Statistics for machine learning notes

:INDEX:

You have the Real world data. You don't know which distribution it is. It is even not fitting into most of the famous distributions. There are two features and you don't know the relation between them. Your data is discrete and you want to find PDF for it.
Population is too big and you want to find mean standard deviation using samples.
you are not sure about the mean you have estimated and find confidence on your answer.

To answer all of the above answers we need to learn some statistics,

It is broadly divided into four parts as per our curriculum-

1. Observing and transforming given RV or data.

2. Estimating different paramters and confidence intervals.

3. Finding relationship between two features (RV).

1. Observing and transforming given RV or data.

1. Is your data symmetric like gaussian distribution? To find this we measure two parameters - Skewness and Kurtosis.

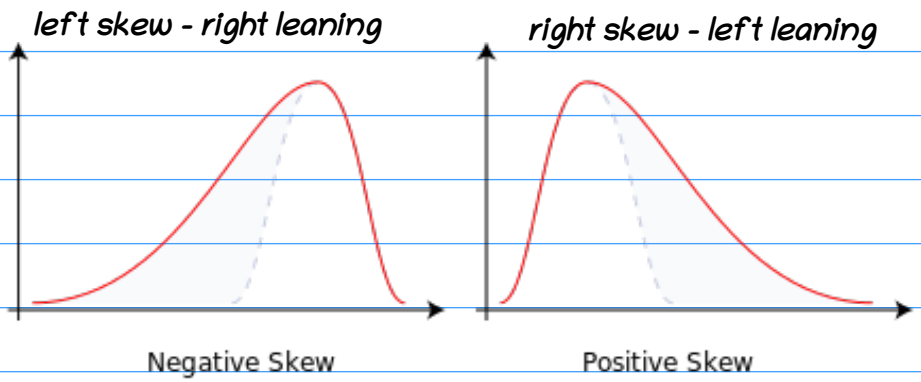
skewness: it is measure of asymmetry of probability distribution of RV about its mean.

sample skewness formula

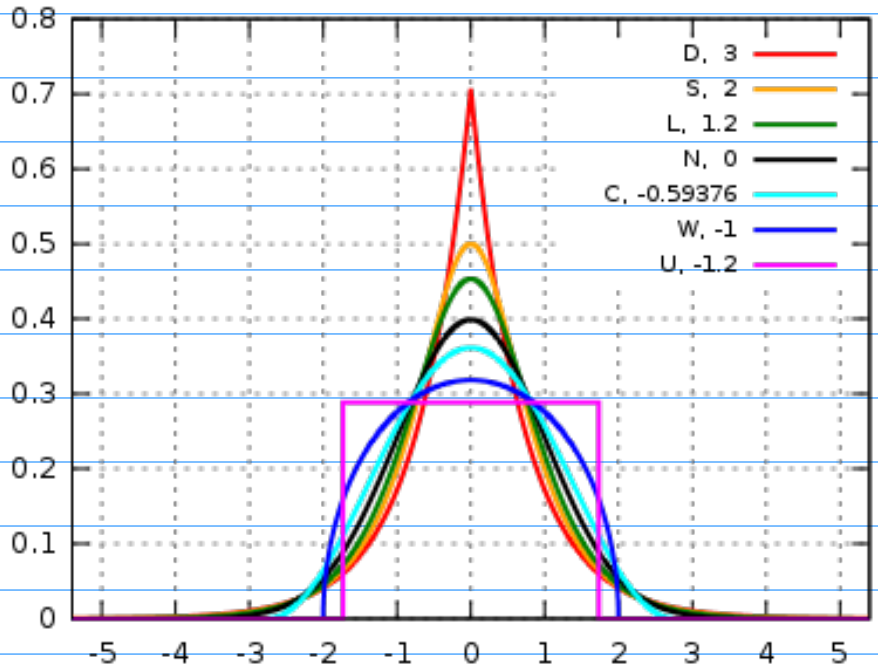
$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}$$

m3 is sample third central moment

s is sample standard devitaion



kurtosis : it is measure of tailedness of probability distribution of RV.



