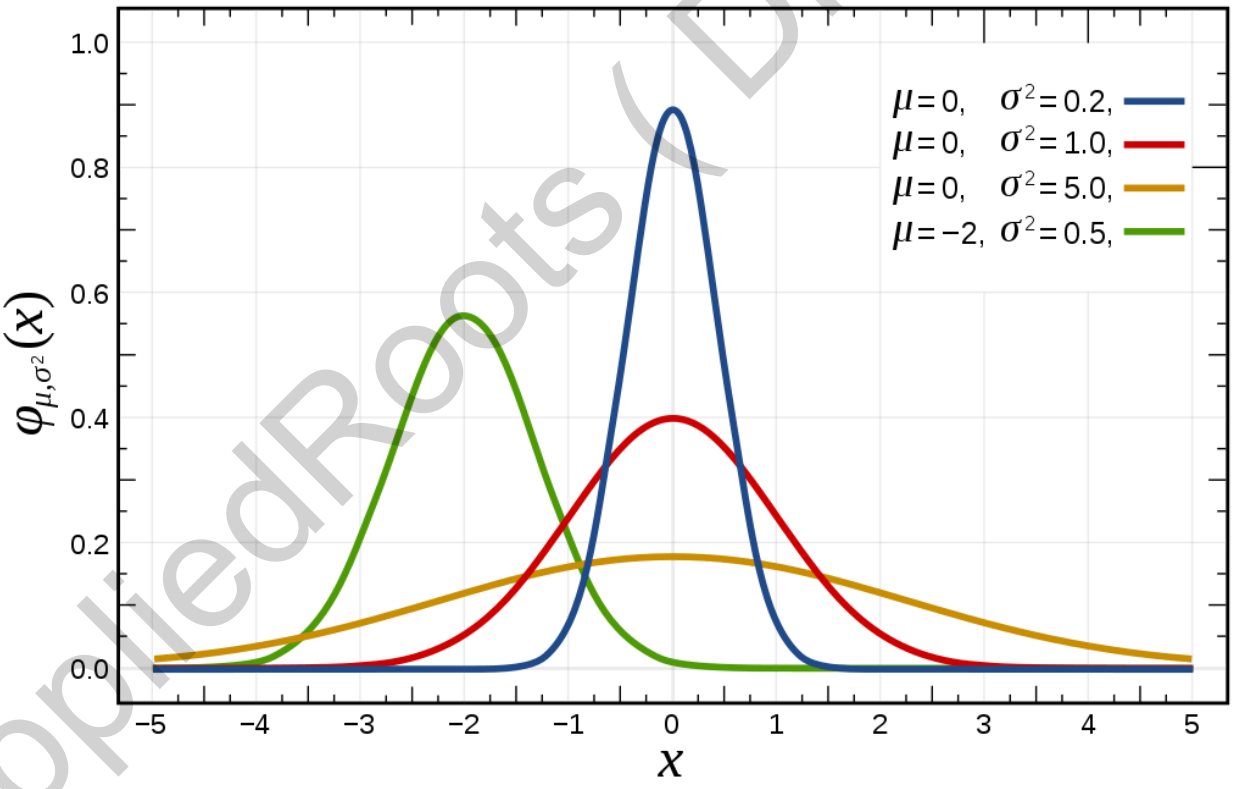
A normal distribution is represented as **$X \sim N(\mu, \sigma^2)$** where

μ → Mean of 'X'
σ^2 → Variance of 'X'

If the given distribution is symmetric, then 50% of the points lie to the left side of the mean and the remaining 50% of them lie to the right side of the mean. So for such distributions, the CDF at the mean is always equal to 0.5.

Note: This logic is applicable only for the symmetric distributions.

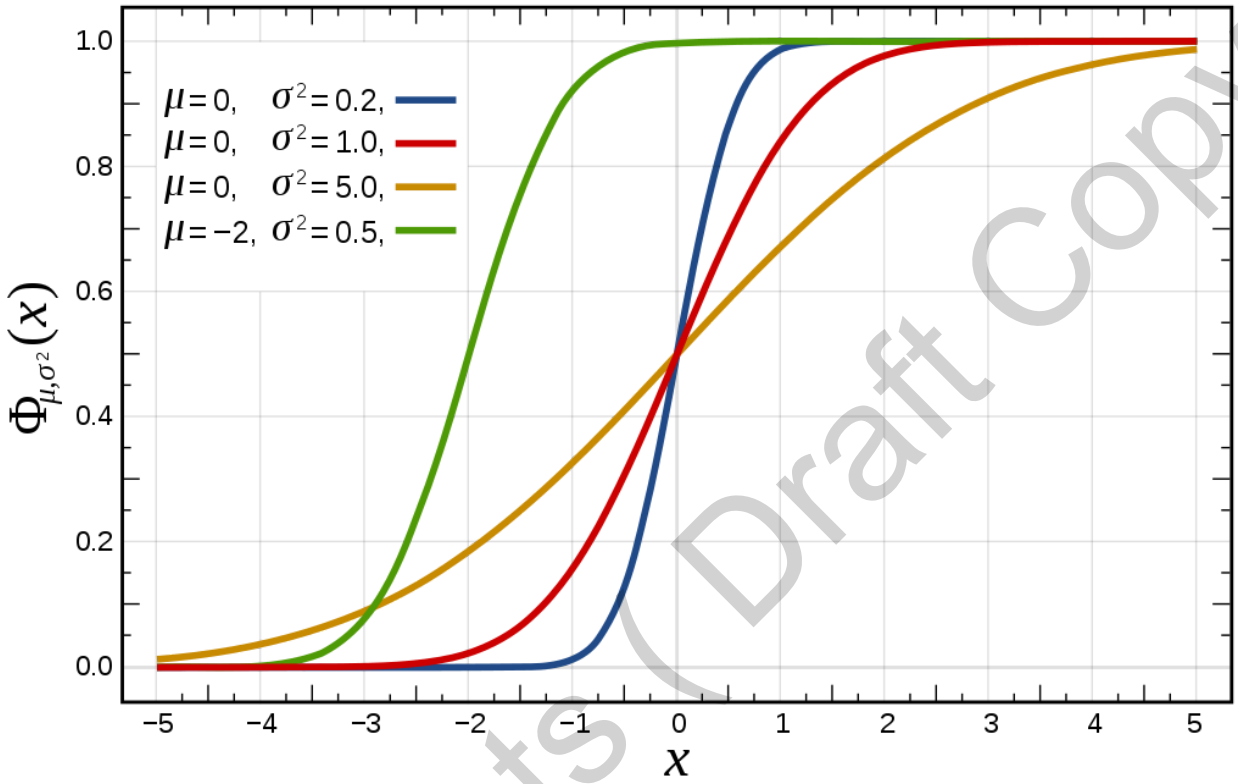
Let us look at the plots of the PDFs and the CDFs of Normal distribution with different mean and variance values, discussed at the timestamp 0.30



Here we can see the PDFs of the gaussian distribution with different mean and variance values. The peak of the curve always lies at the mean and the shape of the curve is symmetric around the mean. Here 50% of the values lie to the left side of the mean and the remaining 50% lie on the right side of the mean.

The curve becomes wider and the peak falls down, if the variance is large. The curve becomes narrower and the peak goes high, if the variance is small.

If the mean increases, then the curve moves towards the right and if the mean decreases, then the curve moves towards the left.



In the above plot,we can see the plot of CDFs of the normal distribution with different mean and the variances. The CDF value for a symmetric distribution at the mean is always equal to 0.5. The alignment of the CDF plot on the axis depends on the position of the mean. If the mean increases, then the CDF curve moves towards the right and if the mean decreases, then the CDF curve moves towards the left. But the value of the CDF always lies in between 0 and 1, irrespective of the position of the mean.

The width of the CDF plot is dependent on the variance. If the variance increases, then the curve moves away from the vertical axis of the mean. If the variance decreases, then the curve moves towards the vertical axis of the mean

Standard Deviation is a measure that is used to quantify the amount of variation or dispersion of the given set of values. A low standard deviation indicates that the data points tend to be close to the mean. If the standard deviation is high, then it indicates that the data points are spread out over a wide range of values.

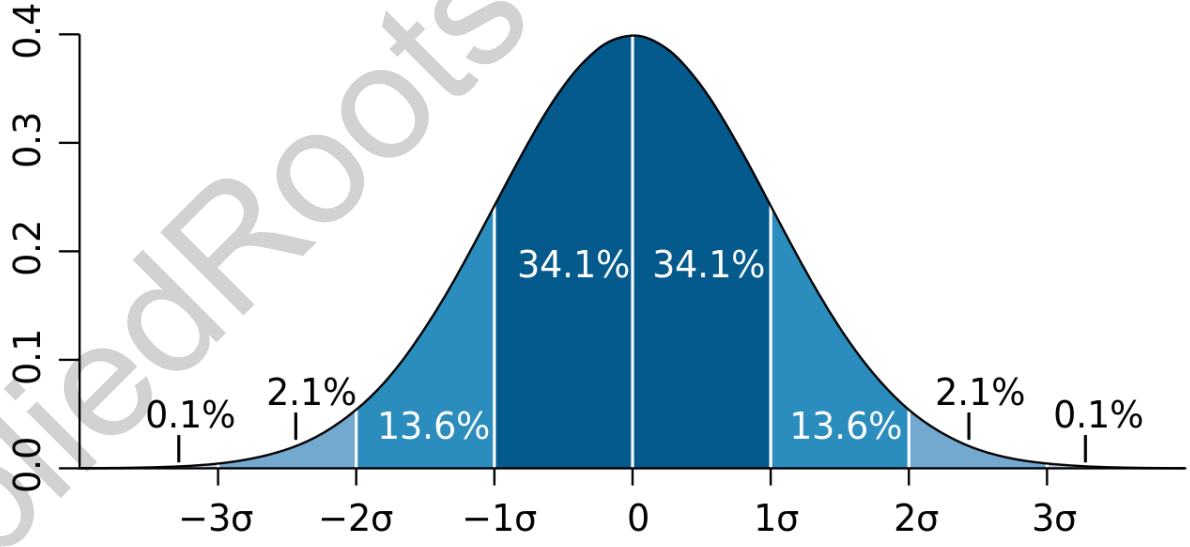
Properties of CDF

- ➢ The CDF function should be non decreasing.
- ➢ The CDF function has to be right continuous.
- ➢ As the value of 'x' keeps tending to negative infinity, the value of CDF should tend to 0.
- ➢ As the value of 'x' keeps tending to positive infinity, the value of CDF should tend to 1.
- ➢ CDF is never symmetrical in nature.
- ➢ CDF is obtained by cumulatively adding the probabilities and as the probabilities can never be negative, the CDF curve never goes down.
- ➢ The CDF curve starts at 0 and ends at 1 and usually has 'S' shape.

68-95-99.7 Rule

- ➢ The 68-95-99.7 rule or the empirical rule is used to remember the percentage of the values that lie in an interval estimate for a normal distribution.
- ➢ 68%, 95% and 99.7% of the values lie in the intervals of first, second and the third standard deviations respectively on both sides.

Below is the representation of the 68-95-99.7 rule, discussed at the timestamp 4:20.



For example, we have a normal distribution with a mean of 150 and a standard deviation of 25, then

1σ = 25; 2σ = 50, 3σ = 75

68% of the values lie in the interval [150-25, 150+25] (ie., [125,175])

95% of the values lie in the interval [150-50, 150+50] (ie., [150,200])

99.7% of the values lie in the interval [150-75, 150+75] (ie., [75,125])

Note: It is always not mandatory for the mean of a normal distribution to be 0. It can also be a non zero value. But once if the values are standardized, then the mean becomes 0 and the standard deviation becomes 1. The concept of Standardization has been discussed in the next set of lectures in this chapter itself.

This 68-95-99.7 rule is applicable only for the normal distribution.

22.2 Symmetric Distribution, Skewness and Kurtosis

Symmetric Distribution

A probability distribution is said to be symmetric, only if there exists a value $\mathbf{x_0}$, such that $\mathbf{f(x_0+\delta) = f(x_0-\delta)}$ for all real numbers of'δ' and '**f**' is the PDF. It means if we choose any value 'δ', then the PDF value at $\mathbf{X=x_0+\delta}$ and $\mathbf{X = x_0-\delta}$ should be the same.

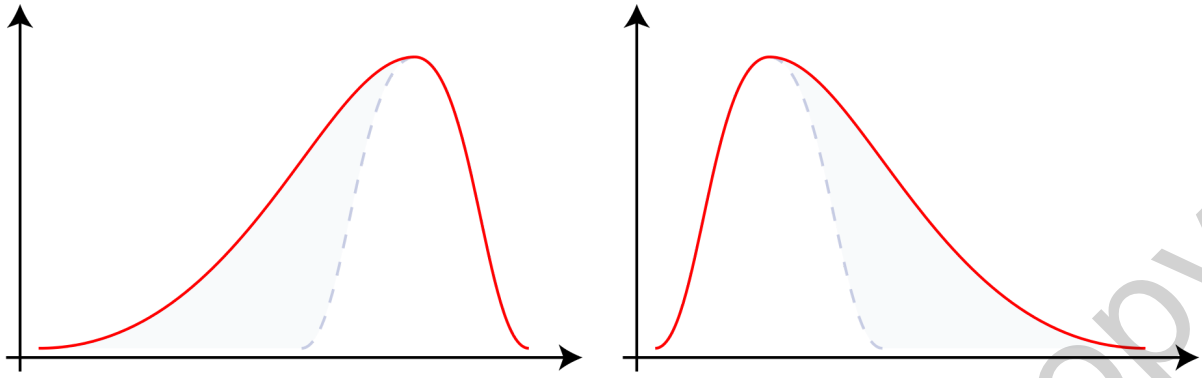
Here if 'f' is symmetric, then the portion of the distribution/curve present to the left side of the point '$\mathbf{x_0}$' is the mirror image of the portion of the distribution/curve present to the right side of the point '$\mathbf{x_0}$'.

In case, if there doesn't exist any point '$\mathbf{x_0}$', such that $\mathbf{f(x_0+\delta) = f(x_0-\delta)}$, then such a distribution is said to be non-symmetric.

Skewness

- Skewness is a measure of asymmetry of the probability distribution of a real valued random variable, around it's mean. It can have either a positive value or a negative value or zero.
- For a symmetric distribution, the skewness value is 0, whereas for a non symmetric distribution, the skewness value could either be negative or positive. For a symmetric distribution, the density of the points on the left side region (ie., before the point '$\mathbf{x_0}$') is same as the density of the points on the right side region (ie., after the point '$\mathbf{x_0}$').
- If a distribution has a longer tail to the left side, then it is called Negatively Skewed Distribution (or) Left Skewed Distribution. The value of skewness for such a distribution is always negative. Here the density of the points on the PDF is more towards the right when compared to the left.
- If a distribution has a longer tail to the right side, then it is called Positively Skewed Distribution (or) Right Skewed Distribution. The value of skewness for such a distribution is always positive. Here the density of the points on the PDF is more towards the left when compared to the right.

Below is the figure that represents the positive and the negative skewed distributions and it was discussed at the timestamp 7:00



Negative Skew Positive Skew

In this figure, we could see that in Negative Skewed Distribution, the tail is present to the left side which means the density of the points is lower on the left side and higher on the right side. Similarly, in Positive Skewed Distribution, the tail is present to the right side which means the density of the points is lower on the right side and higher on the left side.

The more negative the skewness measure value is, the more left/negative skewed the distribution would be and the tail would be longer on the left side. The more positive the skewness measure value is, the more right/positive skewed the distribution would be and the tail would be longer on the right side.

The formula for computing the sample skewness was discussed at the timestamp 10:15 and is given below

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}$$

Here '$\bar{x}$' represents the sample mean and 'n' denotes the sample size.
Note: The skewness of a normal distribution is 0.

Advantages of Skewness

- Skewness is used along with the histogram and the QQ plot to characterize the distribution of the data.
- The magnitude(ie., the absolute value) of the skewness indicates how far is the distribution from symmetry. The sign(ie., positive or negative) of skewness indicates the direction in which the distribution is skewed.

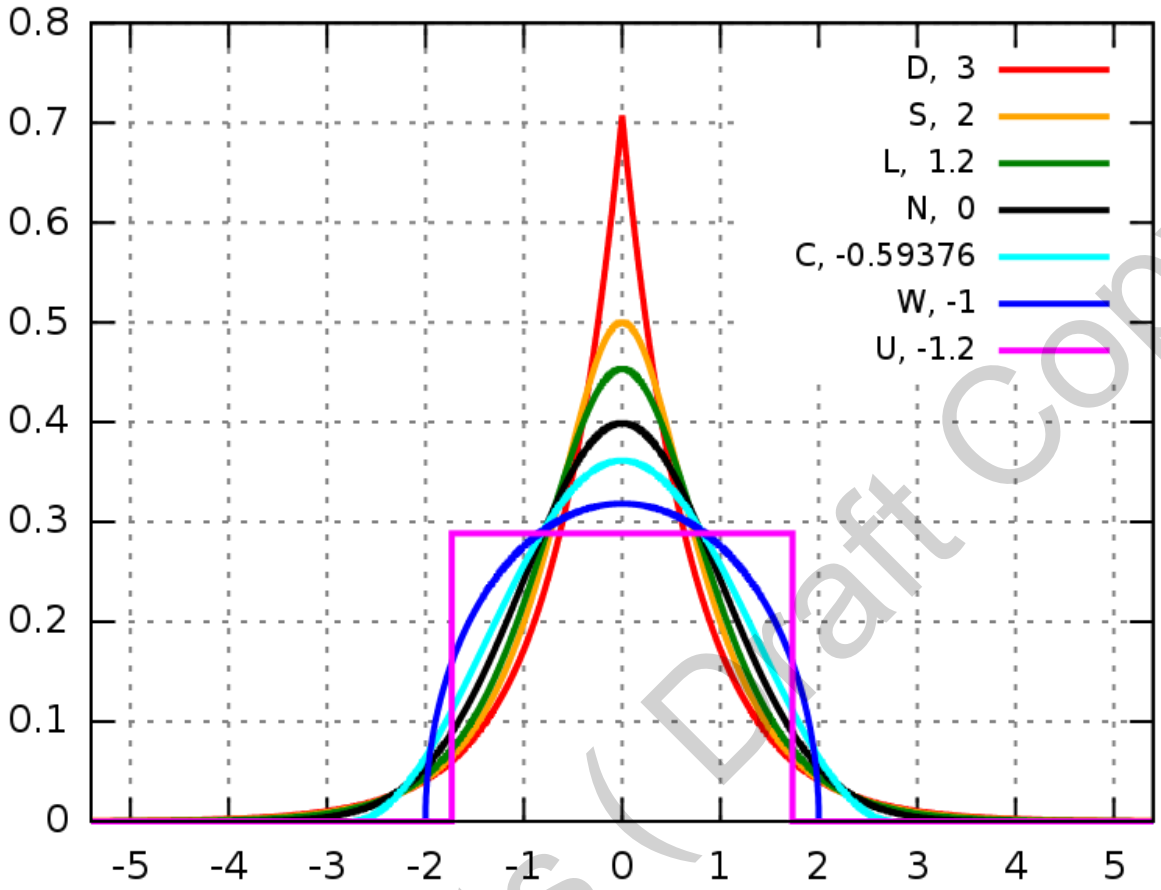
Kurtosis

- ➢ Kurtosis is a measure of tailedness of the probability distribution of a real valued random variable.
- ➢ Skewness only indicates whether the given distribution is left skewed or right skewed whereas Kurtosis gives additional information about the shape of the distribution.
- ➢ Distributions with large Kurtosis have the tails exceeding the tails of normal distribution.
- ➢ Distributions with small Kurtosis have the tails smaller when compared to the tails of normal distribution.
- ➢ There are different ways of quantifying the kurtosis for a theoretical distribution and corresponding ways of estimating it from a sample or population. Different measures of kurtosis may have different interpretations.

