Probability and statistics.

→ Random variable :- variable which take values as outcome of experiment

Ex:- dice $\{1, 2, 3, 4, 5, 6\}$

R.V $X = \{1, 2, 3, 4, 5, 6\}$

tossing of a coin is a Random experiment.
R.V $Y = \{H, T\}$

dice roll:
$P(X=1) = \frac{1}{6} = P(X=2)$

$P(X$ is an even No$) = \frac{1}{2}$

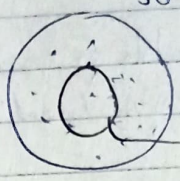
dice roll:
$X = \{1, 2, 3, 4, 5, 6\}$

height of a randomly picked student
$Y = 162.45$
or $Y = 132.62$
or Y can be any real number.

X is discrete RV Y is continous RV.
A discrete RV take discrete value from a finite set of values
A continous RV can take any real value.

Outlier: Y: height of student
$\{122.2, 146.4, 132.5, \ldots 12.26, 156.33\}$
↳ outlier.

* Population and sample:



estimate average human height

$$H = \frac{1}{7B} \sum_{i=1}^{7B} h_i$$

It is a very hard to measure height of each 7B+ human.

But what we can do is we can take a sample out of population and calculate average height

$$\bar{h} = \frac{1}{1000} * \sum_{i=1}^{1000} h_i$$

↳ sample size.

As sample size increases,

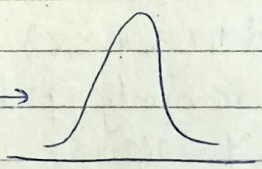
$$\bar{x} = \mu$$

$\bar{x}$ ↓ Sample mean

μ ↳ population-mean.

* Gaussian Distribution:

·PDF of a Gaussian dist R.V →



X: continous R.V

Lot of natural things follows gaussian distribution, hence it becomes very important for us to learn gaussian distribution.

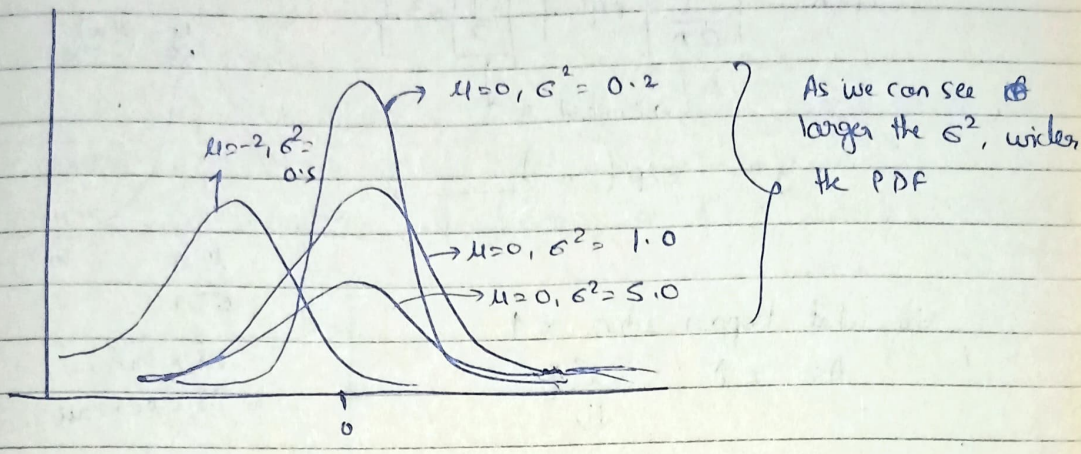
These distribution are very simple model, hence if we know that some feature follows a distribut we can predict lot of things about that feature.

mean and variance are parameter of a distribution because if we know mean is and variance and know which distribution is followed then we can draw PDF for those values.

But if we don't know the distribution followed by the values then mean and var are of no use.

Hence, to draw PDF we need all the three

(i) mean (ii) variance (iii) type of distribution



Parameters of Gaussian distribution : μ, σ^2

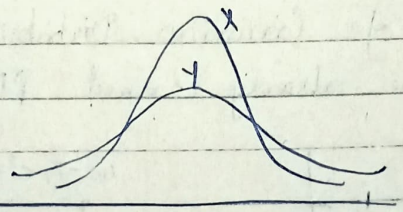
Let assume X is an R.V,

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

X follows Normal/Gaussian distn, parameters.