DISTRIBUTIONS

D : laplace- double exponential

S : hyperbolic secant

L : Logistic distribution

N : Normal distribution

C : Raised cosine distribution

W : Wigner semicircle distribution

U : Uniform distribution

Excess kurtosis = kurtosis - 3 (Guassian RV kurtosis)

2. Is your data following normal/gaussian distribution? To find this we need QQ plot.

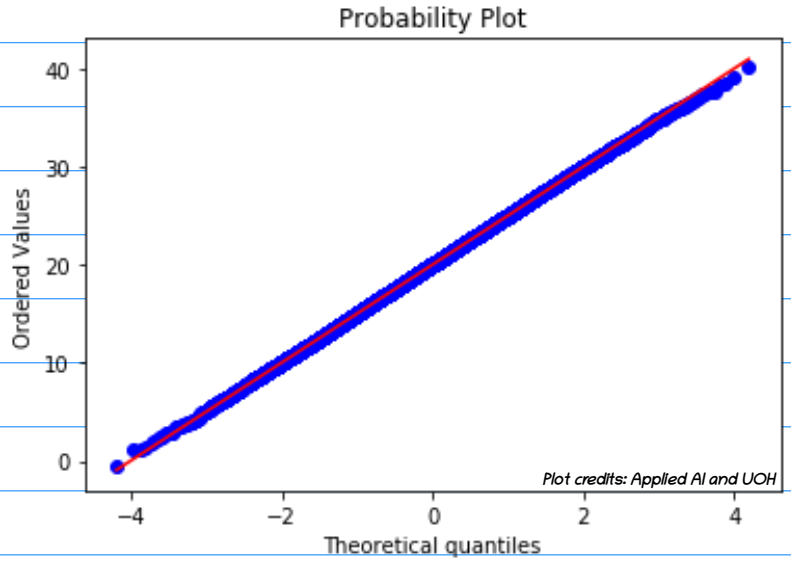
QQ plot: We need to plot percentiles of RV vs normally distributed data (generated).

Step 1: Sort data and calculate percentiles of given RV X.

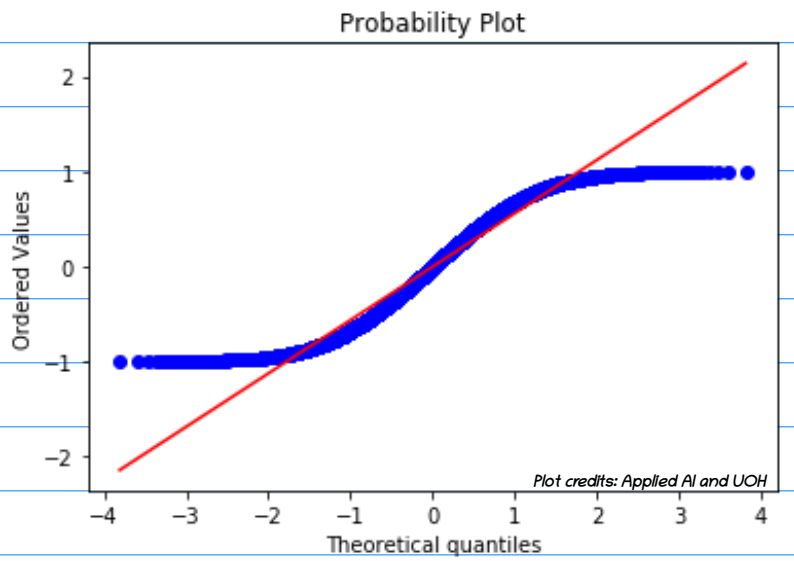
Step 2: Take normally distributed data Y ~ N(0,1) and calculate percentiles.

Step 3: Plot percentiles of X vs percentiles of Y which is called QQ plot.