Below is the formula for computing the excess Kurtosis metric and it was discussed at the timestamp 15:20

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^2} - 3$$

In the above formula, the first term denotes the kurtosis value and the excess kurtosis is given as (**excess kurtosis = kurtosis-3**). Below is the graph of how the PDFs look like with different values of the ‘excess kurtosis’ metric and it has been discussed at the timestamp 17:20.



In this graph of PDFs, we can see the large values of 'excess kurtosis' lead to longer tails of PDF on both the sides whereas smaller values of 'excess kurtosis' leads to quickly falling tails in the PDF. The 'excess kurtosis' value for gaussian distribution is always 0.

The 'excess kurtosis' indicates how different is the shape of the distribution when compared to a gaussian distribution with a kurtosis value of 3.

Applications of Kurtosis

One of the major applications of Kurtosis is in finding the outliers from the given set of values.

Note: Symmetry, Skewness and Kurtosis are three parameters used in understanding the shape of the PDF of a distribution.

22.3 Standard Normal Variate (Z) and Standardization

If we have a random variable 'Z' with a mean of 0 and standard deviation of 1 and following normal distribution, the 'Z' is called a Standard Normal Variate.
It is represented as **Z~N(0,1)**

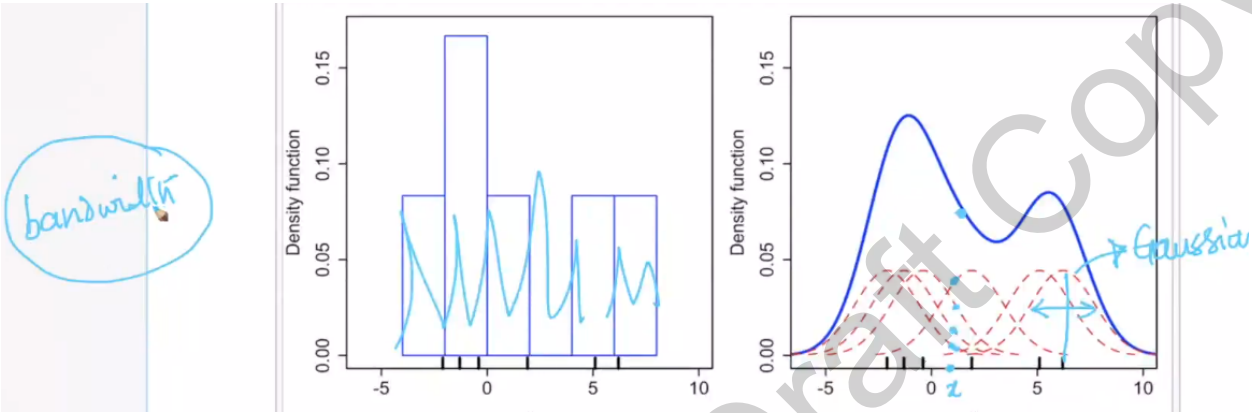
Let us assume we have a random variable 'X' which follows normal distribution and has a mean of '**μ**' and variance of 'σ^2', then it is represented as **X~N(μ, σ^2)**.

Standardization is the process of transforming a given distribution with a mean '**μ**' and variance of 'σ^2', into the same type of distribution with a mean of 0 and standard deviation of 1. (Even the variance also would be 1)

Let the values of 'X' be [x1,x2,x3,....,xn], then after applying standardization $x_i^l = (x_i - \mu)/\sigma$ and the transformed distribution is denoted as $X^l \sim N(0,1)$
Note: Standardization just transforms the given distribution of values onto a new scale with mean = 0 and standard deviation = 1. The nature of the distribution doesn't change at all.(irrespective of whether the distribution is gaussian or non gaussian)

22.4 Kernel Density Estimation (KDE)

So far we have seen the PDF and histograms in the univariate analysis. The PDF is obtained by drawing a smooth curve on the histogram using a technique called Kernel Density Estimation(KDE). Let us look at the plots that were discussed at the timestamp 0:35.

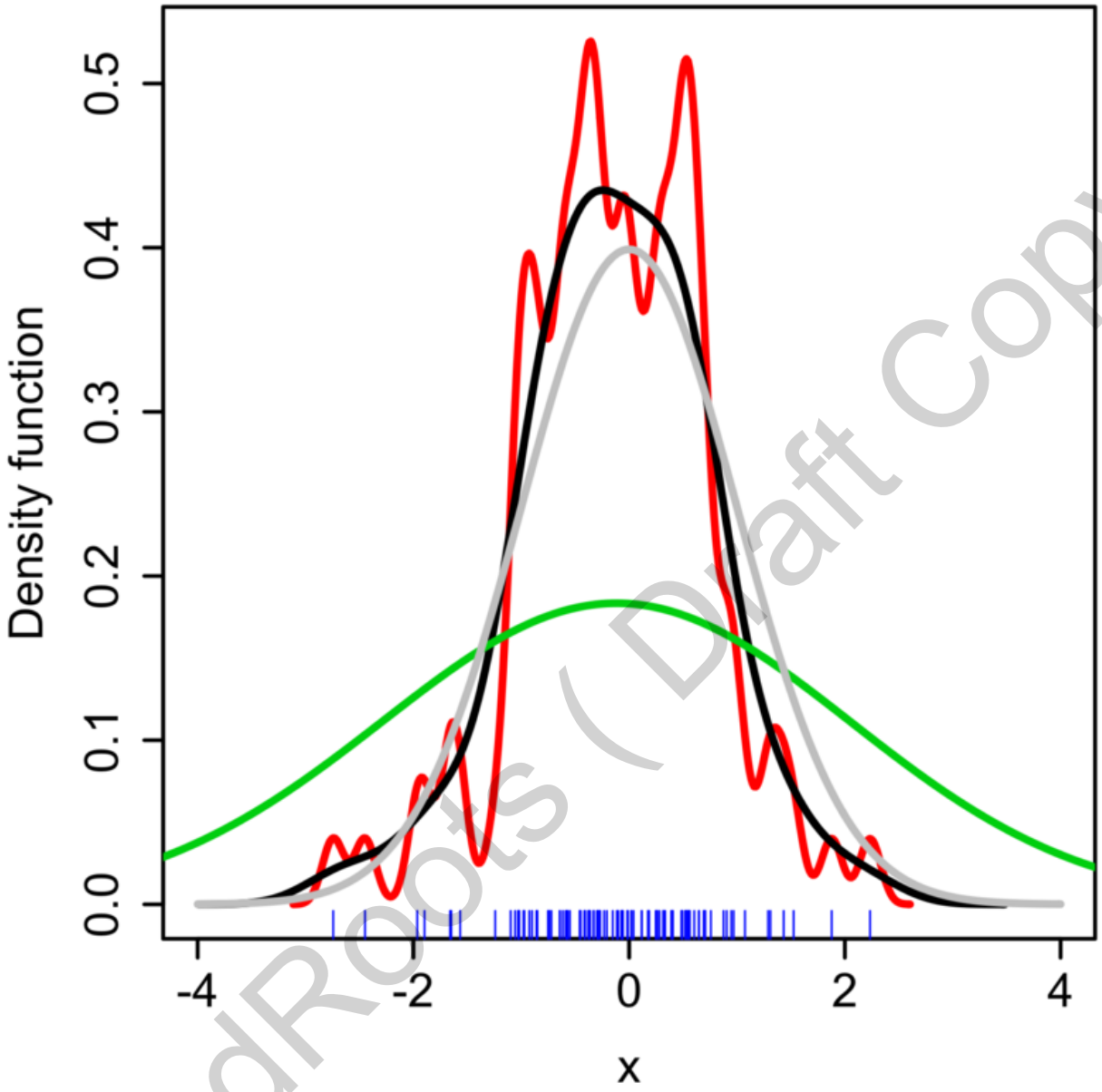


In the right side plot, the number of gaussian kernels drawn is equal to the number of points in the dataset. The kernels used here are the gaussian kernels. Here these gaussian kernels are plotted such that the mean of each of these kernels are the points in the dataset(ie., with each data point as the centre, we plot the gaussian kernels)

For every point on the 'X' axis, we look for how many kernels are passing through that point. The 'Y' axis coordinates of all these kernels are added to get the PDF value at that point.

The bandwidth of these kernels is nothing but the variance of these kernels. If the bandwidth is high, then the kernels will become much wider and thereby the resulting PDFs will be wider. Similarly if the bandwidth is low, then the kernels will become much narrower and the resulting PDF will look jaggered. Hence the bandwidth of these kernels should be chosen properly. The standard deviation of these kernels is equal to the bin width. The height of a gaussian kernel should be such that the area under it is equal to 1.

The smoothness of the PDF depends on the bandwidth chosen. So the value of bandwidth has to be chosen properly. Changing the bandwidth changes the shape of the kernel. If the bandwidth is low, only the points that are very nearer are taken into consideration for computing the PDF, and the curve looks more squiggly. If the bandwidth is high, then more points are taken into consideration for computing the PDF at a point, and the PDF curve looks shallow. You can find it out in the below PDF plot.



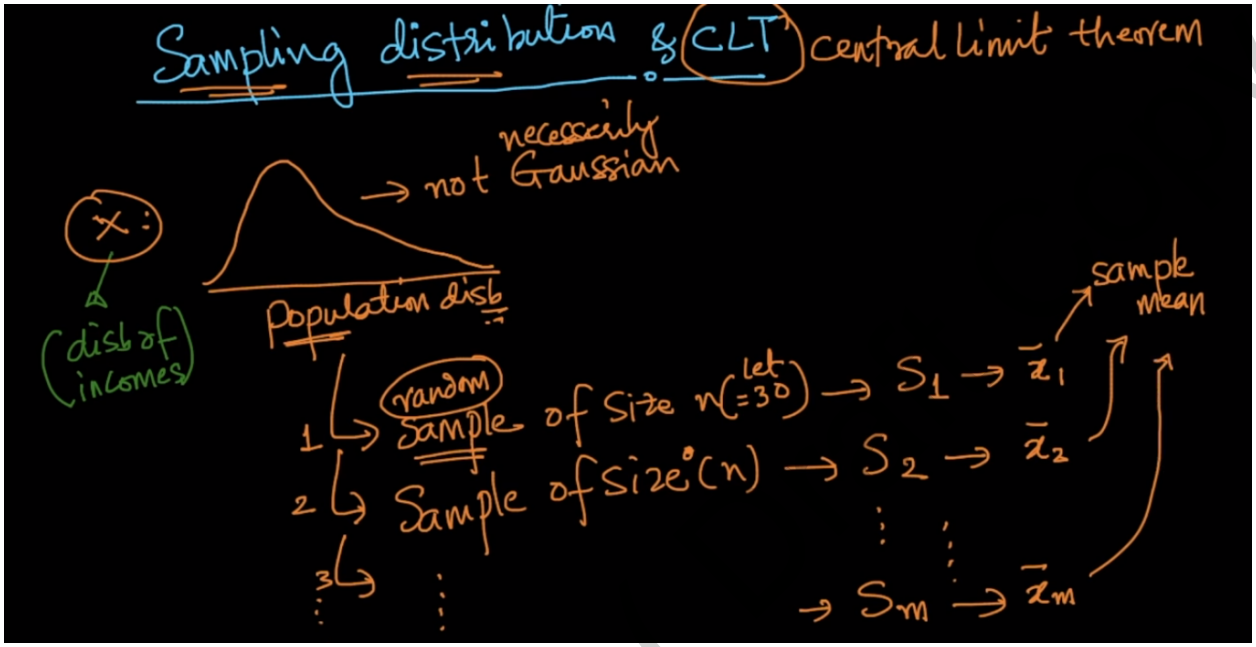
In the above plot, if the bandwidth is high, then the PDF looks like the green curve. If the bandwidth is moderate, then the PDF looks like the black curve. If the bandwidth is low, then the PDF looks like the red curve.

If the density of the points is high, then the height of the PDF increases in that region. If the density of the points is low, then the height of the PDF decreases.

The curves with more variance have more bandwidth and the curves with less variance have lesser bandwidth.

22.5 Sampling Distribution and Central Limit Theorem (CLT)

Sampling Distribution



Let us assume a distribution for the given population. It is not mandatory for this distribution to be gaussian. Let us pick 'm' random samples(which are subsets of the population) from the population and let the size of each of these samples be 'n'. All these samples are independent of each other. Let us calculate the mean of each sample and denote them as

$\overline{x}_1$ → 1st sample mean

$\overline{x}_2$ → 2nd sample mean

.

.

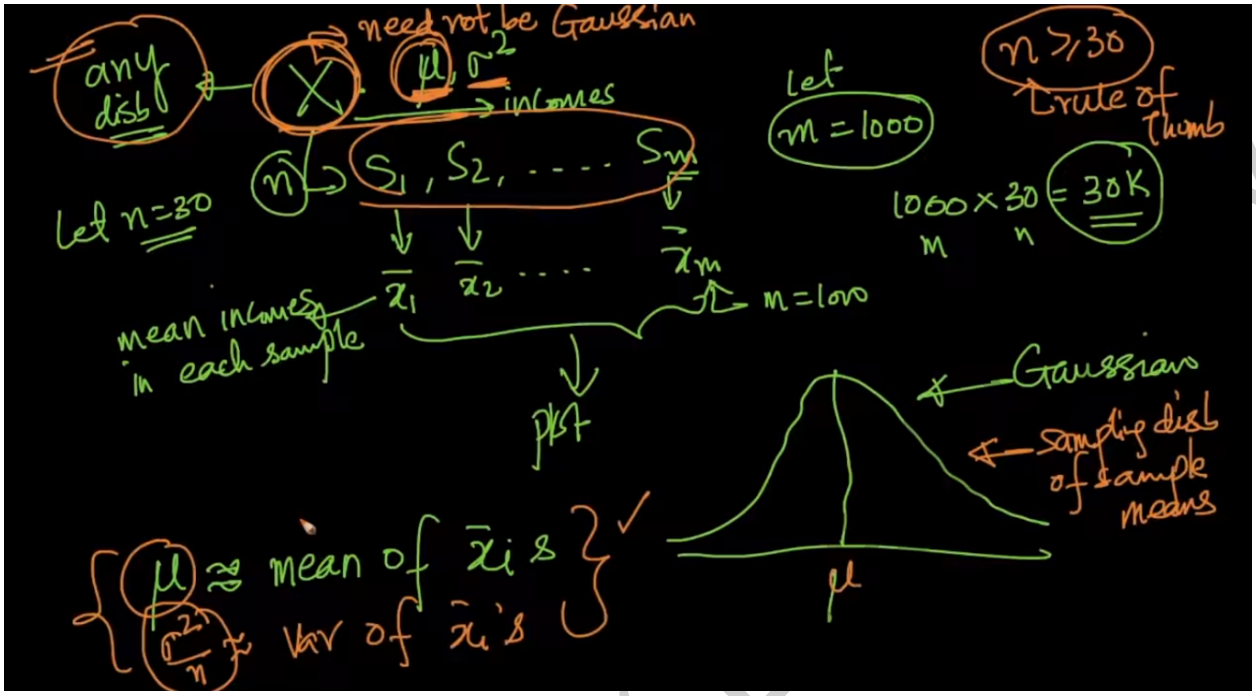
.

$\overline{x}_m$ → m^{th} sample mean

All these samples should be the same size. All the statistical measures of these samples have a probability distribution and it is called **Sampling Distribution**. In this context, as this sampling distribution is obtained from the sample means, we call it **Sampling Distribution of Sample Means**.

We also can have the sampling distributions of sample medians or sample variances as per our problem requirement. Sampling Distributions are very important in statistics as they provide a simplification route to statistical inference.

Central Limit Theorem



Let us assume a random variable 'X' and let it's population distribution have a finite mean and a finite variance. Let us pick 'm' different samples of size 'n' each (ie., S_1, S_2, S_3,...., S_m) from the population. Let us denote the mean of i^{th} sample as $\overline{x}_i$. So $\overline{x}_1$, $\overline{x}_2$, $\overline{x}_3$,....,$\overline{x}_m$ be the respective means of the samples. The distribution of $\overline{x}_i$ is the sampling distribution of sample means. Now the Central Limit Theorem states that

For any distribution of the random variable 'X' with a finite mean 'μ' and a finite variance 'σ^2', the sampling distribution of the sample means is a gaussian distribution with a mean nearly equal to the population mean 'μ' and variance equal to 'σ^2/n'.

CLT works always only for the populations with a finite mean and a finite variance. The population distribution could either be gaussian or not, but sample size should be large, all the samples should be of the same size, and all the samples are picked with replacement. Pareto Distribution is an example of a distribution with an infinite mean and an infinite variance. So CLT doesn't work if the population distribution is pareto.

The 68-95-99.7 rule applies for the sampling distribution of the sample means. If the population distribution of 'X' is gaussian, then the 68-95-99.7 rule applies for both the sampling distribution of the sample means and also the population distribution of 'X', whereas if population distribution of 'X' is not gaussian, then the 68-95-99.7 rule applies only for the sampling distribution of the sample means, but not for the population distribution.

22.6 Q-Q plot: How to test if a random variable is normally distributed or not?

Quantile-Quantile (QQ) plot is a probability plot which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Procedure for QQ plots

Let 'X' = [x_1, x_2, x_3, ….., x_n] be the given random variable for which the probability distribution is unknown.

Let us create a random variable Y~N(0,1). So 'Y' = [y_1, y_2, y_3, ….., y_n].

Step 1

Sort all the x_i's in ascending order and compute the percentiles of 'X'.

Notation

x_i → i^{th} point/value of 'X'

$x^{(i)}$ → i^{th} percentile of 'X'

Step 2

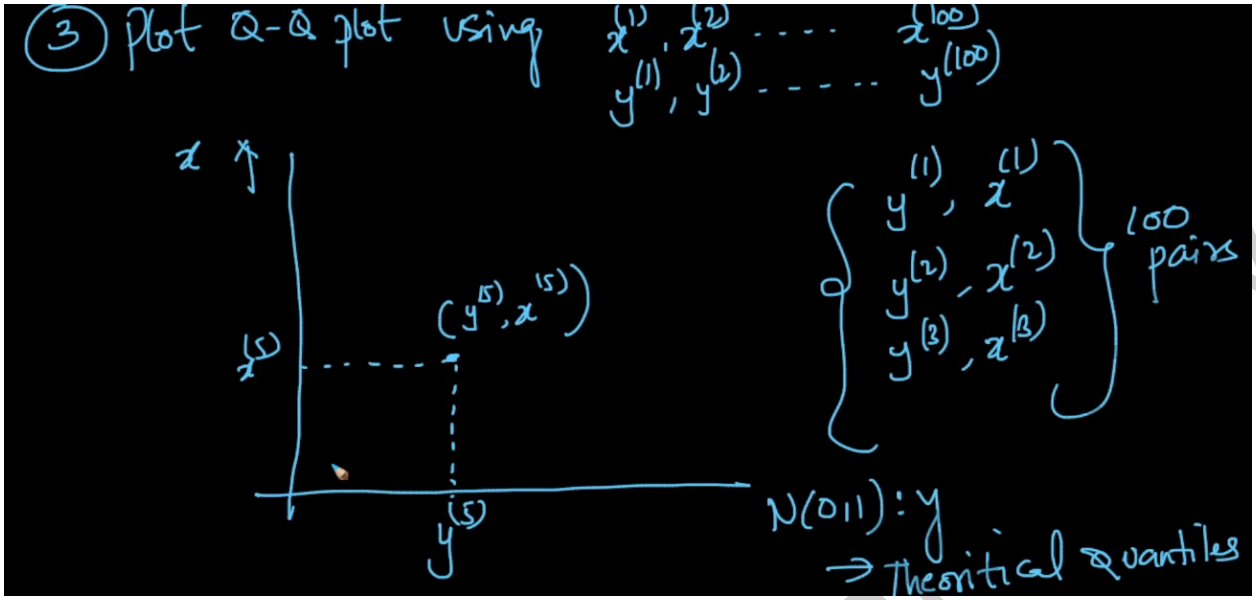
Sort all the yi's in ascending order and compute the percentiles of 'Y'.

Step 3

Pick the 100 percentiles of both 'X' and 'Y' and form them into pairs as ($y^{(i)}$, $x^{(i)}$) and plot these points.