Ex: $X \sim N(0,2)$

$Y \sim N(0,4)$



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

$$P(X=x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

let assume $\mu=0$, $\sigma^2=1$

$\sigma=1$

$$P(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

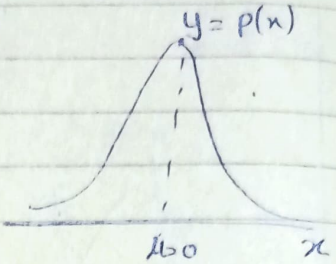
$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}$$

Lets try to plot the above eq.

$$p(x) = \left[\frac{1}{\sqrt{2\pi}}\right] \exp\left\{\left[-\frac{1}{2}\right]x^2\right\}$$

↳ constant

$$\therefore \quad y = \exp(-x^2), \longrightarrow$$



So, what happen when $x \uparrow$

As $x \uparrow$, $-x^2 \downarrow$

$\Downarrow$

$\exp(-x^2) \downarrow$

Conclusion:

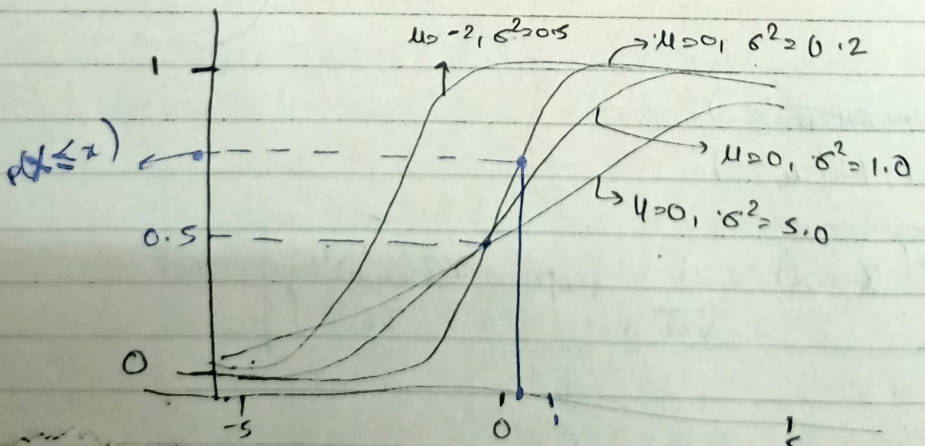
(i) As x move away from μ, $y \downarrow$

(ii) y is a symmetric func^n.

(iii) As x move away from μ, y reduce as $\exp(-x^2)$

→ CDF of Gaussian Distribution:

As we already learned PDF of Gaussian Distribution.



Height of CDF says $P(X \leq x)$

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

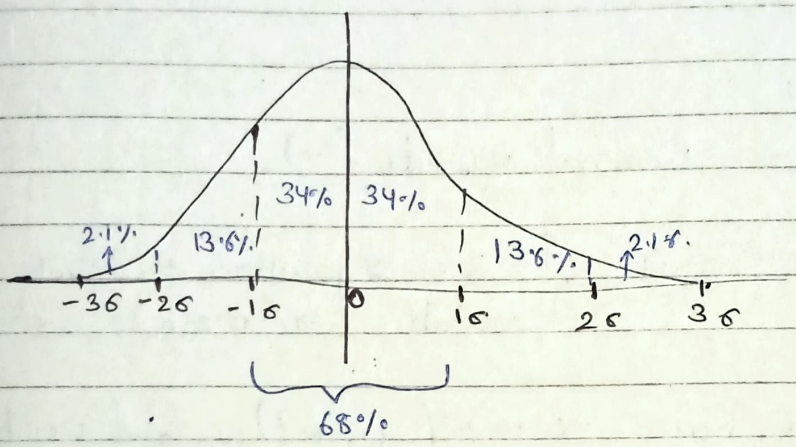
$$P(X \le x) = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

* 68-95-99.7 rule.