QQ plot of Normally distributed data



QQ plot of uniformly distributed data



After confirmation that RV is normally distributed we need to perform standardization,

Standardization of data

We might have different data with different units and there could be so much difference in their range. For example height in meter (around 0.8-2 m) and weight in lb (100-300 lb). While developing ML model one parameter will have more weightage over another. To overcome this we perform standardization process.

Stanadard normal variate (z): z ~ N(0,1) mean=0 and variance=1

let RV X ~ N(μ,σ^2) mean = μ variance= σ^2 and standard deviation = σ

for all values in x we calculate,

$$z = x_i' = \frac{x_i - \mu}{\sigma}$$

Now new X' ~ N(0,1)

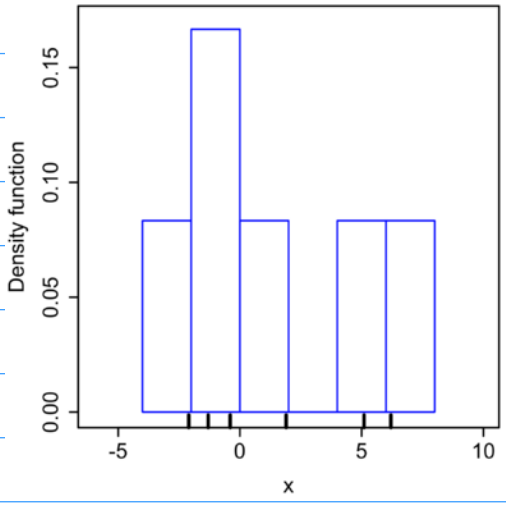
3. You have discrete data. how can you find PDF out of PMF?

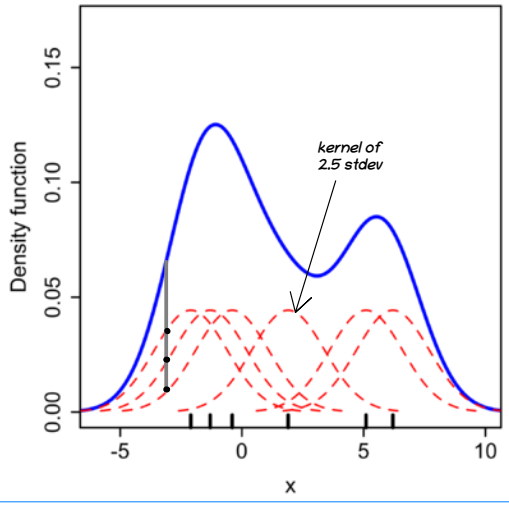
KDE plot (Kernel Density Estimation:

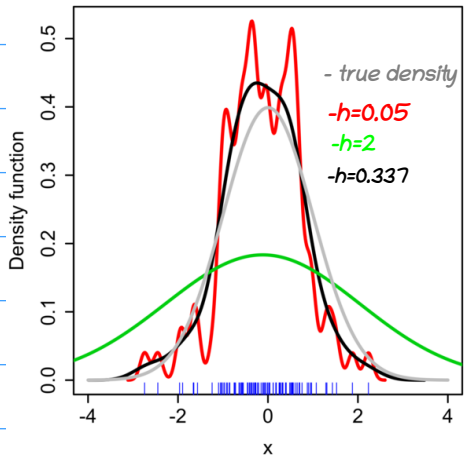
Step 1: Calculate histogram and plot PDF values.

Step 2: For each data point draw a kernel of bandwidth h (or some Standard-deviation)

Step 3: Calculate sum of all kernels at each point, which gives smooth PDF







As we can observe against true density N(0,1) grey line we have tried to plot different KDE estimation with h=2 green line, h=0.05 red line and h=0.337 black line - which is the closest estimation of true density.

4. You noticed that your data is not normally or log normally distributed. It is following Pareto distribution(power law ditribution). How can you normalize it?