After plotting these points, if all these points are on a straight line, then we can say both 'X' and 'Y' follow the same distribution.

Below is the QQ plot that was discussed starting from the timestamp 6:05



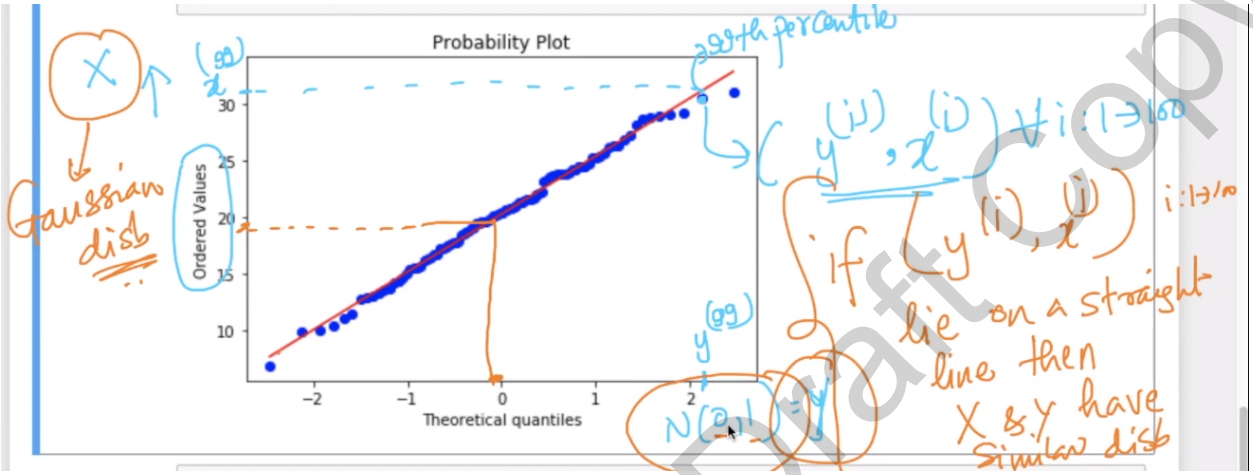
Below is the code snippet for building the QQ plot that was discussed starting from the timestamp 11:40.

Q-Q Plot

```
#Q-Q plot
import numpy as np
import pylab
import scipy.stats as stats

# N(0,1)
std_normal = np.random.normal(loc = 0, scale = 1, size=1000)

# 0 to 100th percentiles of std-normal
for i in range(0,101):
    print(i, np.percentile(std_normal,i))
```

```
# N(0,1)
std_normal = np.random.normal(loc = 0, scale = 1, size=1000)

# 0 to 100th percentiles of std-normal
for i in range(0,101):
    print(i, np.percentile(std_normal,i))
```

```
# generate 100 sanples from N(20,5)
measurements = np.random.normal(loc = 20, scale = 5, size=100)
#try size=1000


stats.probplot(measurements, dist="norm", plot=pylab)
pylab.show()
```



In a QQ plot, if all the points ($y^{(i)}$, $x^{(i)}$) lie along a straight line, then the distribution of 'X' is the same as the distribution of 'Y'.
ie., If 'Y' is gaussian, then 'X' also follows Gaussian distribution. If 'Y' is pareto, the 'X' also follows Pareto distribution.

If we increase the number of points in both the sets of the percentiles of 'X' and 'Y', then the points on the QQ plot move closer to each other and the plot becomes more straight. If the number of points in 'X' and 'Y' is less, then the plot looks like a curve, instead of a straight line, even if both 'X' and 'Y' belong to the same distribution.

If 'X' and 'Y' are of different distributions, then the QQ plot is nonlinear. In such cases, even if you increase the number of points in both 'X' and 'Y', the plot becomes more non linear, as they both are from two different distributions.

When we have two random variables 'X' and 'Y' and if the type of distribution of at least one of them is known, then we can check whether they both belong to the same distribution or not, by building the QQ plot.

In case, if we are not known of the distribution of any one of those two random variables, then we should create a random normal or a random uniform distribution 'Z' and build the QQ plots between 'X & Z' and 'Y and Z' combinations and check whether they are the same.

The more the values are taken in 'X' and 'Y', the confidence of describing a particular distribution is higher. But due to the presence of more number of points(ie., percentiles), the plot looks dense, as if the points are overlapped. But the QQ plots become much more straight.

22.7 Chebyshev's Inequality

Let us assume, we are given a random variable 'X' which follows gaussian distribution. As we know that 68-95-99.7 rule applies for gaussian distribution,
As X~N(μ, σ),
68% of the values of 'X' lie in [μ-σ,μ+σ] => P(μ-σ <= X <= μ+σ) = 68%
95% of the values of 'X' lie in [μ-2σ,μ+2σ] => P(μ-2σ <= X <= μ+2σ) = 95%
99.7% of the values of 'X' lie in [μ-3σ,μ+3σ] => P(μ-3σ <= X <= μ+3σ) = 99.7%

If we know that 'X' follows gaussian distribution and has a finite mean and a finite variance, then we can make the above estimations.

But what if we have a finite mean 'μ' and a finite variance 'σ^2' and if we are not aware of the distribution of 'X', the Chebyshev's Inequality states that

$P(|X-\mu| >= K\sigma) <= (1/K^2)$

$(X-\mu) >= K\sigma$ means the range **$X>=(\mu+K\sigma)$** and **$X<=(\mu-K\sigma)$**
$P(X>=(\mu+K\sigma)$ and $X<=(\mu-K\sigma)) <= (1/K^2)$

Now **$P((\mu-K\sigma)<=X<=(\mu+K\sigma)) >= (1-(1/K^2))$**

Example

Let us assume we have a random variable 'X' that denotes the salaries. Let's say the given mean(μ) is 40K and the standard deviation(σ) is 10K.

1) **If 'X' follows gaussian distribution**
   According to the 68-95-99.7 rule,
   68% of the salaries lie in the interval = [μ-σ,μ+σ] = [40K-10K,40K+10K] = [30K, 50K]
   95% of the salaries lie in the interval = [μ-2σ,μ+2σ] = [40K-20K,40K+20K] = [20K, 60K]
   99.7% of the salaries lie in the interval = [μ-3σ,μ+3σ] = [40K-30K,40K+30K] = [10K, 70K]
2) **If we are not aware of the distribution of 'X'**
   Minimum Percentage of salaries that lie in the interval [μ-Kσ,μ+Kσ] >= $(1-(1/K^2))$

   Minimum Percentage of salaries that lie in the interval [μ-σ,μ+σ] >= $(1-(1/1^2)) = (1-1) = 0$
   (It means minimum 0% of the salaries lie in the interval [μ-σ,μ+σ])

   Minimum Percentage of salaries that lie in the interval [μ-2σ,μ+2σ] >= $(1-(1/2^2)) = (1-(1/4)) = (¾) = 0.75$

(It means minimum 75% of the salaries lie in the interval [μ-2σ,μ+2σ])

Minimum Percentage of salaries that lie in the interval [μ-3σ,μ+3σ] >= $(1-(1/3^2))$ = (1-(1/9)) = (8/9) = 0.89
(It means minimum 89% of the salaries lie in the interval [μ-3σ,μ+3σ])

Note:

Chebyshev's Inequality is weaker when compared to the 68-95-99.7 rule which is applied to normal distribution. The 68-95-99.7 rule gives the exact values of the percentage of points lying within a certain number of standard deviations from the mean, whereas Chebyshev's Inequality gives the minimum percentage of the points. The number of deviations 'K' accepts only integer values.

Chebyshev's inequality applies only when the given distribution has a finite mean and a finite variance.

Note: As the video lecture 22.8 is just a doubt session, we are not adding it here. All the information about Chebyshev's inequality is added in the noted for 22.7.

22.9 Discrete and Continuous Uniform Distributions

Uniform Distribution

Uniform distribution is the distribution in which the probability of occurrence of every value(either in a domain or in an interval) is the same.

Types of Uniform Distribution

There are two types of uniform distributions based on the type of random variable usd in the distribution. They are

a) Discrete Uniform Distribution
b) Continuous Uniform Distribution

Discrete Uniform Distribution

- ➢ It is a symmetric probability distribution where in, a finite number of values are equally likely to be observed. Everyone of the 'n' possible outcomes has equal probability (1/n)
- ➢ A simple example of discrete uniform distribution is throwing a fair die. The possible values are 1,2,3,4,5,6 and each time when the die is thrown, the probability of a given score is 1/6.
- ➢ If two dice are throws and their values are added, then the resulting distribution is no longer uniform because not all sums have equal probability.
- ➢ The uniform distribution has 2 parameters and is represented as U(a,b) (where b>a) and the number of possible outcomes is denoted as 'n' and is equal to b-a+1. (ie., n=b-a+1)
- ➢ In a uniform distribution, all the values are equi-probable. If there are 'n' possible outcomes, then the probability of occurrence of each value is 1/n. The possible outcomes of discrete uniform distribution are from a finite set.

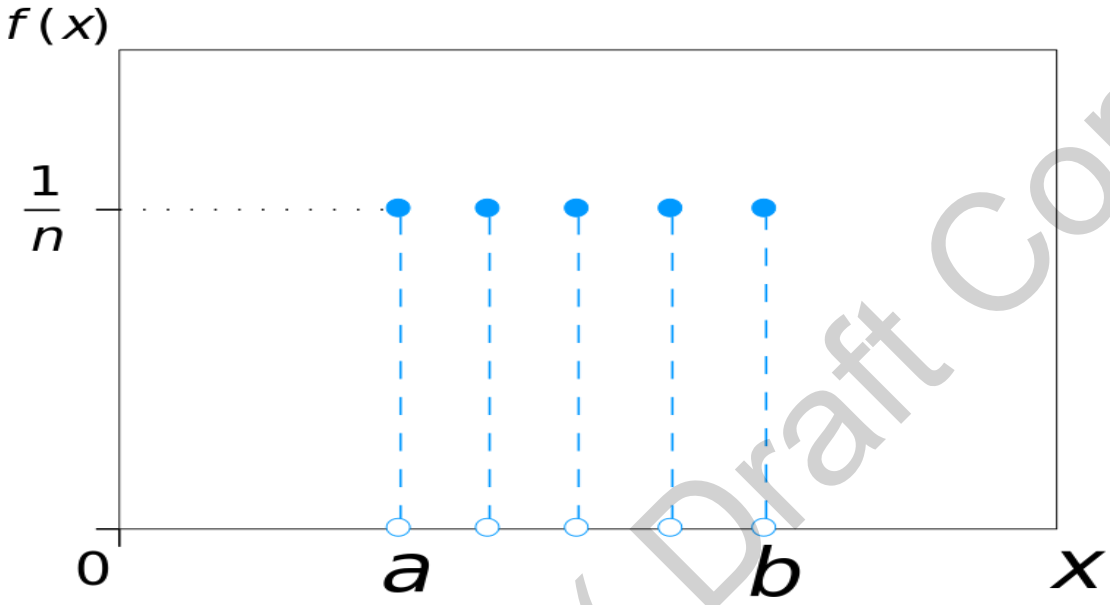
Properties of Discrete Uniform Distribution

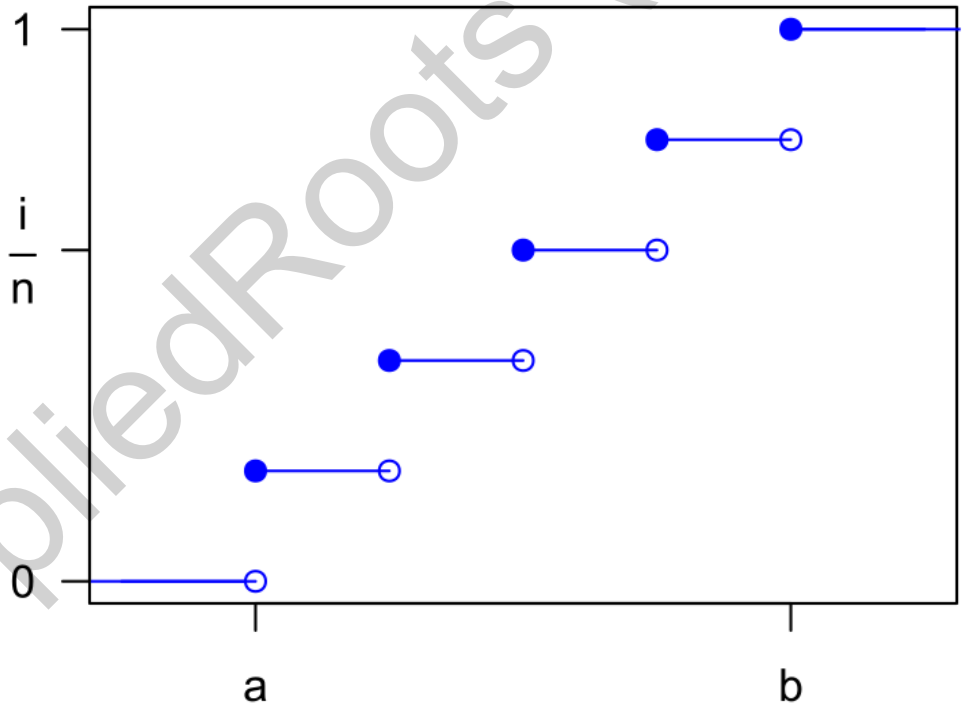
Notation: U(a,b)
Parameters: a,b (where b>=a)
PMF: 1/n
Number of Outcomes(n): b-a+1
Mean and Median: (a+b)/2
Mode: N/A
Variance: $((b-a+1)^2-1)/12$
Skewness: 0
Excess Kurtosis: $-(6*(n^2+1))/(5*(n^2-1))$

CDF(at X=k): (k-a+1)/n

PMF and CDF plots of Discrete Uniform Distribution

Below are the plots of PMF and CDF of a Discrete Uniform distribution that were discussed starting from the timestamp 1:10





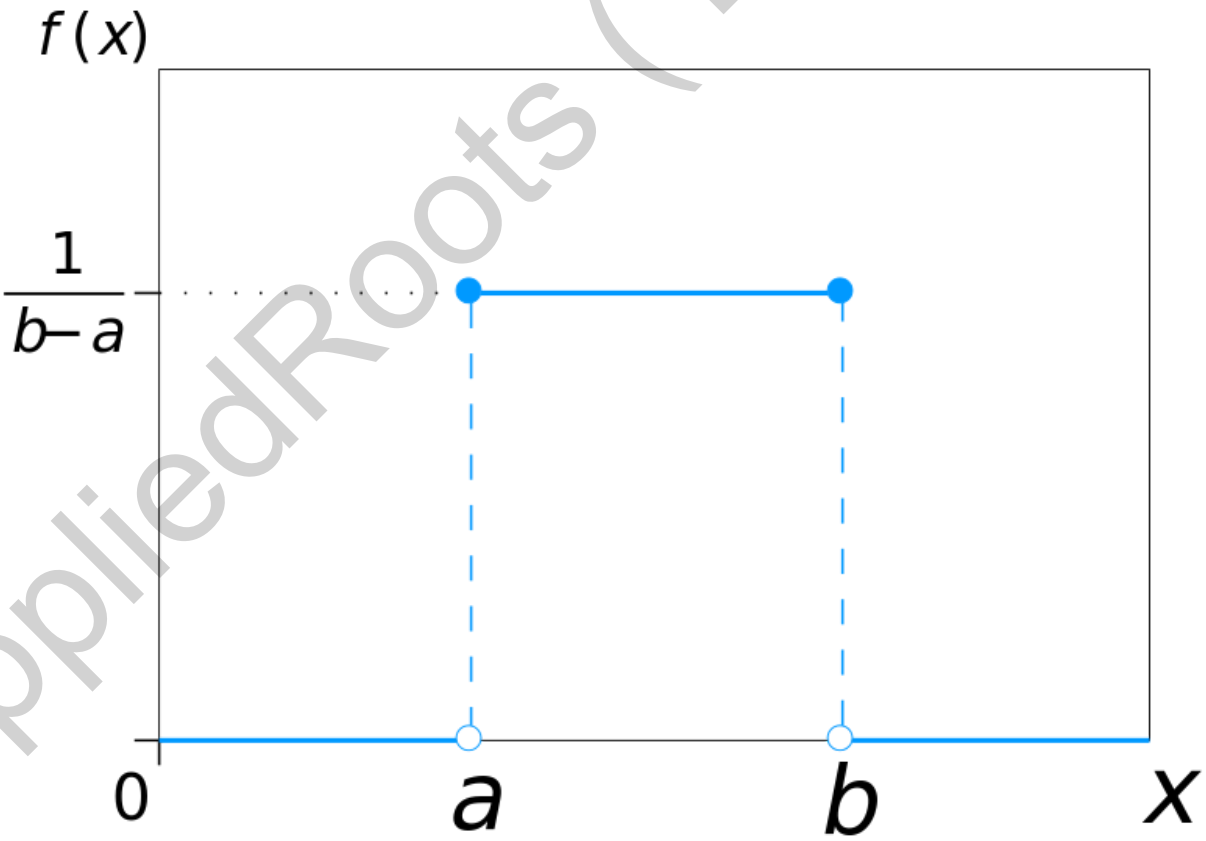
Continuous Uniform Distribution

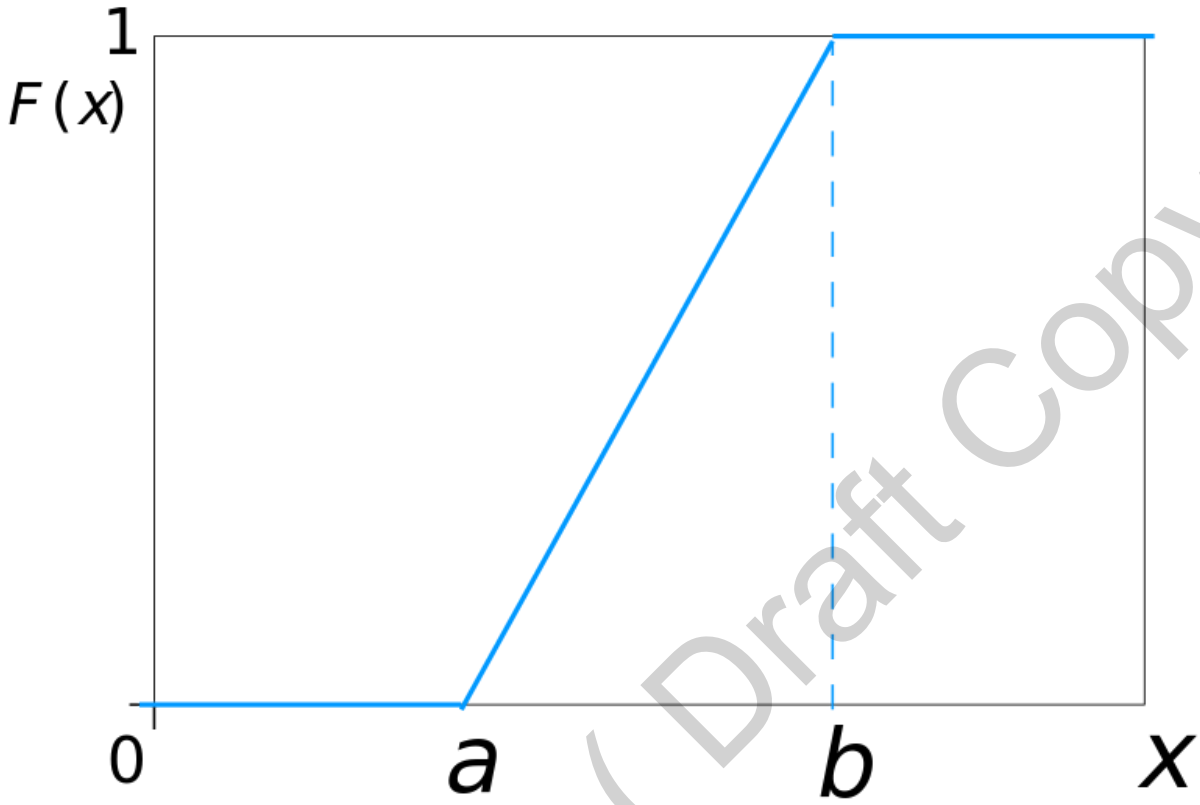
- If a random variable is continuous and follows uniform distribution, then that distribution is called Continuous Uniform Distribution.
- Let us assume the minimum value of the distribution is 'a' and the maximum value of the distribution is 'b', the probability of occurrence of any value in the interval [a,b] is the same. All the possible outcomes are equally probable.