Let we are given $X \sim N(\mu, \sigma^2)$ where $\mu = 0$, $\sigma^2 = 4$

here $\sigma^2 = 4 \rightarrow \sigma = 2$

we can say lot of thing about X using 68-95-99.7 rule



B/w -1σ to 1σ we have 68% of our point

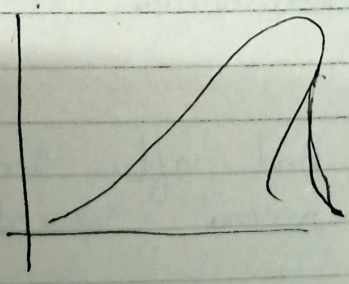
B/w -2σ to 2σ we have 95% " " "

B/w -3σ to 3σ we have 99.7% " " "

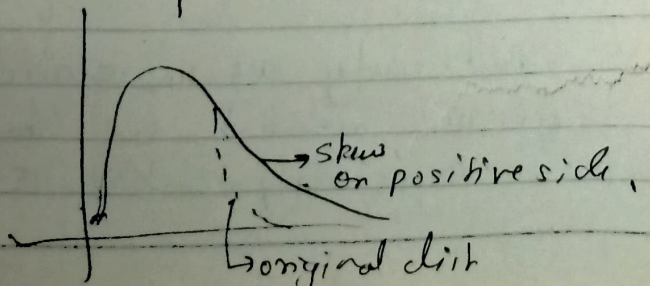
→ Symmetric distribution

→ Non-symmetric distribution.

→ Negative skew

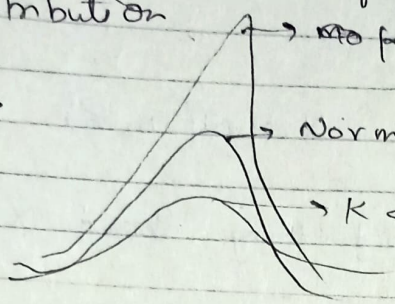


→ positive skew



→ Kurtosis :- says how sharp the peak is in distribution. for a Normal/Gau distr $K = 0$.

Let's see the Kurtosis flatter / or less peaked in distribution



→ Standard Normal variate (z)

① $z \sim N(0,1)$ } z is a gaussion distribute rando variable with mean = 0, $\sigma^2 = 1$

② let we have $X \sim N(\mu, \sigma^2)$, and we have ~~sampled~~ $[x_1 x_2, \cdots x_{50}]$

Given these observation of R.V X we can standardise these value. we can apply standardisation.

Standadisation : $\forall x$ in X, define x' s.t $x'_i = \frac{x_i - \mu}{\sigma}$

And after that we get X' which is standa normal variate.

But why are we standardising, :-

Ans The moment we standadi our data we can apply 68-95-99.7 rule.

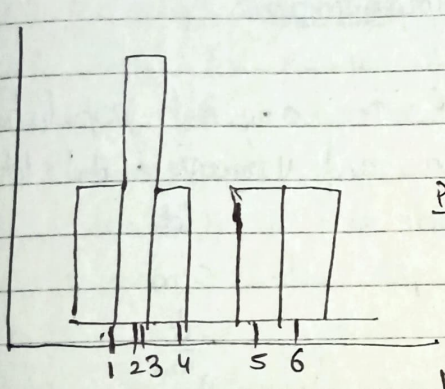
i.e. 68% of our x_i lie b/w -1 and 1 ($\sigma = 1$) here.

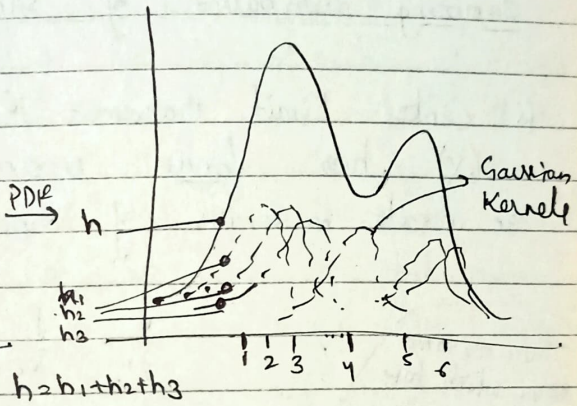
95% of x_i lie b/w -2 and 2

✱ Kernel density estimation

As we already know, that given a histogram we can get PDF from that histogram. We use Kernel density estimation to construct PDF using histogram.