Box-Cox transformation:

RV $X = \{x_1, x_2, x_3, \ldots, x_n\}$ is following power law(pareto) distribution. We want to convert it into normal/gaussian distribution RV $Y = \{y_1, y_2, y_3, \ldots, y_n\}$

Step 1: Apply box-cox transformation to find the value of λ

λ = box-cox(X)

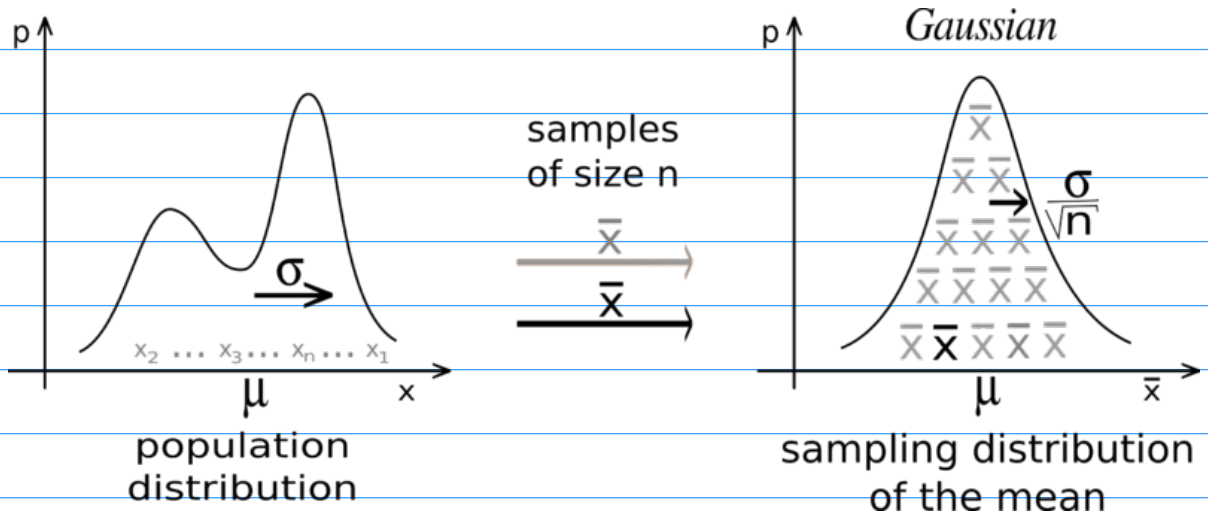
Step 2: Now use this λ to find values of Y for corresponding X,

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

2. Estimating different paramters and confidence intervals.

1. Population is too big that you can not get data of whole population. You can take so many samples. How can you calculate mean and standard deviation now?

Central Limit Theorem (CLT):



let X be the population whose mean is **μ** and standard deviation is **σ**
from which you draw m(1000, more the better) samples of size n(≥30).

Despite of any distribution X follows, sample means will follow gaussian distribution.

Step 1: Out of X take m samples. S1, S2, S3, ..., Sm.

Step 2: Calculate mean of each sample. $\bar{x}1$, $\bar{x}2$, $\bar{x}3$ ...$\bar{x}m$

Step 3: Plot $\bar{x}1$, $\bar{x}2$, $\bar{x}3$ ...$\bar{x}m$ which always follows gaussian distribution $N(\mu, \sigma^2/n)$.
standard deviation = $\sigma / \sqrt{n}$

Now you can calculate population mean and standard deviation,

Population mean = Sampling means' mean = **μ**

Population standard deviation = √n * sampling means' standard deviation

RULE OF THUMB: CLT works for sample size n≥ 30, and more the number of samples better the estimation will be.

To watch CLT in action - https://www.youtube.com/watch?v=8Z9XRrJU9ZM

2. You have sample and you want estimated some parameters but you are not sure 100% on the answer and you want to specify the confidence interval on your answer.