Properties of Continuous Uniform Distribution

Mean and Median: (a+b)/2
Mode: Any value in the interval (a,b)
Variance: $(b-a)^2/12$
Skewness: 0
Excess Kurtosis: -6/5
PDF(at X=x): **1/(b-a)** if $x \in (a,b)$. Otherwise **0**.
CDF(at X=x): **0** if $x<a$. **(x-a)/(b-a)** if $x \in [a,b]$. **1** if $x>b$

Plots of PDF and CDF of a Continuous Uniform Distribution





22.10 How to randomly sample data points (Uniform distribution)

Using Random Number Generator

In general, most of the random number generators generate the random numbers uniformly, unless specified.

```
import random
print(random.random())
```

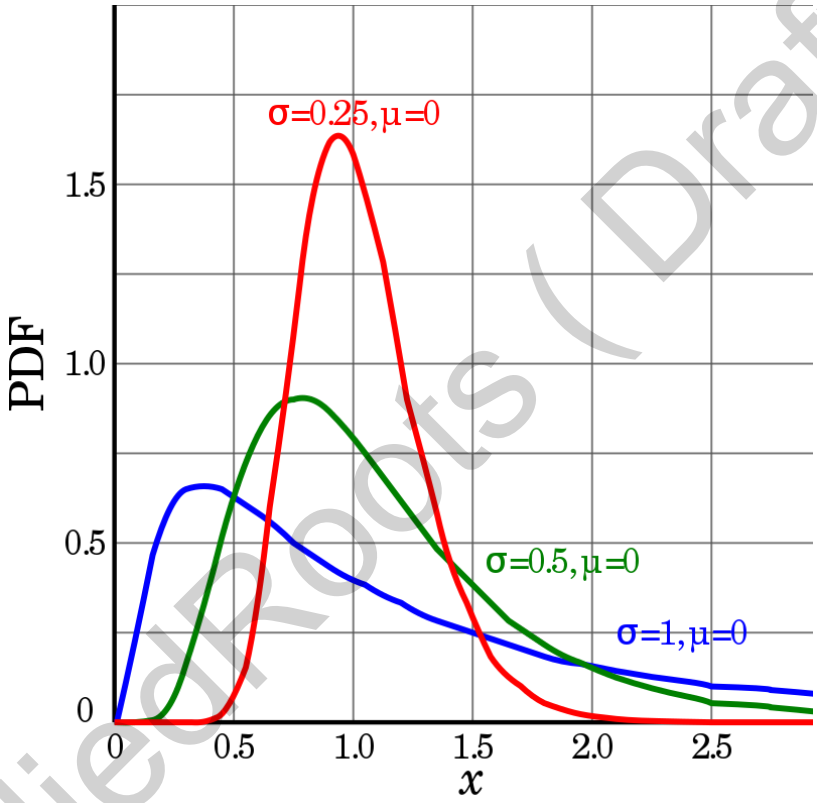
The above statement prints a random value between 0 and 1. These numbers are randomly generated with uniform distribution. Everytime we run this statement, it generates different values in the interval [0,1] with uniform distribution.

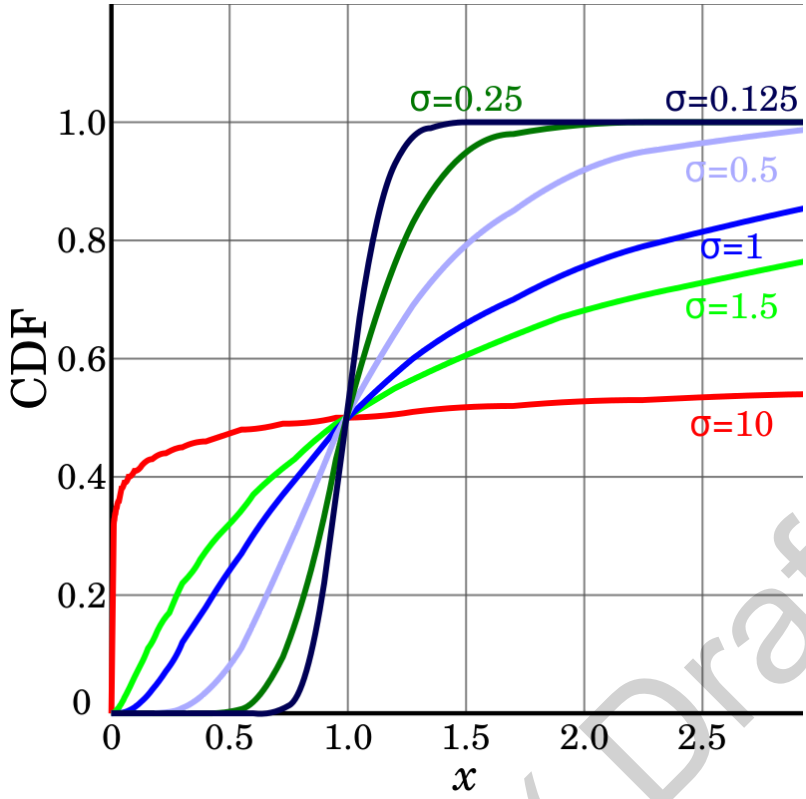
Application of Random Number Generators

Random Number Generators are used in generating continuous random numbers that are uniform in distribution. Random number generators are used in sample selection.

22.11 Log Normal Distribution

- A continuous probability distribution is said to be a log-normal, if it's natural logarithm follows a normal distribution.
- If a random variable 'X' is log-normally distributed, then Y=ln(X) has normal distribution.
- Similarly, if Y=ln(X) has normal distribution, the X = exp(Y) has log-normal distribution.
- A random variable which is log-normally distributed, takes only positive and real values. You can see the PDFs and CDFs of a few log normal distributions below with the same mean and different standard deviations, that were discussed at the timestamp 1:20.





- As the standard deviation increases, the tail factor increases and the distribution becomes skewed. (specially in log-normal distribution)
- Log Normal Distribution is always positive skewed. There is no chance of being negative skewed. Every Log Normal Distribution is a skewed distribution, but every skewed distribution is not a log normal distribution.
- The value of PDF at any point can exceed 1(as PDF is computed by adding many gaussian kernels), but the area under the curve should always be equal to 1. PDF height of more than 1, indicates the density of points in that region is heavy.
- If we have a gaussian distribution, 1σ, 2σ help us get a sense of spread of the data. If a distribution is long tailed and not gaussian, we can expect to find some large observations very far away from the mean/median. This type of information helps us get a better sense of what we can expect from our data.
- In a gaussian distribution, the peak of the PDF indicates the mean of the distribution. But in log-normal distribution, the mean is not the same as the peak of the PDF.
- For a log normal distribution, we cannot apply the 68-95-99.7 rule, as it is not a gaussian distribution. But once if we apply a natural logarithm on a log-normal distributed variable, the feature then after the transformation, follows normal distribution and then the 68-95-99.7 rule is applicable.

- In the scenarios where the log-normal distribution has its applications, we see log-normal distribution most of the time being used, whereas the gaussian distribution was less frequently used.
- If 'X' follows log normal distribution, then Y=ln(X) follows a normal distribution.
  $Y \sim N(\mu, \sigma)$
  According to normal distribution,
  68% of the data lies in $[\mu-\sigma, \mu+\sigma]$
  95% of the data lies in $[\mu-2\sigma, \mu+2\sigma]$
  99.7% of the data lies in $[\mu-3\sigma, \mu+3\sigma]$
  As $Y=\ln(X) \Rightarrow X = e^Y$
  So 68% of the data in 'X' lies in $[e^{\mu-\sigma}, e^{\mu+\sigma}]$
  So 95% of the data in 'X' lies in $[e^{\mu-2\sigma}, e^{\mu+2\sigma}]$
  So 99.7% of the data in 'X' lies in $[e^{\mu-3\sigma}, e^{\mu+3\sigma}]$
  Here μ,σ are the parameters of 'Y'.
- The values in a log normal distribution are positive and hence they are in right skewed form. In case, if we come across any negative values in the log-normal distributed values, we can convert all of them into non negative values, by adding some constant 'a', to all the values. This constant 'a' can be the largest magnitude among all the negative values.

How to check if a given distribution is log-normal

Let 'X' be the given input distribution. Let us compute the natural logarithm for the values of 'X' and let them be denoted as 'Y'. (ie., $Y = \ln(X)$)

Then we should go for the QQ plot with 'Y' values on the 'Y' axis and a randomly generated normal distribution $N(\mu', \sigma'^2)$ on the 'X' axis.

If the plot looks like a straight line, then we can confirm that 'Y' is normally distributed and 'X' is log normally distributed.

Why is Log Normal Distribution having skewness

It is because of the exponential function applied on the top of the gaussian distribution. If we generate some data points from a gaussian random variable and apply exponential function on top of them, plot the resultant values, it looks like a log-normal distribution. Log Normal distribution gets more skewed if the standard deviation 'σ' increases. Larger 'σ' spreads the points wider in gaussian distribution.

Applications of Log Normal Distribution

1) The length of comments posted in internet discussion forums follow log normal distribution
2) The time spent by the users on the internet to read articles/blogs, etc also follows log normal distribution.

Why is the peak of a log normal distribution not considered as it's mean?

Normal/Gaussian distribution is symmetric and 50% of the values lie on one side and the remaining 50% of the values lie on the other side. So we can say the highest peak of a gaussian distribution is it's mean.

But in case of a log normal distribution, we can't guarantee that the mean is exactly present at the highest peak, as the curve is asymmetric.

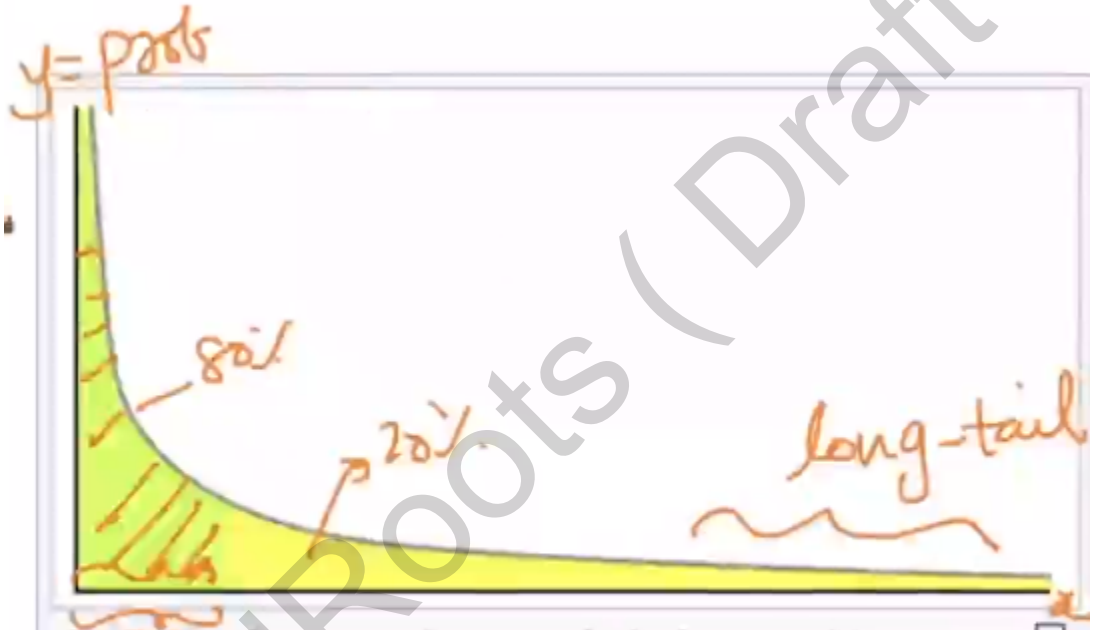
Properties of Log Normal Distribution

Notation: $X \sim \text{lognormal}(\mu, \sigma^2)$
Parameters: $\mu \in (-\infty, \infty)$, $\sigma > 0$
Support(Region in which this distribution is applicable): $x \in (0, \infty)$
PDF(at X=x): $\frac{1}{x\sigma\sqrt{2\pi}} * \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$
CDF(at X=x): $\frac{1}{2}+\left[\frac{1}{2} * \text{erf}\left(\frac{\ln(x)-\mu}{\sigma\sqrt{2}}\right)\right]$
Mean: $\exp(\mu+(\sigma^2/2))$
Median: $\exp(\mu)$
Mode: $\exp(\mu-\sigma^2)$
Variance: $[\exp(\sigma^2)-1]*\exp(2\mu+\sigma^2)$
Skewness: $(\exp(\sigma^2)+2)*\sqrt{\exp(\sigma^2)-1}$

22.12 Power Law Distributions

A power law is a functional relationship between two quantities where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities. (ie., one quantity varies as the power of the other).

Power Law follows the 80-20 rule. That in the given distribution 'X', 80% of the values of the distribution lie below 20% of the values of 'X'.Whenever a distribution follows Power Law, that distribution is called **Pareto Distribution**. There is a long tail for Power Law functions. Pareto Distributions are for the continuous random variables. You can find the PDF of a Power Law distribution below, which was discussed starting from the timestamp 0:20.



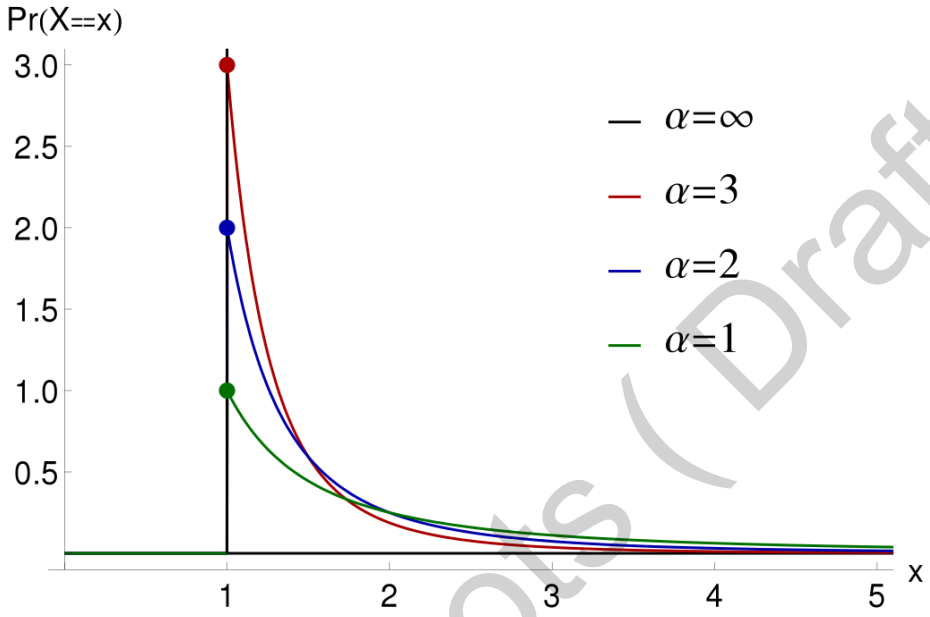
An example power-law graph, being used to demonstrate ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).

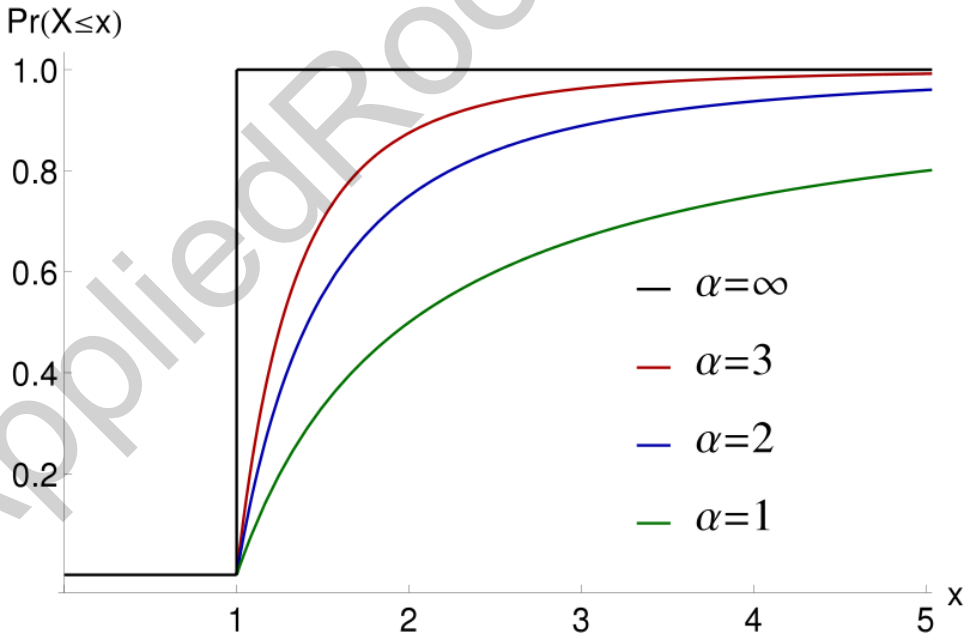
Pareto Distribution

The parameters of the Pareto distribution are

1) $x_m>0$. This parameter is called 'scale' and takes only real values. It is similar to 'μ' in a gaussian distribution.
2) α>0. This parameter is called 'shape' and takes only real values. It is similar to 'σ' in a gaussian distribution.

Let us look at the PDF and CDF plots of a pareto distribution that were discussed starting from the timestamp 2:35.





From the PDF plots, we observe that as the ‘α’ value keeps decreasing, the tails become less fatter. For α → infinity, the PDF becomes a delta function(ie., a straight vertical line with a single value). Here this delta function has a value only at one point, whereas on the other points, it takes the value as 0. Such a function is called **Dirac Delta Function**.

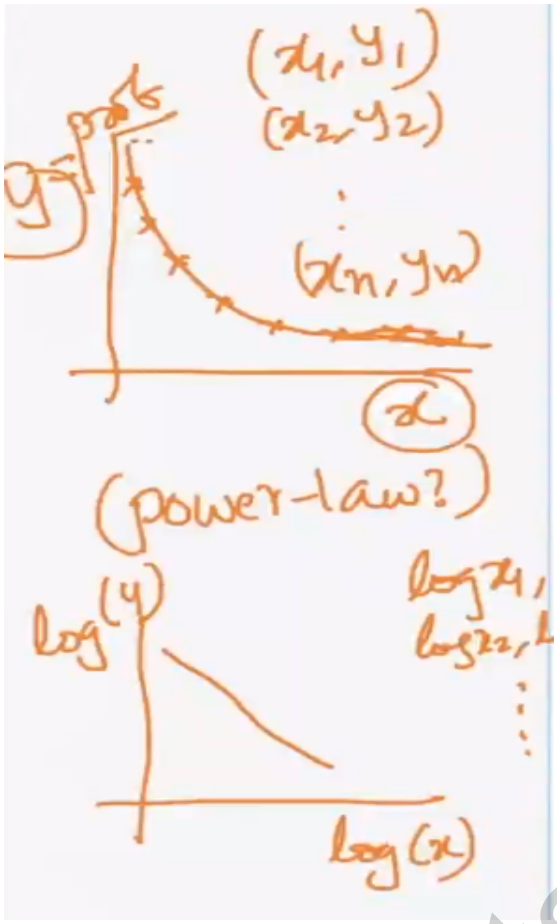
The common point in both Pareto and Log Normal Distributions is “Both the distributions have a small number of larger values and a large number of smaller values. But the major difference is the Pareto Distribution does not have an increasing PDF.

How to check if the two given variables follow Power Law?