Ex:-



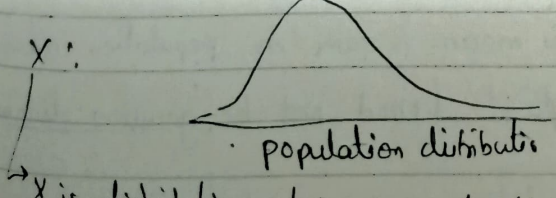


h = h1 + h2 + h3

For each data point we draw Gaussian Kernel function. At any point we sum up all the Kernels height at that point to get the height of PDF at that point. variance of Kernel is called bandwidth.

✱ Sampling distribution:-

Assume we have distribution of R.V X (not necessarily gaussian distribution)



X is distribution of incomes of all people in the world.

Now suppose we pick a sample S from the population, suppose sample size is n.

let say n = 30, Sample S_1 = sample of size 1000 observations

let again sample S_2 of size 30

30 Person

Let we keep doing the random sampling of 30 people and got in samples. and computing means of each sample

$S_1 \rightarrow \bar{x}_1$

$S_2 \rightarrow \bar{x}_2$

$S_3 \rightarrow \bar{x}_3$

$\vdots$

$S_m \rightarrow \bar{x}_m$

So we have $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_m$, these m values also have a distribution., this distribution is called Sampling distribution of sample mean.

☆ Centre Limit theorem :- If our original population dist. (X), has finite mean and variance :- And lets assume we create m samples of sample size n, i.e.

$S_1, S_2, \ldots S_m$

$\downarrow \bar{x}_1 \quad \downarrow \bar{x}_2 \quad \downarrow \bar{x}_m$

Sample mean.

→ distri of $\bar{x}_i$ = sampling dist of sample mean

There are some distri. which have infinite mean & variance like paretodistr

So, CLT says:

$\bar{x}_i$ is distributed with a gaussian distribution with mean μ & variance $\frac{\sigma^2}{n}$, $n \to \infty$

$$\bar{x}_i \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad n \to \infty$$

σ^2 → variance of population

n → sample size

This mean is same as population mean

Ex: X : income of people : (Need Not be gaussian distributed)

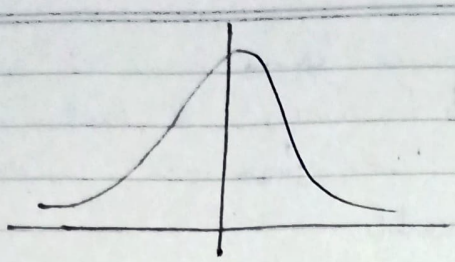
$\rightarrow \mu, \sigma^2$

$\rightarrow S_1, S_2, S_3 \ldots S_m \quad (|S| = 30)$

$\downarrow \bar{x}_1 \quad \downarrow \bar{x}_2 \quad \downarrow \bar{x}_3 \quad \ldots \quad \bar{x}_m \quad m = 1000$

Ans Take mean income in each sample.

↓ plot a distri of means :



3. mean of $\bar{x}_i$ = population mean.

Hence we can predict the mean of population using mean of sample mean.

Hence, for any distribution, to get their mean & variance we just need to ensure that, the population have finite mean & variance and we are done, we can calculate the mean & vari of a very large popula. by just using CLT in which we just take small samples, calculate their means and the distri of that Sample mean follows a Normal/Gauss dist with mean = popula mea & vari = $\frac{\sigma^2}{n}$.

★ Quantile - Quantile (Q-Q) plot :-

Let we have R.V $X : x_1\ x_2\ x_3 \cdots x_{500}$, How can we know if X is gaussian distributed, to Ans this Question we have various techniques, Q-Q plot is one of them., QQ plot is graphical methode.

→ How to plot Q-Q plot :

(i) Sort x_i & compute percentiles.