You might know or don't know the underlying distribution of RV. Let's see how can we find the confidence interval in both cases.

Normal distribution:

If RV is following Gaussian distribution we can easily find confidence interval using normal ditribution table.

μ lies in (μ-2σ , μ+2σ) range with 95% probability.

To find confidence interval of 90% we need to check corresponding value in Normal ldstribution table or calculate using formula.

CI for mean of Random Variable:

Case 1: Population standard deviation is known.

We can use CLT to estimate the mean and standard deviation by drawing large number of samples with sample size≥30.

Population mean (μ) ≈ Sampling means' mean ($\bar{x}$) = $\frac{1}{m}\sum_{i=1}^{m} xi$

We already know the population standard deviation = σ

Now, $\mu \in \{ \bar{x} - 2\sigma', \bar{x} + 2\sigma'\}$ with 95% Confidence

Sample St.dev = $\sigma' = \sigma / \sqrt{n}$

$\mu \in \{ \bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n}\}$ with 95% Confidence

Case 2: Population standard deviation is unknown. (NOTE: RV is gaussian distributed)

In this case we can use Student's t distribution to estimate the population standard deviation and then estimate the CI for population mean.

WHY? when we have really small sample size we can't use CLT; In that case we need to use student's t-distribution to estimate the mean of RV which is gaussian disb.

3. You know the mean and standard deviation of RV but you don't know the distribution. Now how can you answer about x% data lies within yσ range?

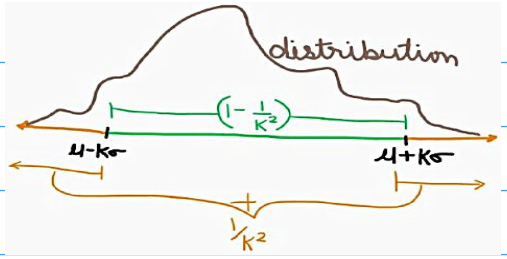
For given data we have calculated the mean μ and standard deviation σ but we are not sure about the underlying distribution

Chebyshev's inequality:

$P(|X - \mu| \geq k\sigma) \leq 1 / k^2$

$P(X \leq \mu - k\sigma \text{ and } X \geq \mu + k\sigma) \leq 1 / k^2$

$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - 1 / k^2$



3. As we have seen methods to estimate mean and standard deviation, what if we want to estimate other parameters like median, 90th percentile with confidence interval?

BootStrapping:

From given random variable X (size n) we draw k samples of size m (m <= n). Now for all samples we calculate the parameter value we are interested (median, variance ...). After that we need to sort these values and use percentiles to find the C.I.

example:

X = {x1, x2, x3, ..., xn} out of which we draw k=1000 samples S1, S2, ..., Sk

Si = {si1, si2, si2,..., sim}

for each sample we calculate median (or any othher parameter) Mi and sort them.

sorted median values: M1', M2', M3', ..., M1000'

Confidence interval of 95% = (2.5th Percentile , 97.5th Percentile)

= (M'(25%), M'(97.5%))

= (M25' , M975')

We can apply bootstrapping technique to estimate anything using samples.

3. Finding relationship between two features (RV).

1. You have two features in the data (two RV) and you want to see if the are following same kind of distribution. how to do it?