The one approach to check if the two given variables follow Power Law is using Log-Log plot.

If ‘X’ and ‘Y’ are two given variables, then if we make a plot with Log(X) on X-axis and Log(Y) on Y-axis and if the plot converges to a straight line as shown in the figure, then we can say that the distribution has power law tail. (ie., both the variables follow power law)

A straight line on a log-log plot is a strong evidence for power laws, and the slope of the straight line corresponds to the power law exponent.



Disadvantages of Log-Log Plot

1) Log-Log plot is insufficient evidence for a power law relationship as many non power law distributions also appeared like straight lines on log-log plots.
2) In order to provide reliable results using a log-log plot, we need huge amounts of data.
3) Log-Log plots are appropriate for discrete data.

How to check whether a given distribution is a Pareto Distribution?

Let us assume 'X' is a continuous random variable and we have to check whether it has Pareto Distribution using QQ plot.

We have to take a reference Pareto distribution with continuous random variable 'Y'. We shall calculate the percentile values of 'X' and 'Y' and then we have to plot the points $(y^{(i)}, x^{(i)})$ where

$y^{(i)} \rightarrow i^{th}$ percentile of 'Y'

$x^{(i)} \rightarrow i^{th}$ percentile of 'X'

If the result is a straight line, then we can say both follow the same distribution.

20.13 Box Cox Transform

So far we have seen random variables following log-normal distribution and when we apply natural logarithm on it, it turns into gaussian distribution. We generally expect the numerical features to follow gaussian distribution because most of the ML models work on the assumption that all the numerical features in the data follow gaussian distribution.

Power Transform is a data transformation technique used to stabilize variance, make the data more like a normal distribution and improve the validity of measures of association between variables and for other stabilization procedures. One such popular transform is Box Cox Transform.

Box Cox transform is a power transform technique used to convert a non gaussian distribution into a gaussian distribution.

Procedure of Box-Cox Transform

Let us assume, we have a random variable 'X'(say it follows pareto distribution) and we want to transform it into a gaussian distribution denoted by 'Y'.

Let $X = [x_1, x_2, x_3, \ldots\ldots, x_n]$ is a pareto distribution. The output should be $Y = [y_1, y_2, y_3, \ldots.., y_n]$ that follows a gaussian distribution. (Here 'X' can be any non gaussian distribution. It is not mandatory for 'X' to be a pareto distribution).

1) When we apply box-cox transform on a random variable 'X', then it returns 'λ'.
   Box-Cox(X) = λ
2) Compute $\mathbf{yi = (x_i^{\lambda}-1)/\lambda}$, if λ≠0 (or) $\mathbf{\log_e(xi)}$ if **λ=0**.($\forall$ i $\rightarrow$ 1 to n)
   If λ=0, then the given original distribution is log-normal.

In the conversion process, if we get λ=0, then it means 'X' is following log-normal distribution and we have to convert 'X' into gaussian form by applying natural logarithm on every element.

If we get λ≠0, we can go ahead in converting 'X' into gaussian form by the formula

$$\mathbf{yi = (x_i^{\lambda}-1)/\lambda}$$

Note: Box Cox Transform is not guaranteed to work on all pareto or power law distributed data. It works only on some of them and we need to perform the box-cox transform and observe the QQ plot and confirm if it is working well on our data.

We are finding the best 'λ' using the data we currently have. If the data we have is a good representative sample of all the possible values, the 'λ' we computed using the current data would be reasonably food for future unseen data also.

Given any new data, we can apply the transformation using the 'λ' already obtained to transform the new data into gaussian distributed data. Box Cox Transform is a one-one transform and we can obtain the original data back from the transformed data.

λ value	Transformation used in Box Cox Transform
λ = -1	Reciprocal Transform
λ = -0.5	Reciprocal Square Root Transform
λ = 0	Log Transform
λ = 0.5	Square Root Transform
λ = 1	No Transform

Below is the code snippet that is used for applying box-cox transform on a given distribution of the data and it was discussed at the timestamp 7:05.

```
>>> from scipy import stats
>>> import matplotlib.pyplot as plt
```

We generate some random variates from a non-normal distribution and make a probability plot for it, to show it is non-normal in the tails:

```
>>> fig = plt.figure()
>>> ax1 = fig.add_subplot(211)
>>> x = stats.loggamma.rvs(5, size=500) + 5
>>> prob = stats.probplot(x, dist=stats.norm, plot=ax1)
>>> ax1.set_xlabel('')
>>> ax1.set_title('Probplot against normal distribution')
```

We now use boxcox to transform the data so it's closest to normal:

```
>>> ax2 = fig.add_subplot(212)
>>> xt, _ = stats.boxcox(x)
>>> prob = stats.probplot(xt, dist=stats.norm, plot=ax2)
>>> ax2.set_title('Probplot after Box-Cox transformation')

>>> plt.show()
```

Note: As the 20.14 video lecture was only about examples, we are not adding any notes for it. You can just go through the video just to get an idea of it.

20.15 Covariance

In our dataset, sometimes there exists relationships between two or more variables(ie., increase/decrease in the value of a variable may decrease/increase the values of another variable). In order to quantify these relationships between the variables, we have 3 measures.

1) Covariance
2) Pearson Correlation Coefficient
3) Spearman Rank Correlation Coefficient

Covariance

Let us assume we have 'N' observations with random variables 'X' and 'Y' in pairs. Then the covariance between them is defined as

$$\textbf{Covariance(X,Y)} = (1/N)*\Sigma^{N}_{i=1} (x_i-\mu_x)*(y_i-\mu_y)$$

Let us recall the variance formula.

$$\textbf{Variance(X)} = (1/N)*\Sigma^{N}_{i=1} (x_i-\mu_x)*(x_i-\mu_x)$$

When we compare both the equations, then we can conclude that

$$\textbf{Covariance(X,X) = Variance(X)}$$

Covariance measures only the linear relationships between the random variables 'X' and 'Y'. So when we take the covariance between two variables 'X' and 'Y', we have

$$\textbf{Covariance(X,Y)} = (1/N)*\Sigma^{N}_{i=1} (x_i-\mu_x)*(y_i-\mu_y)$$

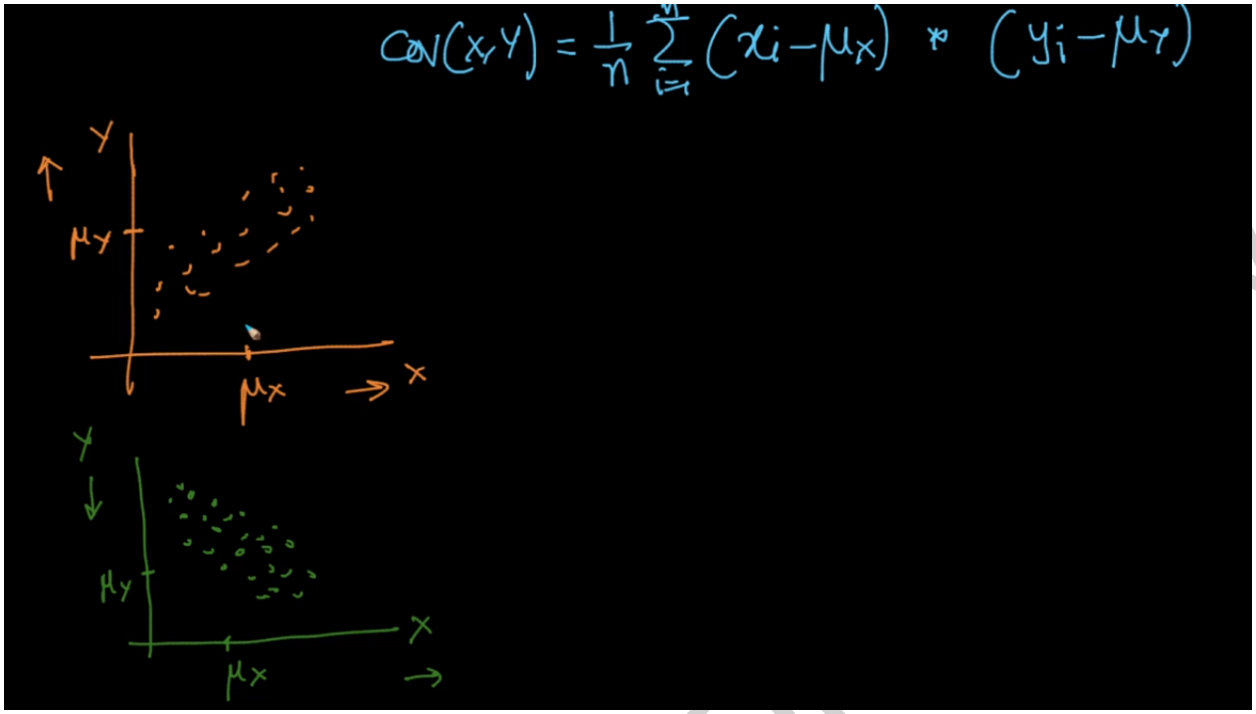
The covariance between the two random variables 'X' and 'Y' is said to be positive only

a) If the value of 'X' increases with an increase in the value of 'Y'.
b) If the value of 'X' decreases with a decrease in the value of 'Y'.

Similarly the covariance between the two random variables 'X' and 'Y' is said to be negative only

a) If the value of 'X' increases with a decrease in the value of 'Y'.
b) If the value of 'X' decreases with an increase in the value of 'Y'

Below are the plots that show how the relationship between two variables exists for positive and negative covariances. It was discussed starting from the timestamp 6:20

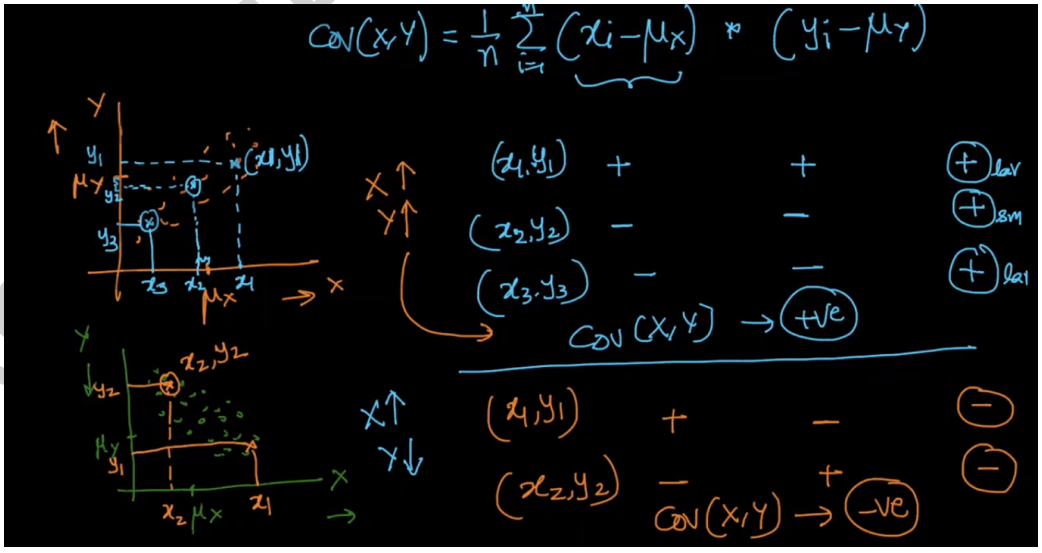


In the above picture, the orange plot shows the existence of the positive covariance and the green plot shows the existence of the negative covariance between the random variables ‘X’ and ‘Y’.

In the orange plot, as the value of ‘X’ increases, the value of ‘Y’ also increases. It indicates positive covariance between ‘X’ and ‘Y’. In the green plot, as the value of ‘X’ increases, the value of ‘Y’ decreases. It indicates negative covariance between ‘X’ and ‘Y’.

Graphical Representation

Below are the plots that were discussed starting from the timestamp 6:20.



Pairs(X,Y)	Sign of $(x_i-\mu_x)$	Sign of $(y_i-\mu_y)$
(x_1, y_1)	+ve	+ve
(x_2, y_2)	-ve	-ve
(x_3, y_3)	+ve	-ve
(x_4, y_4)	-ve	+ve

In the first two combinations, both the terms(ie., $(x_i-\mu_x)$ and $(y_i-\mu_y)$) are positive, so the covariance becomes positive. In the 3rd and 4th combinations, one of the terms is negative and the other term is positive, so the covariance becomes negative, as the product of these two terms is negative.

Disadvantage of Covariance

When we want to find out the covariance between two variables 'X' and 'Y', if the scale/metric changes, then the covariance score changes drastically. This drawback is overcome by Pearson Correlation coefficient.

Whenever we have 2 variables in different units(ex: kg and pounds), if we want to compute covariance, we first should bring the variables onto the same units, and then compute the covariance.

Whenever the covariance between two random variables is 0, then it doesn't mean that the two random variables are independent of each other. It only means that there doesn't exist any linear relationship between the two random variables. There might exist a non linear relationship between these two random variables.

Significance of Magnitude and Sign of Covariance

The sign of the covariance shows the tendency in the linear relationship between the variables. The magnitude of the covariance is not easy to interpret because it is not standardized and hence it depends on the magnitudes of the variables.

The standardized version of the covariance shows the strength of the linear relationship by its magnitude. If we standardize the data, then the covariance becomes correlation.

If the scale of the variables change, only the magnitude of the covariance changes, but not the sign of the covariance. The sign of the covariance remains the same. When it comes to checking whether the correlation between two random variables is positive or negative, we have to focus only on the sign of the covariance, but not on the magnitude.

When we have to check how strongly/weakly the variables are correlated, we have to consider the magnitude also. For example, if we have 3 random variables 'X', 'Y' and 'Z', then if a small change in 'X', results in a drastic change in 'Y', then we say that 'X' and 'Y' are strongly correlated. If a small change in 'X', results in a small change in 'Z', then we say that 'X' and 'Z' are weakly correlated.

So whether two random variables are strongly correlated or weakly correlated, is decided by the magnitude of the covariance. Having a large magnitude indicates a strong correlation and a small magnitude indicates a weak correlation. In this case, the magnitude of the covariance plays a key role.

Similarly, if we have a positive covariance, then we can say that correlation between two quantities is positive. If we have a negative covariance, then we can say that correlation between two quantities is negative. In this case, the sign of the covariance plays a key role.

20.16 Pearson Correlation Coefficient(ρ)

Pearson Correlation Coefficient is a measure of the linear correlation between two random variables 'X' and 'Y'. It has a value between -1 and 1.

1 → Total Positive Linear Correlation
0 → No Linear Correlation
-1 → Total Negative Linear Correlation

Pearson Correlation Coefficient is the covariance of the two variables divided by the product of their standard deviations.

Pearson Correlation Coefficient when applied to a population is commonly represented by 'ρ' and may be referred to as the Population Correlation Coefficient (or) Population Pearson Correlation Coefficient.

Given a pair of random variables (X,Y), the formula for 'ρ' is

$$\rho_{x,y} = (1/(\sigma_x * \sigma_y)) * \text{Covariance}(X,Y)$$

σ_x → standard deviation of 'X'
σ_y → standard deviation of 'Y'

Pearson Correlation Coefficient when applied to a sample is commonly represented by 'r' and may be referred to as the Sample Correlation Coefficient (or) Sample Pearson Correlation Coefficient.

Given a sample of points for random variables (X,Y), the formula for 'r_{xy}' is given as

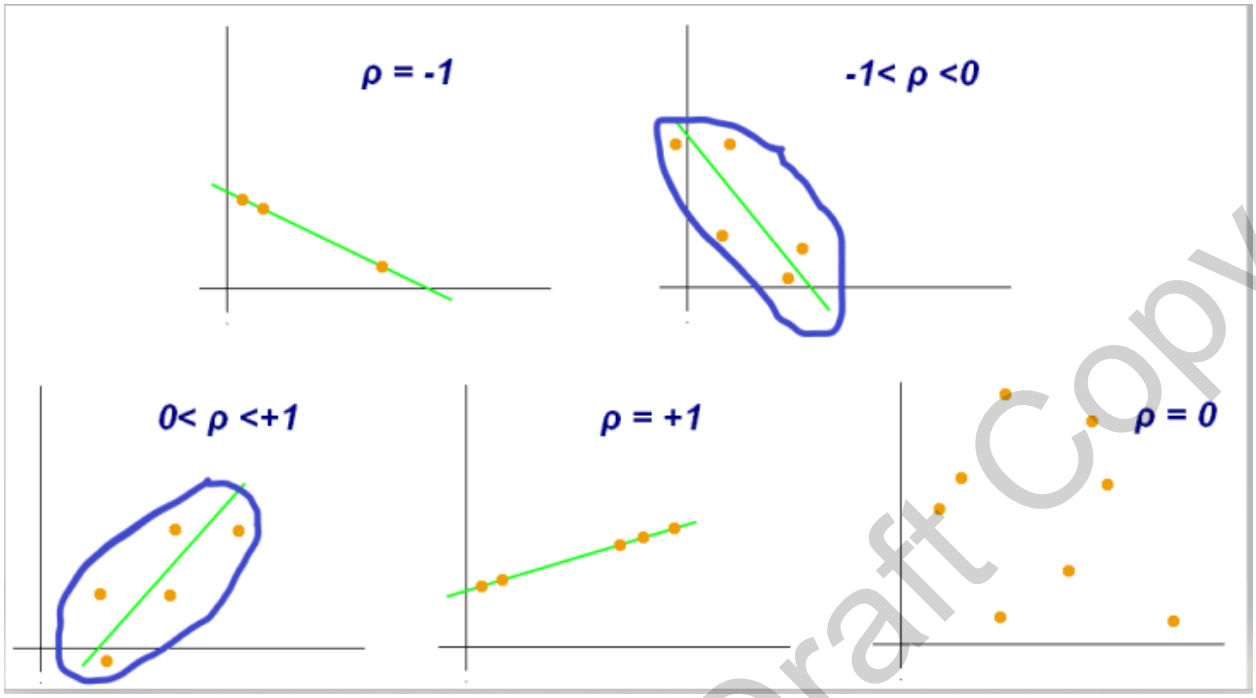
$$r_{xy} = (\Sigma^{n}_{i=1} (xi-\bar{x})*(yi-\bar{y}))/(\text{sqrt}(\Sigma^{n}_{i=1} (xi-\bar{x})^2)*\text{sqrt}(\Sigma^{n}_{i=1} (yi-\bar{y})^2))$$

n → number of points in the sample
$\bar{x}$ → sample mean of 'X'
$\bar{y}$ → sample mean of 'Y'

The problem with the Covariance is that it doesn't take the standard deviations of the random variables into consideration, whereas the Pearson Correlation Coefficient takes the standard deviations into consideration.

Different cases that show how the value of the Pearson Correlation Coefficient Changes

Below are the cases that show how the PCC value changes with different correlations among the data. It has been discussed starting from the timestamp 2:00



If the points are more spread on both sides of a straight line, but still in an oval shape, the 'ρ' value lies in between (-1,0) or (0,1). If all the points lie on a straight line, then only we can expect the 'ρ' value to be either +1 or -1.

There is also a limitation with PCC. Since both Covariance and PCC are biased towards the linear relationships, we could get a better score of PCC, when the points lie on a straight line. If the points are little dispersed, we get a low score of PCC.

We also can see different scenarios of correlation, that was discussed starting from the timestamp 5:15

a) Type 1 (on the basis of the deviation from the straight line)

The maximum variance is seen when ρ=0 and the least variance is seen when ρ=1 and ρ=-1

b) Type 2 (on the basis of the slope)

The value of ‘ρ’ will be either 1 or -1, if there exists a linear relationship between ‘X’ and ‘Y’. It doesn’t bother about the angle the lines make with the ground. It only bothers about whether a linear relationship exists between the variables or not.

c) Type 3 (on the basis of linearity)

In all the above nonlinear relationships, **ρ=0** because PCC cannot find the correlation if both the given variables are non linear. In order to estimate non linear distributions, we have to build a full fledged model.