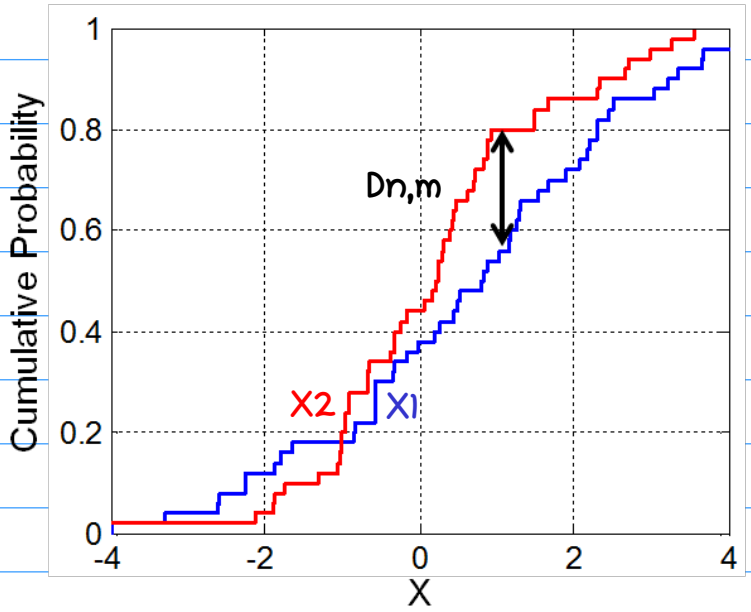
KS test (Kolmogorov-Smirnov test): Two sample KS test

We use hypothesis testing to prove if two RV are following the same distribution.

We use CDF to perform KS test.

KS statistic is the maximum (supremum) difference between two RV in CDF.

Example:



Sample X1 is of size n and X2 is of size m.

and F1,n(x) is cummulative probability at x for X1 sample

F2,m(x) is cummulative probability at x for X2 sample

$$\text{K-S statistic} = D_{n,m} = \sup_x |F_{1,n}(x), F_{2,m}(x)|$$

The null hypothesis (X1 and X2 are of same distribution) is rejected at level α if,

$$D_{n,m} > C(\alpha) \cdot \sqrt{\frac{n+m}{n \cdot m}}$$

Most common values,

α	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.073	1.138	1.224	1.358	1.48	1.628	1.731	1.949

and in general,

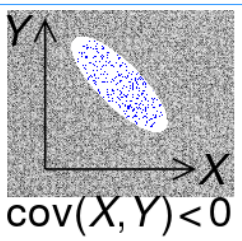
$$C(\alpha) = \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1}{2}} \quad \text{and} \quad D_{n,m} > \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1+\frac{m}{n}}{2m}}$$

2. We have two features in the data (two RV) and we want to see if there is any realation between those two variables. how can we do that?

Covariance:

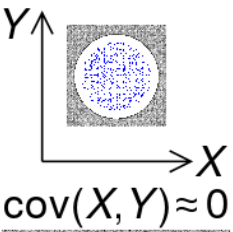
Covariance between to random variable X and Y (Height and weight of student),

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)\cdot(y_i - \mu_y)$$

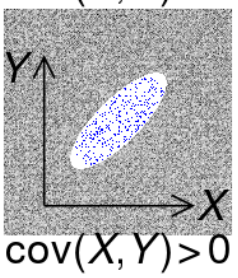


We can also notice that,

$$cov(X,X) = Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)\cdot(x_i - \mu_x)$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2$$



Sign of cov(X,Y) shows the tendency in linear relationship between x and y



cov(X,Y) > 0, X is directly proportional to Y, X ∝ Y

cov(X,Y) <0, X is inversly proportional to Y, X ∝ 1/Y

NOTE: As covariance depends on the scale of RV, it changes with changes in unit.

Pearson Correlation Coefficient (PCC):

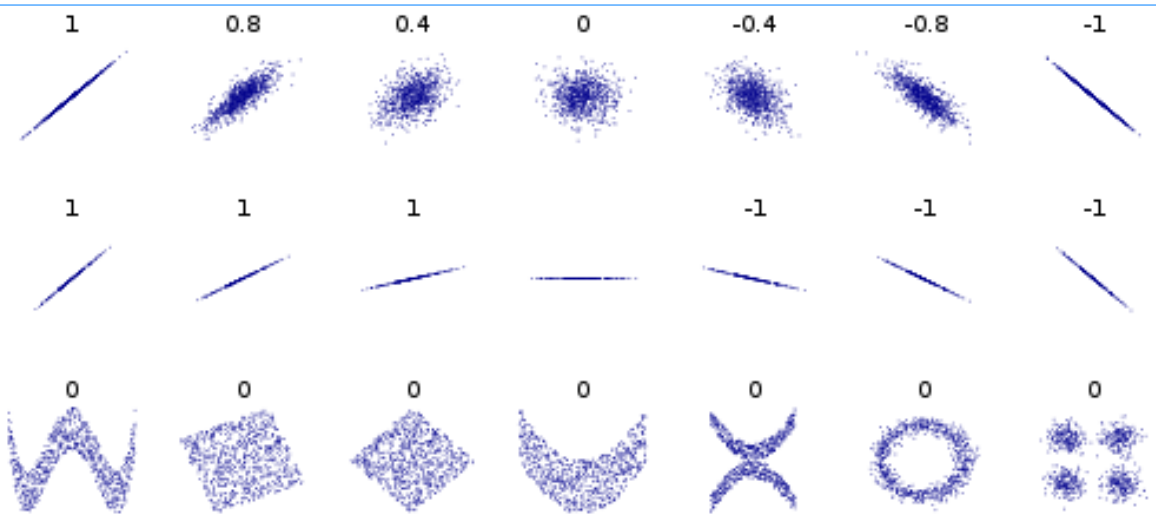
To overcome the problems with covariance we can use PCC.

PCC gives us the exact estimation of how much X and Y are correlated lineraly.

Value of PCC lies between -1 to 1. It is not dependent on the units.

Pearson Correlation Coefficient $\rho(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$

PCC does not work for non-linear correlation between X and Y and gives value 0.



Spearman rank correlation co-efficient (SRCC):

To overcome the problems with PCC we can use Spearman rank CC.

SRCC gives us the estimation of how much X and Y are correlated linearly or non-linearly.

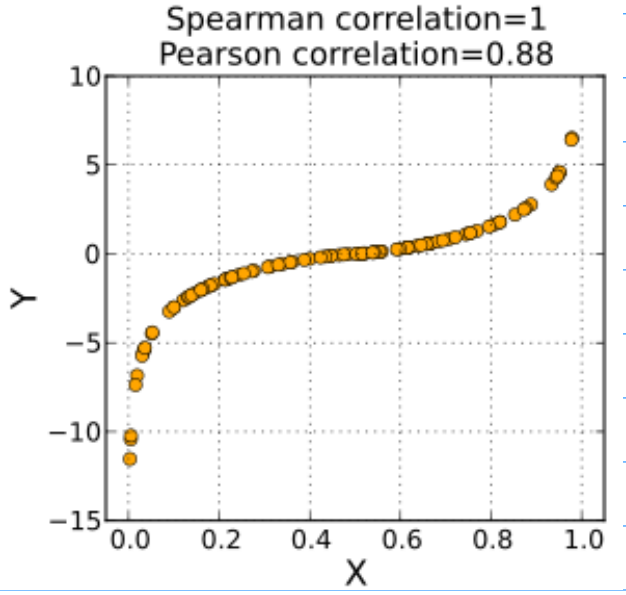
Value of SRCC lies between -1 to 1. It is not dependent on the units.

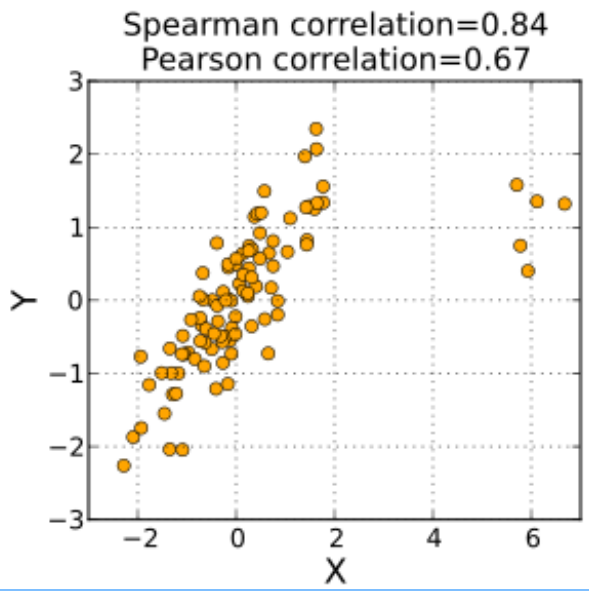
Spearman rank Correlation Coefficient $r_s = \rho(rg_X, rg_Y) = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_y}}$