Note:

Covariance and PCC do not work for nonlinear relationships. When it comes to linear relationships, Covariance doesn’t take the variability into consideration, due to which we see huge differences in the magnitude of covariance when the distance of the points from the mean varies. Whereas PCC takes the standard deviation into consideration.

We cannot make any inferences on the basis of the covariance score, whereas we can make inferences on the basis of the PCC score.

PCC = 1 indicates a perfect positive linear relationship between the features.

PCC = -1 indicates a perfect negative linear relationship between the features.

In the above two cases, all the points fit exactly on a straight line. But in real time, it is always not possible for all the points to fit on a straight line.

PCC cannot handle monotonic functions such as non decreasing or non increasing functions because it can handle only linear relationships whereas Spearman’s Rank Correlation Coefficient (SRCC) can handle monotonic functions. When the data doesn’t follow normal distribution, then SRCC may be more appropriate. Since PCC works only if the relationship is linear, it doesn’t give an appropriate value when the two given random variables ‘X’ and ‘Y’ are not linearly related. If ‘X’ and ‘Y’ are not linearly related, then **Covariance(X,Y) = PCC(X,Y) = 0**.

20.17 Spearman Rank Correlation Coefficient (SRCC)

Pearson Correlation Coefficient is useful when the relationship between two random variables is linear. When we have non linear relationships (like monotonically increasing/ decreasing/ non increasing / non decreasing functions), the best measure for correlation is Spearman Rank Correlation Coefficient.

Let us look at the below example that was discussed starting from the timestamp 1:00..

X	Y	Rank of values of 'X' (r_X)	Rank of values of 'Y' (r_Y)
160	52	4	3
150	66	2	4
170	68	5	5
140	46	1	1
158	51	3	2

The value with the least number will have less rank. Instead of computing the PCC on 'X' and 'Y', SRCC says to compute the PCC between the ranks of 'X' and 'Y'.

SRCC(r) = $\rho_{rX, rY}$

In case of SRCC

- If 'Y' increases with an increase in 'X', irrespective of whether the relationship between 'X' and 'Y' is linear or not, the value of SRCC = 1.
- If 'Y' decreases with a decrease in 'X', irrespective of whether the relationship between 'X' and 'Y' is linear or not, the value of SRCC = 1.
- If 'Y' increases with a decrease in 'X' (or) if 'Y' decreases with an increase in 'X', then irrespective of whether the relationship between 'X' and 'Y' is linear or not, the value of SRCC = -1.

In case of PCC

- If 'Y' increases with an increase in 'X' (or) if 'Y' decreases with a decrease in 'X', then if the relationship between 'X' and 'Y' is linear, then PCC = 1.
- If 'Y' increases with a decrease in 'X' (or) if 'Y' decreases with an increase in 'X', then if the relationship between 'X' and 'Y' is linear, then PCC = -1.
- If the relationship between 'X' and 'Y' is non linear, then PCC is not applicable.
- If $\rho_{X,Y}$ = 0, it doesn't mean that there is no relationship between 'X' and 'Y'. It only means that 'X' and 'Y' are not linearly related to each other.
- Similarly if r = $\rho_{rX, rY = 0}$ = 0, it doesn't mean that there is no relationship between 'X' and 'Y'. It only means that 'X' and 'Y' are not monotonically related to each other.
- While assigning the ranks to the data of random variables while estimating SRCC, if two or more numbers are same, then some implementations use the

same rank and some other implementations use fractional ranks.

For example, if the rank ‘5’ is repeated two times for two numbers then we skip the rank ‘6’ and give the next entity, a rank of ‘7’. For these two values, we distribute the ranks between ‘5’ and ‘6’, which is 5.5, to both of them. This is called Fractional Ranking and it is mostly used.

- ➢ SRCC is the PCC of ranks. If the ranks have a covariance of 0, then the SRCC also will be 0. So when the PCC of ranks is 0, then the SRCC also will become 0.
- ➢ The assumptions of SRCC are that data must be at least ordinal and the scores on one variable must be monotonically related to the other variable. In an ordinal scale, the levels of variables are ordered such that one level can be considered higher/lower than another.
- ➢ SRCC works only if two variables are monotonically increasing or decreasing, but not both at a time. In case, if we have both monotonically increasing and monotonically decreasing natures together, then we should split the function into intervals and then apply SRCC on each nature separately.

Disadvantages of SRCC

SRCC cannot completely overcome the problem of non-linear relationships, as it uses only the information of the order through ranks.

For example, if we take Y=Sin(X), if we plot ‘X’ (vs) ‘Y’, visually we’ll be able to see a correlation between ‘X’ and ‘Y’. But SRCC fails to understand the correlation between the values as it is concerned only with the ordering using the ranks and monotonicity.

We are losing some information by taking the ranks into consideration, but SRCC is still used in real world scenarios where we do not care about exact numeric values, but only about the order.

If SRCC(X,Y) = 0, then it only means ‘X’ and ‘Y’ are monotonically independent. They also could be related in non-monotonic ways like Y = Sin(X) (or) Y = Cos(X), etc.

20.18 Correlation vs Causation

Correlation doesn't imply causation. We sometimes see as 'X' increases, 'Y' also increases and as 'X' decreases, 'Y' also decreases. Here due to the presence of correlation between 'X' and 'Y', we cannot say that 'X' causes 'Y'.

Even though the correlation between 'X' and 'Y' is very high, we cannot say 'X' causes 'Y'. In order to confirm whether 'X' causes 'Y', we need to have the domain expertise. If 'X' and 'Y' are correlated, then sometimes 'X' may cause 'Y' and sometimes it may not.

Correlation doesn't imply Causation all the time. It implies only sometimes.

20.19 How to use Correlations?

Example 1: Is salary correlated with the square feet of home?

Here high salaried people look for homes with higher square feet. It is not always mandatory for high salaried employees to purchase only homes with higher square feet, but mostly the high salaried people prefer higher square feet houses. Also if the people are highly salaried, then if they prefer higher square feet houses, it makes the business easier and gain huge profits. Also if we mention our salary range, then the real estate company can estimate what kind of house we are interested, depending on the budget and shows us only those houses.

Example 2: Is the number of years of education correlated with the income?

For the education or labour department, if they know that the number of years of education fetches more income, then the department makes sure every individual gets as many number of years of education in order to generate more income.

Example 3: Is the time spent on an e-commerce web page in the last 24 hours correlated with money spent on the website?

From the data we have, for multiple users, if we can draw the conclusions that the total money spent on the website by each customer increases, as the time spent by him/her browsing the web page increases, then the e-commerce company can design the website in a much more attractive way such that the customers spend more time browsing it and spend more money.

Also one more correlation can be made between the number of unique customers (vs) the total sales. If the number of unique customers increases, the total sales also increase. So the business focuses on increasing the number of customers, so that it could increase the sales.

Note

If the correlation coefficient is positive, then the two variables are directly proportional. If the correlation coefficient is negative, then the two variables are inversely proportional.

First of all, we have to check if correlation exists between the variables. Once the presence of correlation is confirmed, then we can perform tests for the presence of causality.

Normalizing the covariance to the range [-1,1] gives the correlation.

Causation indicates a relationship between two events where one event is affected by the other. In statistics, when the value of one event or variable increases or decreases, as a result of other events, then there is said to be causation.

20.20 Confidence Interval (CI): An Introduction

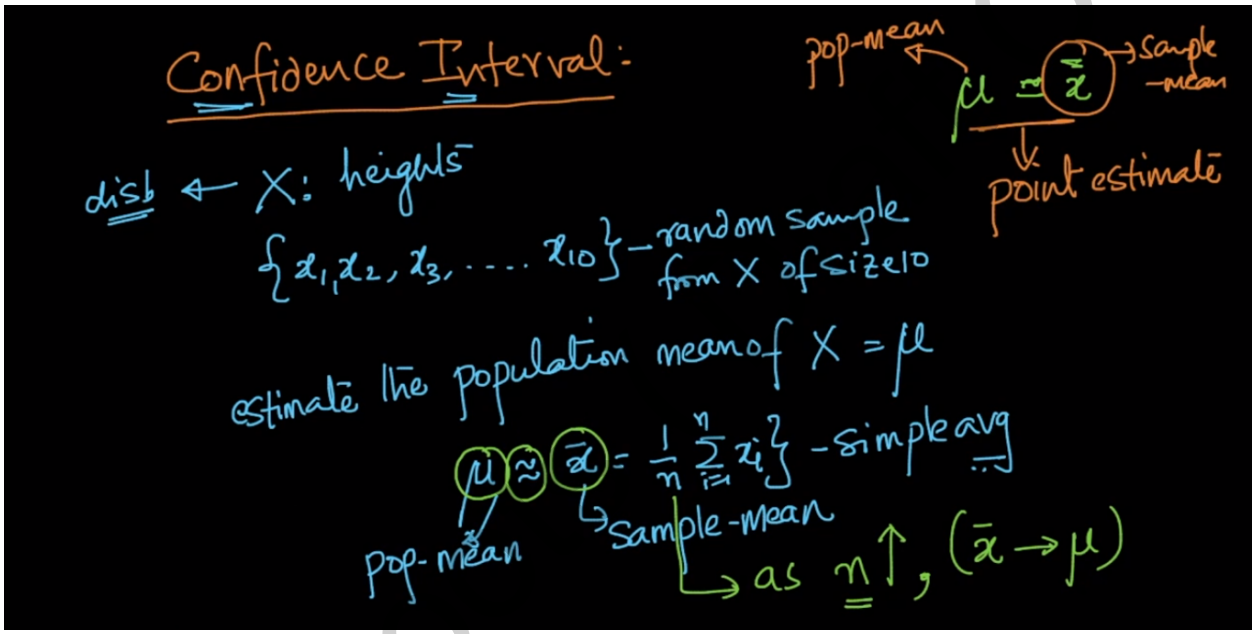
Let us assume we have a random variable 'X' that represents the heights of the people. Let us assume we have a sample of 10 points.

X = {x1, x2, x3, …., x10} (random sample of size 10)

Our job is to estimate the population mean 'μ' of 'X'.

$\mu = \bar{x} = (1/n) * \Sigma^{n}_{i=1} x_i$ ($\bar{x}$ → sample mean, 'n' → sample size)

As the value of 'n' increases, the sample mean becomes more and more nearer to the population mean. If the estimate of the population mean(μ) is exactly equal to the sample mean($\bar{x}$), then it is called **point estimate**.



Point Estimate of Population Mean(μ) = Sample Mean($\bar{x}$) = $(1/n) * \Sigma^{n}_{i=1} x_i$ ('n' → sample size)

In case, if we can make a statement that the population mean(μ) lies in an interval [a,b] with 'x%' probability, then it means we are saying with x% confidence that the population mean lies in the interval [a,b].

M∈[a,b]. Here 'x%' is the percentage of confidence.

[a,b] is the interval estimate of the population mean and the technical name given to this interval is called **Confidence Interval**.

In the above example, we can say that in 'x%' of the cases, the population mean(μ) lies in the interval [a,b] and in the remaining '(100-x)%' of the cases, the population mean(μ) lies outside the interval [a,b].

Making an interval estimate is much more richer in terms of population estimate when compared to a point estimate.

Need for Confidence Interval over the point estimates

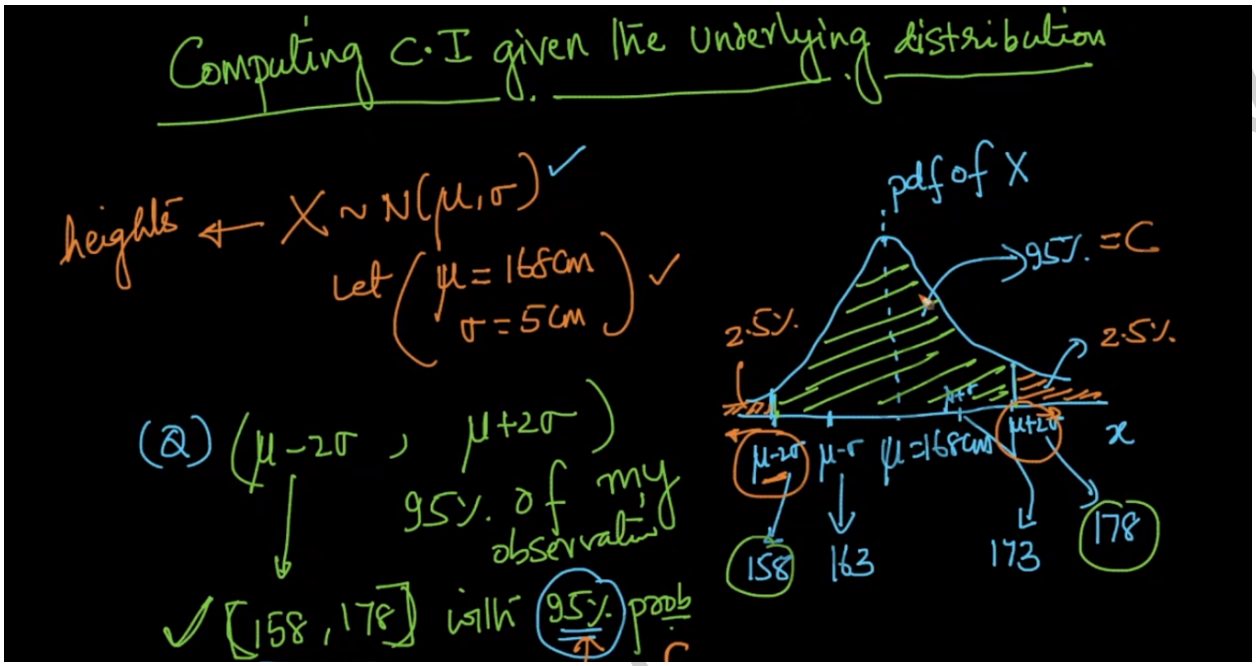
It is always not possible to consider the whole data and find the mean and standard deviation values in real time, so we work with samples many times.

The reason why we prefer interval estimates over the point estimates is that, every time if we keep changing the points in the samples, then the point estimate (ie., sample mean) keeps changing every time. So it is better to give an interval estimate than a point estimate and hence we choose confidence intervals.

Interval estimates are preferred by the statisticians because the interval estimates are accompanied by a statement concerning the degree of confidence that the interval contains the population parameter being estimated.

It is desirable for a point estimate to be consistent so the larger the sample size, the more accurate the estimate would be (or) we can say the population mean is nearly equal to the sample mean, but in general the sample mean varies from one sample to another.

20.21 Computing the Confidence Interval given the underlying distribution



Let us assume we have a random variable 'X' and it follows normal distribution. X~N(μ, σ)

From the 68-95-99.7 rule, we know that 95% of the values lie in the interval [μ-2σ, μ+2σ]. In this case, we shall denote 95% with 'C'.

So **C = 95%**. 'C' denotes confidence.

In case, if we have to find the confidence interval of 90% of the values, then **C = 90%**. In order to compute the confidence intervals for the values C = 90%, 80%, 70%, etc, we have to check the normal distribution table.

In case, if the given distribution is not a gaussian distribution and we are asked to find out the confidence interval, then for the confidence interval estimation of non gaussian distribution, we have to go for **Bootstrapping**.

If the given distribution is gaussian, then irrespective of the number of points in the distribution, the confidence intervals are the same. That is

I. 68% confidence interval → [μ-σ, μ+σ]
II. 95% confidence interval → [μ-2σ, μ+2σ]
III. 99.7% confidence interval → [μ-3σ, μ+3σ]

In case of the gaussian distribution, if the data points increase, then there are chances for the confidence interval limits to change. If the sample size is increased, the confidence interval range reduces. Similarly, if the sample size is decreased, the confidence interval range increases.

According to the Central Limit Theorem, the population mean(μ) lies in the interval [$\bar{x}$-(2σ/sqrt(n)), $\bar{x}$+(2σ/sqrt(n))] with 95% confidence. (where $\bar{x}$ → sample mean, n → sample size).

What is the use of Confidence Interval?

The purpose of confidence intervals is to give us a range of values for our estimated population parameter rather than a single value or a point estimate. The estimated confidence interval gives us a range of values within which we believe with certain probability that the true population value falls. This probability is confidence level.

For instance, if repeated samples were taken and 95% confidence interval for the mean was computed for each sample, then 95% of the intervals would contain the population mean. We expect that 5% of the interval would not contain the true value.

Computing the Confidence Interval using CDF

For example, if we want to compute the range of values within which 95% of the values lie, then we have to first compute the 2.5 percentile (say ‘a’) and 97.5 percentile (say ‘b’) values from the CDF. So we can conclude that the random variable has a 95% confidence interval as [a,b] which is computed using the CDF.

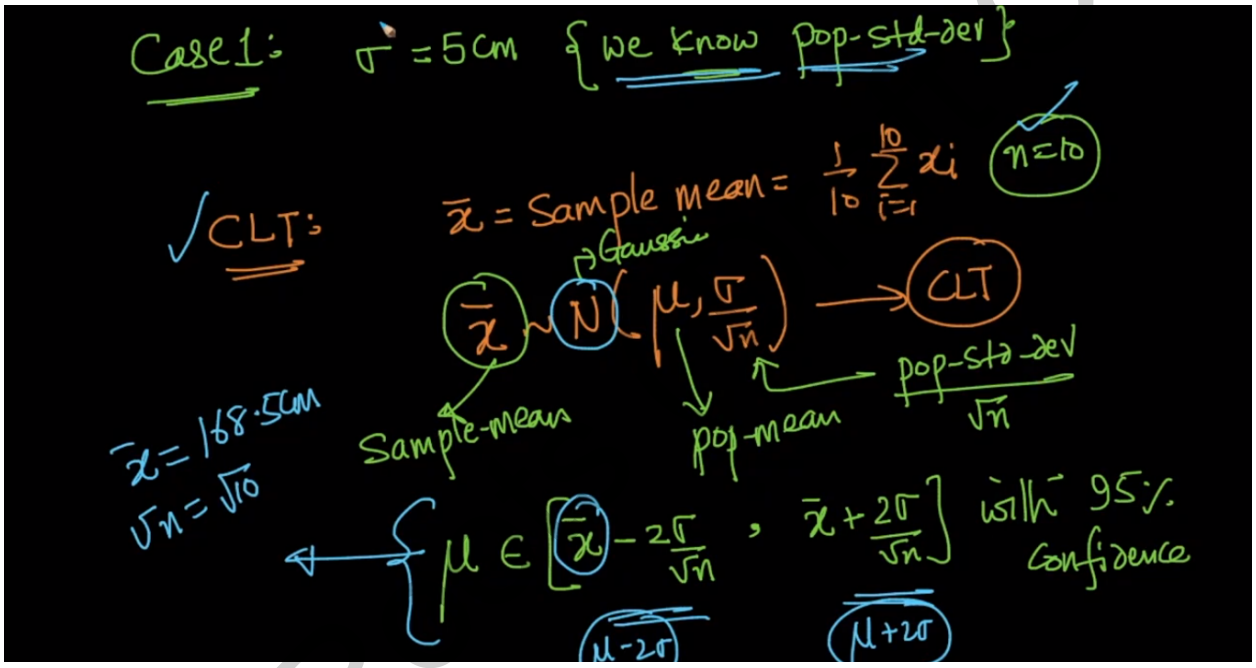
20.22 Confidence Interval for mean of a random variable

Let us assume we have a distribution 'X' (not sure about the type of distribution) with a population mean of 'μ' and a standard deviation of 'σ'. Let us assume we have picked a sample of size 10 (ie., x_1, x_2, x_3, ……, x_{10}).

Let us assume we have to find out the 95% confidence interval estimate for the population mean.