where, rgX and rgY are rank variables of RV X and Y

How to calculate rank variable?

arrange your data in ascending order and at whatever place the value of data point come is the rank of that datapoint. By doing this process on RV X we get another rank variable rgX and similar process is done on Y.





Another way to calculate SRCC if all ranks are distinct integers,

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \text{ where } d_i = x_i - y_i$$

xi and yi are ith rank values

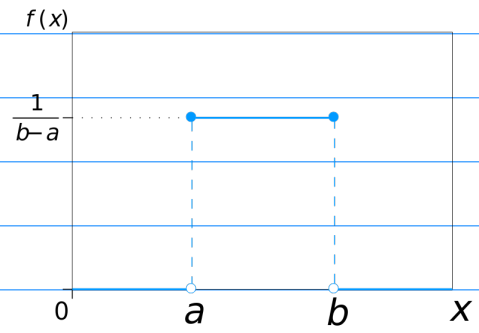
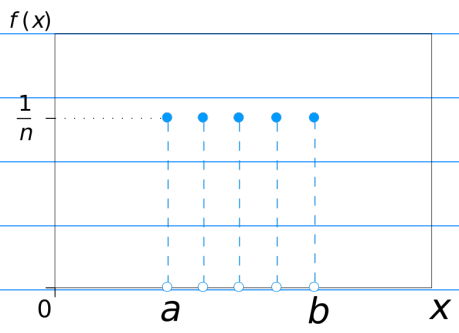
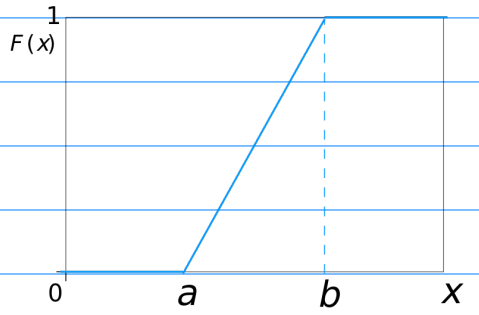
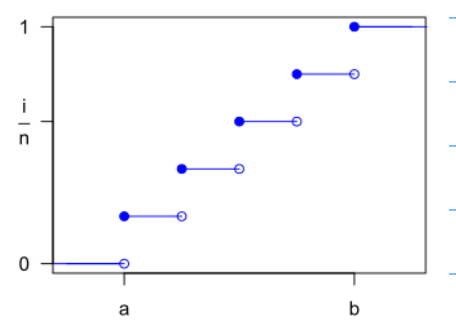
2. We have found correlation between two RV. Can we say that X causes Y?

If two random variable are correlated, We can't say that one is causes another.
OR due to value of one variable second is happening.

Example: Amount of chocolate eaten per person is correlated to number of Nobel Laureates per 10 million population.
But we can't say eating more chocolate causes increase in Nobel Laureates.

4. miscellaneous topics

1. Continuous and discrete uniform distribution

	Continuous	Discrete
PDF		
CDF		
Expectation (mean)	(a+b)/2	(a+b) / 2
Variance	$(b-a)^2 / 12$	$(n^2 - 1) / 12$

2. Weibull Distribution

Weibull distrubution is used to predict the maximum rainfall in a day in hydrology while constructing dam, to decide some parameters like height of a dam.

It is also used in survival analysis, weather forecasting, extreme value theory.