Case 1: (If we already knew the population standard deviation 'σ')

Discussed starting from the timestamp 2:30



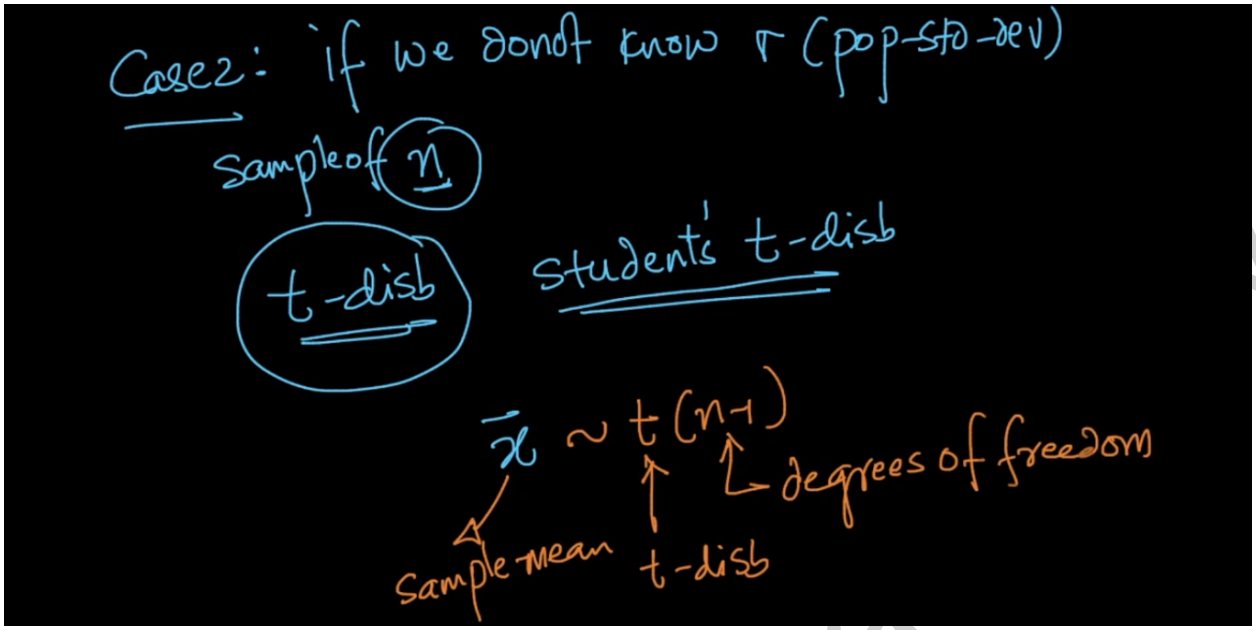
From Central Limit Theorem, we have learnt that if we have any distribution (whether it is gaussian or non gaussian) with a finite and valid values of the population mean and standard deviation, then

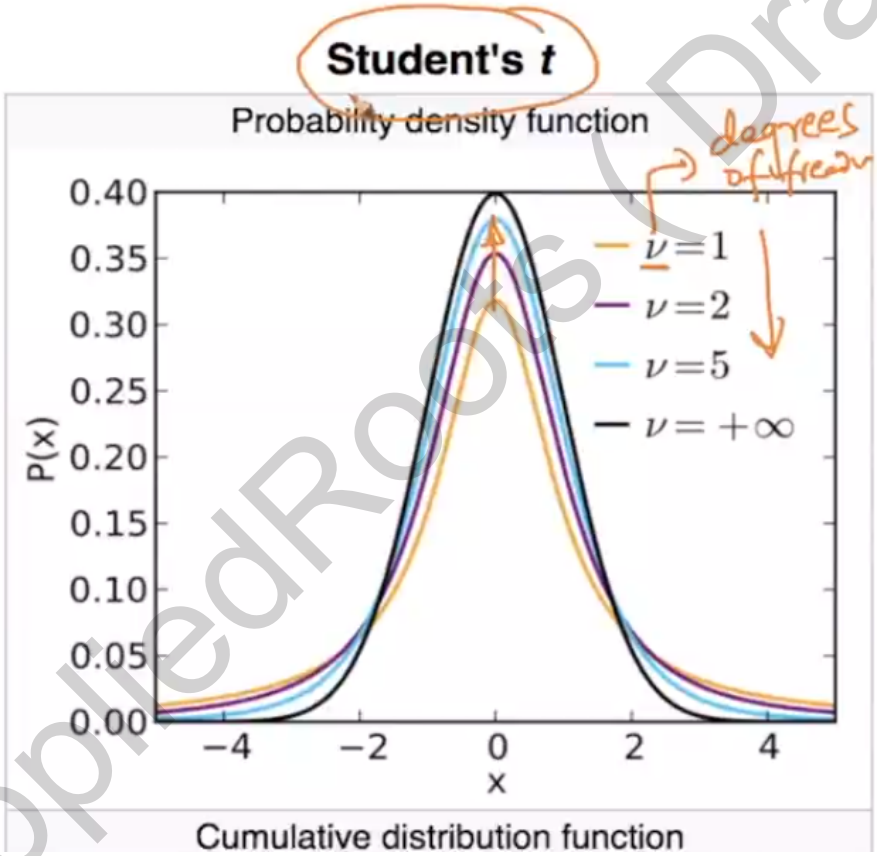
the random variable **$\bar{x}$ ~ N(μ, σ/sqrt(n))** where 'n' → sample size, $\bar{x}$ → mean of all sample means.

The sample mean follows a gaussian distribution with the population mean(μ) as its mean and a standard deviation of 'σ/sqrt(n)'. So now, we can say that the population mean(μ) lies in **μ∈[$\bar{x}$-2σ/sqrt(n), $\bar{x}$+2σ/sqrt(n)]**.

Case 2: (If the population standard deviation 'σ' is not known)

Discussed starting from the timestamp 7:45.





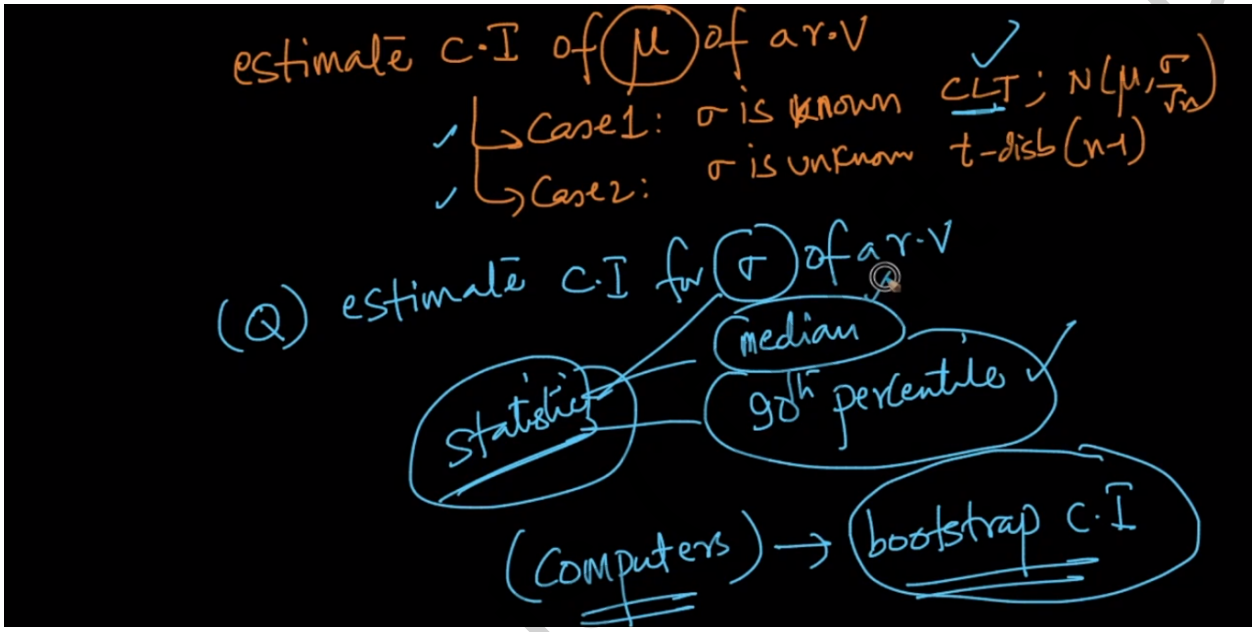
If we do not know the population standard deviation(σ), then we could not apply the Central Limit theorem, as the value of 'σ' is unknown. So we need to go for the t-distribution for estimating the confidence interval of the population mean. This t-distribution is also known as **Student's t-distribution**.

Theoretically the student's t-distribution states that the sample mean follows t-distribution with (n-1) degrees of freedom.

$\bar{x} \sim t(n-1)$

Where 't' denotes the t-distribution and (n-1) denotes the degrees of freedom.

The PDF of student's t-distribution looks similar to the gaussian curve, but not exactly a gaussian curve. The student's t-distribution is exclusively developed for computing the confidence intervals. The major application of student's t-distribution is computing the confidence intervals, when the population standard deviation is unknown.



So in order to estimate the confidence intervals of the population mean,

I. If the population standard deviation is known, we have to apply CLT.
II. If the population standard deviation is unknown, we have to apply the student's t-distribution.

But when it comes to the estimation of other population statistics(like population standard deviation, median, 90th percentile, etc), the above 2 techniques do not work. So we have to go for a technique called **Bootstrapping**.

22.23 Confidence Interval using Bootstrapping

Bootstrapping can be used to compute the confidence intervals not only for the mean, but also for the other statistics like median, variance, standard deviation, 90th percentile, etc.

Let us assume we are given a random variable 'X' whose distribution is unknown. Let us pick a sample of size 'n' and we have to compute the confidence interval of the median of the population of 'X'.

Let 'S' be the sample and its size be 'n'.

$S = \{x_1, x_2, x_3, \ldots, x_n\}$

Let us now pick a sample 'S_1' of size 'm' from 'S', such that m<=n.

$S_1 = \{x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots, x_m^{(1)}\}$ (m<=n)

Here 'S' → Random sample of size 'n' generated from the population.

'S_1' → Random sample of size 'm' generated from the sample 'S'.

The points in 'S_1' are selected randomly one after the other from 'S' while sampling. Here we may also get repetition of the points in 'S_1' while sampling from 'S'. This is called Sampling with Repetition.

While sampling the points from 'S' into 'S_1', the chances of selecting a point is equal for all the points. Hence we generate a uniform random variable U(1,n) and we run a loop for 'm' times, as we have to pick 'm' points into the sample 'S_1'. While running this loop for 'm' times and picking 'm' points randomly, we may get some repeated points. Here 'U' is a discrete uniform random variable.

So now from the sample 'S', we generate 'K' different samples, each of size 'm' using a random distributed discrete random variable using sampling with replacement.

$S_1 = \{x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots, x_m^{(1)}\}$

$S_2 = \{x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, \ldots, x_m^{(2)}\}$

$S_3 = \{x_1^{(3)}, x_2^{(3)}, x_3^{(3)}, \ldots, x_m^{(3)}\}$

.

.

.

$S_k = \{x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_m^{(k)}\}$

These samples are created artificially from the sample 'S' and hence they are called **Bootstrap Samples**. As our task is not to compute the confidence intervals for the population median, we have to compute the medians of each of these samples.

$S_1 = \{x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots, x_m^{(1)}\}$ → 'm_1' is the median

$S_2 = \{x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, \ldots, x_m^{(2)}\}$ → 'm_2' is the median

$S_3 = \{x_1^{(3)}, x_2^{(3)}, x_3^{(3)}, \ldots, x_m^{(3)}\}$ → 'm_3' is the median

.

.

$S_k = \{x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_m^{(k)}\}$ → 'm_k' is the median

These are the medians computed for the bootstrap samples. We now have to sort all these medians and let those sorted medians be

$m_1' <= m_2' <= m_3' <= ..... < m_k'$

From these sorted medians, we have to obtain the confidence intervals.

For example if k=1000, then it means we have 1000 bootstrap samples and 1000 medians associated with them. So then the sorted means would be m_1', m_2', m_3',....,m_{1000}'.

In case, if we want to compute the 95% confidence interval, as the remaining is 5%, we shall divide it into two halves and it would be 2.5% each. (It is because the CLT is applicable for not only mean, but also for other statistics like median, 90th percentile, etc. That is if we pick multiple samples and then compute the median for each of these samples, then all these sample medians follow normal distribution. So 68-95-99.7 rule applies for these sample medians.)

Among the sorted medians, pick the lower bound such that 2.5% of the medians are to its left side and pick the upper bound such that 2.5% of the medians are to its right side. The interval with these lower and upper bounds will be the 95% confidence interval.

(Here out of 1000 medians, 2.5% is 25. So the lower and the upper bounds should be chosen in such a way that 25 medians are present on the left side and 25 on the right side. So the lower and upper bounds are m_{25}' and m_{975}' respectively.).

If we want to compute the confidence intervals for other population statistics such as variance, standard deviation, 90th percentile, etc, then instead of computing the medians, we have to compute the respective statistics for each of the bootstrap samples.

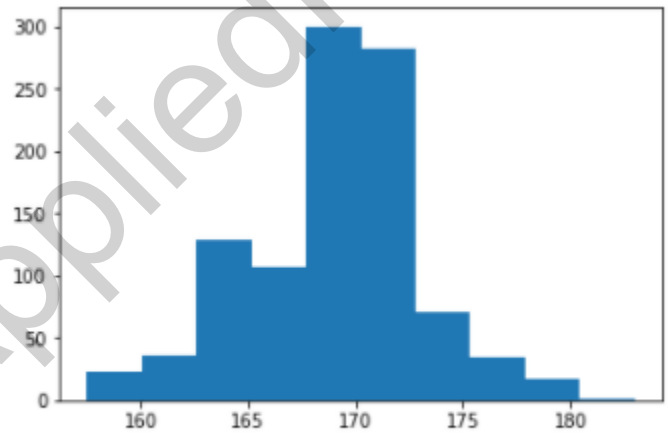
Bootstrapping is a non parametric approach which doesn't make any assumptions about the distribution of the data used. If the bootstrap sample sizes are small, then the confidence intervals are wider. If the bootstrap samples are larger, then the confidence intervals are narrower.

How is Bootstrapping better over the other techniques?

The reason for choosing bootstrapping is that it is a very generic technique that is extremely powerful and makes no assumptions about the underlying distribution. With more and more data as we have nowadays and availability of computational resources, bootstrapping is a super powerful technique which is very popular in data science as compared to t-distribution.

Below is the code snippet that was discussed at the timestamp 13:24 in the video.

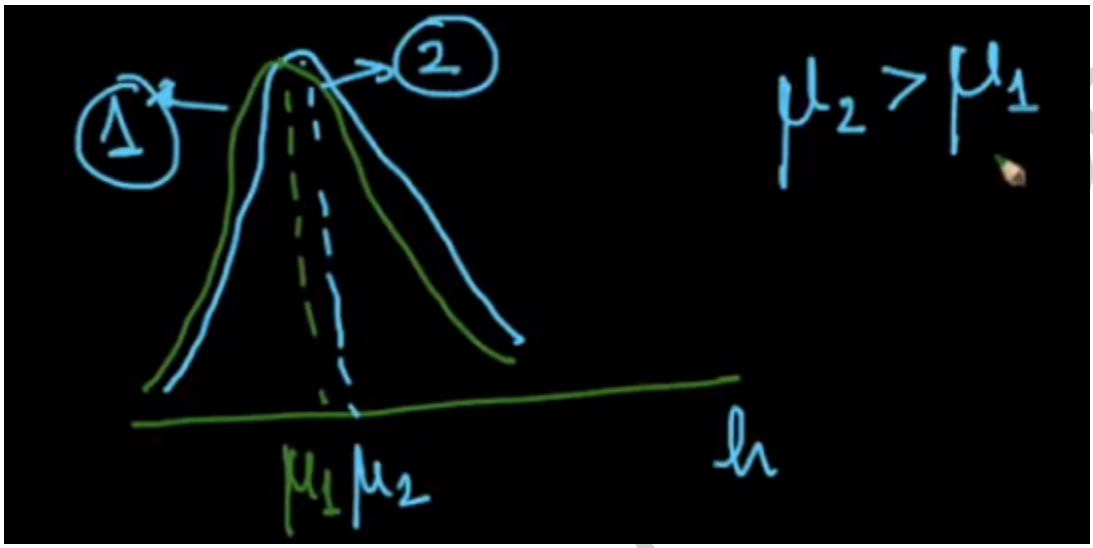
```
import numpy
from pandas import read_csv
from sklearn.utils import resample
from sklearn.metrics import accuracy_score
from matplotlib import pyplot

# load dataset
x = numpy.array([180,162,158,172,168,150,171,183,165,176])

# configure bootstrap
n_iterations = 1000
n_size = int(len(x))

# run bootstrap
medians = list()
for i in range(n_iterations):
    # prepare train and test sets
    s = resample(x, n_samples=n_size);
    m = numpy.median(s);
    #print(m)
    medians.append(m)

# plot scores
pyplot.hist(medians)
pyplot.show()

# confidence intervals
alpha = 0.95
p = ((1.0-alpha)/2.0) * 100
lower =  numpy.percentile(medians, p)

p = (alpha+((1.0-alpha)/2.0)) * 100
upper =  numpy.percentile(medians, p)
print('%.1f confidence interval %.1f and %.1f' % (alpha*100, lower, upper))
```




```
95.0 confidence interval 161.5 and 176.0
```

22.24 Hypothesis Testing Methodology - Null Hypothesis, p-value

Let us assume we have two classes and we have the heights of the students in each class. Let us assume the distributions of the heights of both these classes are as shown below at the timestamp 2:43.



We can say that $\mu_2 > \mu_1$ from the plot, but how confidently can we say that height difference exists between class 1 and class 2 (Because if the scale of 'h' is large, then both the means look as if they are overlapping). We can test whether height difference exists or not using Hypothesis Testing.

Process for solving this problem

Step 1: Choosing a test statistic

Here we have to come up with a number that can say there exists a difference in the heights of two classes. This number is the test statistic and in this context, we choose the difference between the mean of the heights of classes 1 and 2, as the test statistic. (ie., μ_2-μ_1).

μ_2 → Mean of the heights of the sample from class '2'.

μ_1 → Mean of the heights of the sample from class '1'.

If the value of μ_2-μ_1=0, then we can say that there exists no difference in the heights of the students in the classes.

Note: If the populations of class 1 and class 2 are different, then we should pick samples of the same size from both the populations.

Step 2: Null Hypothesis (H_0)

This whole method follows proof by contradiction. Null Hypothesis is an assumption we are making for solving a problem and working on the problem towards it.

For this given problem, the null hypothesis is

H_0: There is no difference between 'μ_1' and 'μ_2'.

There is another concept called Alternative Hypothesis which is an inverse of Null Hypothesis. The Alternative Hypothesis is denoted as 'H_1'.

For this given problem, the alternative hypothesis is

H_1: There exists a difference between 'μ_1' and 'μ_2'.

As this is a proof by contradiction, if we believe that null hypothesis is True, we have to prove that alternative hypothesis is False and should accept Null Hypothesis. (or) In case, if we are unable to prove that the null hypothesis is true, we have to accept the alternative hypothesis.

Step 3: Computing 'p' value (Computation part which is done using Resampling and Permutation test)

'p' value is the probability of observing a value for (μ_2-μ_1) if our null hypothesis is true. Let us assume the null hypothesis (H_0) is true.

1) Let us assume a value of 10cm for μ_2-μ_1(ie., μ_2-μ_1 = 10)
   If p-value = 0.9, it means the probability of (μ_2-μ_1 = 10) is 0.9, if our null hypothesis is true. It means, if there is not much difference in the heights of the means, we still see a difference of 10cm with 90% probability.

   Here if the sample size is small, the difference between the mean heights may be large,because the samples might be biased sometimes. But if the samples are large in size, then the samples become more appropriate representatives of the population and both the sample means will be very much near to each other, than the sample means of smaller samples. This could reduce the differences in the sample means, when the sample sizes are large.

   If the sample sizes are small, then we see huge variations in the sample means and whereas if the sample sizes are large, then we do not see much variation among the sample means(as both of them will be nearer to each other). So the difference between the samples is low.

   So if the samples are larger the difference between the sample means will become almost nearer (or) equal to zero and in such cases, if p=0.9, then we can easily confirm that the null hypothesis is true. We then accept the null hypothesis.

2) Let us assume p-value = 0.05 which means there is 5% chance that (μ_2-μ_1 = 10), if the null hypothesis is true. So here if 'p' value is very low, then we say that the chances for the null hypothesis to be true are very low. In such cases, we have to reject the null hypothesis, and accept the alternative hypothesis.

22.25 Hypothesis Testing Intuition with coin toss example

Given a coin, we have to determine if the coin is biased towards the head or not. It means we have to determine if P(H)>0.5

If a coin is not biased, then P(H)=0.5. Let us perform an experiment of tossing a coin (say 5 times). Let the count of the number of heads be 'X'. (This is the test statistic)

So let us assume the occurrences of tossing a coin for 5 times be {H,H,H,H,H}.

So now X = Number of Heads = 5 (This is an observation from the experiment)

We shall now work on computing P(X=5|coin is not biased towards the head) (ie., $P(X=5|H_0)$.

Where $H_0 \rightarrow$ Null Hypothesis which says the given coin is not biased towards the head.

As it is given that the coin is not biased towards the head, P(H) = ½

So $P(X=5|H_0) = P(H)*P(H)*P(H)*P(H)*P(H) = (½)^5 = 0.03 = 3\%$

It means there are 3% chances of getting 5 heads, if the coin is not biased towards the head.

So here p-value = P(observation by experiment|assumption)

Here $P(X=5|H_0) = 0.03$. It means the probability of the observation to be true, when the given assumption is true, is 0.03(ie. 3%)

(In general, 5% is the rule of thumb. If the percentage of p-value is <5%, then we reject the null hypothesis). Here in this case, the null hypothesis is "Coin is not biased towards the head". So we reject it. The alternative hypothesis(H_1) in this case is, "The coin is biased towards the head" and we accept it.

For example, if we reduce the number of tosses to 3, then

$P(X=3|H_0) = (½)^3 = 12.5\%$

Here as 12.5% > 5%, we accept the null hypothesis. Here changing the number of tosses has changed the result completely.

Note: If the 'p' value is less than 0.05(say), then we can't reject the observation as it has already been observed. In such cases, we can only reject the null hypothesis.

While performing Hypothesis Testing, we should carefully

1) Choose the sample sizes (in this experiment, the sample size is the number of coin tosses).
2) Define the null hypothesis, in such a way that P(observation|assumption) is easy and feasible to compute.
3) Choose the test statistic.

22.26 Resampling and Permutation Test

Let us consider the example of the heights of students in classes 1 and 2(which was the first example discussed in Hypothesis Testing). Let us assume, we have picked samples of size 50(say) from each of these classes and have computed the means for both the samples, and let them be denoted as 'μ_1' and 'μ_2' respectively. Initially let the difference between these two sample means be denoted by 'Δ'.

$\Delta = \mu_2 - \mu_1$

Now let us combine all the elements present in both these samples into a single set, and then pick two random samples(these samples should be of the same size 50 throughout). Let us compute the means for the new randomly picked samples and compute the difference between them. Let this difference be denoted as 'δ_1'.

We shall repeat this process for a certain number of times(say 10000) and obtain all the values of the sample mean differences. Let them be denoted as 'δ_1', 'δ_2', 'δ_3',..... 'δ_{10k}'.

We should now sort all these sample mean differences and let these sample means in the sorted order be denoted as 'δ_1|', 'δ_2|', 'δ_3|',..... 'δ_{10k}|'. Now from the increasing order of these sample means, we have to check what percentage of the values are greater than 'Δ'. (ie., $\Delta = \mu_2 - \mu_1$)

Let's say 5% of the values are greater than 'Δ', then

P - value = 5/100 = 0.05

If x% of the values are greater than 'Δ', then

P - value = x/100

In the olden days, the statisticians assumed that the sample mean differences follow gaussian distribution.

$\delta_i \sim N(0,1)$

So if Δ = 2, then $P(\delta_i >= 2) = 0.025$ (It is because 2.5% of the points lie to the right side of 2σ on the 'X' axis for a normal distribution). So p-value = 0.025.

Note: The notes for the video lecture 22.27 is ignored as the same examples were discussed in this course.

22.28 K-S Test for Similarity of Two Distributions

Let us assume we have two random variables 'X_1' and 'X_2' and we have to pick samples from them of sizes 'n1' and 'n2' respectively.

Sample of $X_1 = \{x_1^{(1)}, x_2^{(1)}, x_3^{(1)},.....,x_{n1}^{(1)}\}$

Sample of $X_2 = \{x_1^{(2)}, x_2^{(2)}, x_3^{(2)},.....,x_{n2}^{(2)}\}$

We have to check if the distributions of 'X_1' and 'X_2' are the same or not. For any distribution, let us say the population mean is 'μ' and the population standard deviation is 'σ', then if we standardize the distribution and if the standardized form is N(0,1), then we can say that the given population follows gaussian distribution.

Objective: To determine if both 'X_1' and 'X_2' have the same distribution.

'X_1' has 'n_1' observations and 'X_2' has 'n_2' observations. Let us now plot the CDF of 'X_1' and 'X_2'.

For this problem, the null hypothesis and the alternative hypothesis would be

H0: 'X_1' and 'X_2' have the same distribution.

H1: 'X_1' and 'X_2' don't have the same distribution.

If we pick any point on the 'X' axis, we find the differences in the CDF values. In case, if the given two random variables follow the same distribution, then the CDFs of both the plots will overlap. When the CDFs overlap completely, then at any point on the 'X' axis, the difference in the CDF is zero.

If both 'X_1' and 'X_2' follow the same distribution, then

a) If both the samples are of smaller size(ie., 'n_1' and 'n_2' are small), then we see the differences in the CDFs, even if both the random variables follow the same distribution.
b) If both the sample sizes are large(ie., 'n_1' and 'n_2' are large), then if both the random variables follow the same distribution, the more the CDFs would overlap.

So the test statistic in this context is

$D_{n1,n2} = Sup_x|F_{1,n1}(x) - F_{2,n2}(x)|$ which is pronounced as "Supremum over all x's of "$F_{1,n1}(x) - F_{2,n2}(x)$". Here supremum means the maximum value.

$F_{1,n1}(x) \rightarrow$ CDF of random variable 'X_1' over 'n_1' points.

$F_{2,n2}(x) \rightarrow$ CDF of random variable 'X_2' over 'n_2' points.

If both 'n_1' and 'n_2' are large and 'X_1' and 'X_2' come from the same distribution, then the difference between the CDFs at any point is zero, and finally the value of $D_{n1,n2} = 0$. (We can accept the null hypothesis).

After a thorough research, it is found that we can reject the null hypothesis at a level 'α' if $\mathbf{D_{n1,n2} > c(\alpha)*sqrt((n_1+n_2)/(n_1*n_2))}$

Where $n_1, n_2 \rightarrow$ sample sizes of 'X_1' and 'X_2' respectively.

$\alpha \rightarrow$ 'p' value

22.29 p-value for K-S Test

So far in hypothesis testing, we have learnt to obtain the 'p' value first and then decide whether to accept or reject the null hypothesis. In Hypothesis testing, we reject the null hypothesis at a significant level 'α', if the p-value < α.

But here in KS test, we reject the null hypothesis at a significance level 'α', if

$$\mathbf{D_{n1,n2} > c(\alpha)*sqrt((n_1+n_2)/(n_1*n_2))}$$

The value of $\mathbf{c(\alpha) = sqrt(-½)*ln(\alpha/2))}$

So for the time being we shall denote $\mathbf{D_{n1,n2}}$ as 'D'. So now the above condition becomes

$$\mathbf{D > c(\alpha)*sqrt((n_1+n_2)/(n_1*n_2))}$$

We are now substituting the value of 'c(α)' here and then it becomes

$$\mathbf{D > sqrt(-½)*ln(\alpha/2)) * sqrt((n_1+n_2)/(n_1*n_2))}$$

$$\mathbf{D * sqrt((n_1*n_2)/(n_1+n_2)) > sqrt((-½)*ln(\alpha/2))}$$

Applying square on both the sides, then it becomes

$$\mathbf{D^2 * ((n_1*n_2)/(n_1+n_2)) > ((-½)*ln(\alpha/2))}$$

Multiplying with a minus on both sides.

$$\mathbf{-D^2 * ((n_1*n_2)/(n_1+n_2)) < ((½)*ln(\alpha/2))}$$

Multiplying with '2' on both sides.

$$\mathbf{-2*D^2 * ((n_1*n_2)/(n_1+n_2)) < (ln(\alpha/2))}$$

In order to remove the logarithm on the right hand side, we have to apply exponential function on both sides/

$$\mathbf{exp(-2*D^2 * ((n_1*n_2)/(n_1+n_2))) < (\alpha/2)}$$

Multiplying by 2 on both sides, so that the denominator in the right hand side gets cancelled.

$$\mathbf{2*exp(-2*D^2 * ((n_1*n_2)/(n_1+n_2))) < \alpha}$$

When we compare this to the hypothesis testing (where null hypothesis gets rejected if p<α), if we apply the same logic here, then we can say that

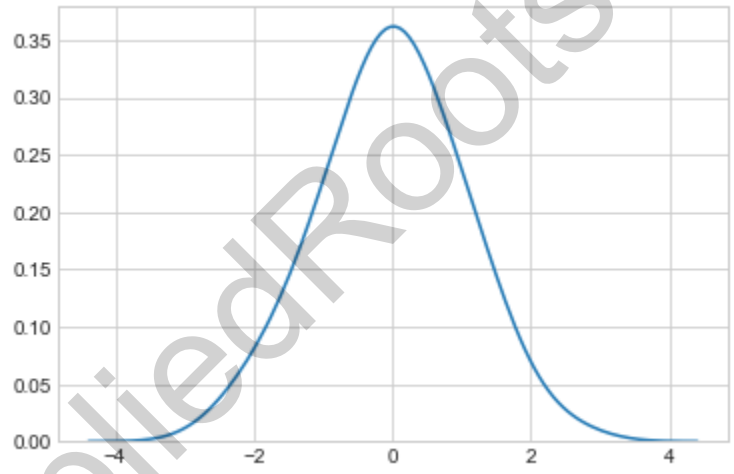
$$\mathbf{P\text{-}value = 2*exp(-2*D^2 * ((n_1*n_2)/(n_1+n_2)))}$$

22.30 Code Snippet - KS Test

So far we have learnt the theoretical concept of KS test and now we shall look at the code part.

```
import numpy as np
import seaborn as sns
from scipy import stats
import matplotlib.pyplot as plt

#generate a gaussian r.v X
x = stats.norm.rvs(size=1000);
sns.set_style('whitegrid')
sns.kdeplot(np.array(x), bw=0.5)
plt.show()
```

In the above code snippet, we are first importing the libraries and then creating a random variable ‘x’ that follows normal distribution. After that, we are plotting its PDF. The plot looks as below



Let us now perform the KS test using the below code snippet.

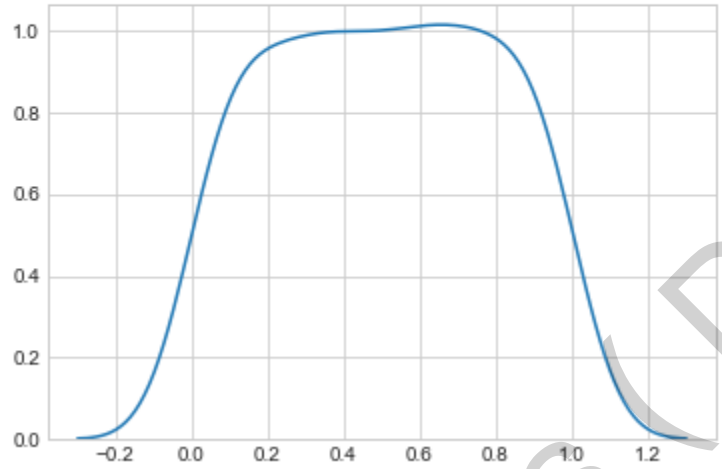
```
stats.kstest(x, 'norm')
```

```
KstestResult(statistic=0.021308397286061931, pvalue=0.75424031453335627)
```

After the execution of the above statement, in the result, we get two parameters. One is the 'p' value and the other is the 'statistic' parameter which indicates the '$D_{n1,n2}$' value. Here we accept the null hypothesis.

Now we shall perform the same KS test for a uniform distributed random variable.

```
# Y ~ Continous Uniform Distribution(0,1)
y = np.random.uniform(0,1,10000);
sns.kdeplot(np.array(y), bw=0.1)
plt.show()
```



Here Using the above code snippet, we are creating a continuous uniform random variable and are plotting its PDF. Let us now perform the KS test.

```
stats.kstest(y, 'norm')
```

```
KstestResult(statistic=0.50159644397739489, pvalue=0.0)
```

Here we get the 'p' value and the statistic as the result. 'P' value of 0.0 indicates that the two distributions are not the same and we can reject the null hypothesis.

22.31 Hypothesis Testing - Another Example

Let us assume we have two cities 'C_1' and 'C_2'. Our task here is to determine if the population means of heights of people living in these two cities is same or not.

Let us now pick two random samples(one from each city) of size(say 50) and let them be denoted as 'C_1' and 'C_2'.

$C_1 = [h_1, h_2, h_3, ..., h_{50}]$

$C_2 = [h_1', h_2', h_3', ..., h_{50}']$

Let 'μ_1' and 'μ_2' be the sample means heights of the people living in the cities 'C_1' and 'C_2' respectively. For example, we have observed the values as μ_1 = 162cm and μ_2 = 167cm. So the difference between these sample means = μ_2-μ_1= 167-162 = 5cm.

So the test statistic here would be X = μ_2-μ_1

Null Hypothesis(H_0): There is no difference between the population means.

Now we have to compute the value of P(X=5cm|H_0) which is the probability of observing a difference of 5cm in the sample means of sample size 50, if the null hypothesis is true.

Case 1:

If P(X=5cm|H_0) = 0.2 (say), it means there is a 20% chance of observing a difference of 5cm in the sample mean heights of 'C_1' and 'C_2' with a sample of 50, if there is no population mean difference. As this probability is true, we can confirm that our assumption is true and we accept the null hypothesis.

Case 2:

If P(X=5cm|H_0) = 0.03 (say), it means there is a 3% chance of observing a difference of 5cm in the sample mean heights of 'C_1' and 'C_2' with a sample of 50, if there is no population mean difference. This implies our assumption is wrong and we reject the null hypothesis and accept the alternative hypothesis.

22.32 How to use Hypothesis Testing?

Application 1

KS Test is one of the best applications where hypothesis testing is used to check if two random variables follow the same distribution or not.

Application 2 (Designing medicine/drugs)

Let us assume we have two drudge 'D_1' and 'D_2' for curing a disease. 'D_1' is manufactured by company 'C_1' and is in use whereas 'D_2' is recently manufactured by a company 'C_2'.

Let's say 'D_1' cures the disease in 4hours and the company 'C_2' claims that 'D_2' can cure the same disease faster than 'D_1'.

Claim by 'C_2': 'D_2' cures the disease faster than 'D_1'.

Task/Experiment: To determine if the claim is true or not.

In order to conduct this experiment, let's pick a sample of 100 patients and divide them into 2 groups. They are set 'S_1' and set 'S_2' with a size of 50 patients each.

Let the patients in 'S_1' be given the drug 'D_1' and the patients in 'S_2' be given the drug 'D_2'. Let the values in the sets be filled with the duration taken by the patients to recover.

'μ_1' → mean time of 'S_1' (say 4 hours)

'μ_2' → mean time of 'S_2' (say 2 hours)

Here we got these mean values for samples of smaller sizes. If the sample sizes are large, then these values will change. So instead of going with these sample means, we can go for hypothesis testing.

Null Hypothesis (H_0): 'D_1' and 'D_2' are not different. (ie., they both take the same time)

Test statistic(X) = (μ_2-μ_1)

μ_2-μ_1 = 2 hours

if $P(X>=2|H_0) = 0.01$ (say it is small), then we reject the null hypothesis and accept alternative hypothesis.

Significance Level (α):

We have to be really sure when we reject the null hypothesis. If the case is really critical like in healthcare, we have to choose α = 1%.

If it is something like an e-commerce domain problem, then we have to choose α = 3%. But typically 5% of 'α' is used in most of the cases.

Note

One best example of hypothesis testing in e-commerce is usage of credit cards for purchasing.

Sample 1 (S_1): Customers who have a Visa Credit Card.
Sample 2 (S_2): Customers who have a Mastercard Credit Card.
Let 'μ_1' be the mean of the sample 'S_1' and 'μ_2' be the mean of the sample 'S_2'.

Visa claims that it's customers make more purchases than Mastercard.

Null Hypothesis (H0): (μ_2-μ_1) is large.
Test Statistic: (μ_2-μ_1)

22.33 Proportional Sampling

It is one of the important topics in machine learning. Let us assume we are given an array of values. Let the array be d = [2.0, 6.0, 1.2, 5.8, 20.0] and let these numerical values be denoted as d_1, d_2, d_3, d_4 and d_5 respectively.

The task of proportional sampling is to pick an element among the given 'n' elements such that the probability of picking an element is proportional to the d_i's.

If we randomly pick a value, then all these values are equally probable. When we look at the values in the array, the value of 'd_5' is the highest and the value of 'd_3' is the least. So in case of proportional sampling, the probability of picking 'd_5' should be the highest and the probability of picking 'd_3' should be the least.

Procedure for Proportional Sampling

Step 1

a) Compute the sum of all the values of the array.
   $S = \Sigma_{i=1}^{n} d_i$ (For the given values in the array, S = 35)

b) Normalize all the values by dividing them by the sum.
   $d_i^{l} = d_i/S$
   $d_1^{l} = d_1/S$, $d_2^{l} = d_2/S$, $d_3^{l} = d_3/S$, $d_4^{l} = d_4/S$, $d_5^{l} = d_5/S$
   So after normalization, the values are
   $d_1^{l} = 0.0571$, $d_2^{l} = 0.171428$, $d_3^{l} = 0.034$, $d_4^{l} = 0.165$, $d_5^{l} = 0.571$
   All the above normalized values lie in [0,1] and also all these values sum to 1.
   (ie., $\Sigma_{i=1}^{n} d_i^{l} = 1$)

c) Compute Cumulative Normalized Sum
   d_1_tilda = $d_1^{l} = 0.0571$
   d_2_tilda = d_1_tilda + $d_2^{l} = 0.228528$
   d_3_tilda = d_2_tilda + $d_3^{l} = 0.262828$
   d_4_tilda = d_3_tilda + $d_4^{l} = 0.428528$
   d_5_tilda = d_4_tilda + $d_5^{l} = 1.00$

Step 2

We have to sample one value from a uniform random variable U(0,1). Let this value be denoted as 'r'.

r = numpy.random.uniform(0.0, 1.0, 1)

Step 3

Proportional Sampling happens now.

If **r<=d_1_tilda**, then **return 1**
If **r<=d_2_tilda**, then **return 2**
If **r<=d_3_tilda**, then **return 3**

If **r<=d_4_tilda**, then **return 4**
If **r<=d_5_tilda**, then **return 5**

Now
d_1_tilda < $d_2^{|}$ < d_2_tilda (because **d_2_tilda = d_1_tilda + $d_2^{|}$**)
d_2_tilda < $d_3^{|}$ < d_3_tilda (because **d_3_tilda = d_2_tilda + $d_2^{|}$**)
d_3_tilda < $d_4^{|}$ < d_4_tilda (because **d_4_tilda = d_3_tilda + $d_3^{|}$**)
d_4_tilda < $d_5^{|}$ < d_5_tilda (because **d_5_tilda = d_3_tilda + $d_4^{|}$**)

The probability of picking '4' = probability of 'r' lying between 'd_3_tilda' and 'd_4_tilda' = $d_4^{|}$
But **$d_4^{|} \propto d_4$** (because $d_4^{|} = d_4/5$). So
The probability of picking '4' in U(0,1) = d_4
The probability of picking '3' in U(0,1) = d_3
The probability of picking '2' in U(0,1) = d_2
The probability of picking '1' in U(0,1) = d_1
The probability of picking '5' in U(0,1) = d_0

Higher weightage is given to the larger values in proportion sampling.