$x_1, x_2, \cdots x_{500}$

↓ sort

$x_1', x_2', \cdots x_{500}'$ (s.t. $x_1' \le x_2' \le x_3' \ldots$)

↓ Percentile

① ② ③ ... 1000

$x_5'\ x_{10}'\ x_{15}' \cdots x_{500}'$

$(x^{(1)})\ (x^{(2)})\ (x^{(3)}) \cdots (x^{(1000)})$

(ii) Create a R.V $Y \sim N(0,1)$, create 1000 observ from Y : $y_1, y_2, \cdots y_{1000}$

$y_1, y_2, y_3 \ldots y_{1000}$

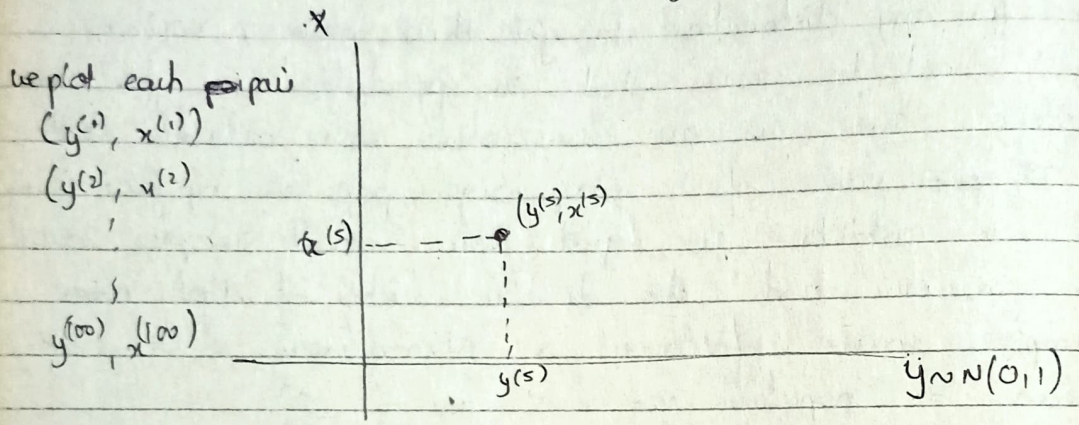
↓ sort

$y'_1, y'_2, y'_3 \ldots y'_{1000}$

↓ Percentile

$y^{(1)}, y^{(2)}, y^{(3)}, \ldots y^{(100)}$

③ plot Q-Q plot using $x^{(1)}, x^{(2)}, \ldots x^{(100)}$
$y^{(1)}, y^{(2)}, \ldots y^{(100)}$



* If the plot is roughly create a straight line then X has a gaussian distribution
* If plot is not straight line the X has not gaussian distribution.

So, we can check for any distribut ; we change y distribut and check if X follow y's distrib

Code:

```
std_normal = np.random.normal(loc=0, scale=1, size=1000)
```
(loc = mean, scale = σ^2)

```
measurements = np.random.normal(loc=20, scale=5, size=100)

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

* Limitation of Q-Q :- If no. of ~~sample~~ obsern in R.V is small then it becomes hard to interpret if R.V really follow Gaussia distn or not.

Using Q-Q we can check any 2 R.V if they have same distribution or not.

* Till now we have learned, R.V, PDF, CDF, Gaussian distn.
* But question is How/ where to use these distn in real wd These concepts are used in Exploratory Data Analysis.

* Chebyshev's Inequality :-
As we already know, if X is Normal dist and has mean μ and std- σ

the $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$.

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

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

Q) what if, we don't know the distribution, but we know mean $= \mu$, and std- σ, and we also know that mean is finite and $\sigma \neq 0$ and finite

In this case we can use chebyshev's Inequality:
chebyshev's Inequality :- If X is a R.V with finite mean μ and a non-zero and finite std $= \sigma$ .. We don't know the distribution of X. Then.

$$\boxed{P(|X - \mu| \geq K\sigma) \leq \frac{1}{K^2}}$$

OR

$$\boxed{P(\mu - K\sigma < X < \mu + K\sigma) \geq 1 - \frac{1}{K^2}}$$

Uniform distribution

There are 2 type of UD

(i) Discrete uniform distn

(ii) Continous " "

* similar to continous distribution, discrete distribution have Probability mass function (PMF) it is similar to PDF, but PDF is for continous R.V !

There are 2 parameter of Uniform distn, a and b

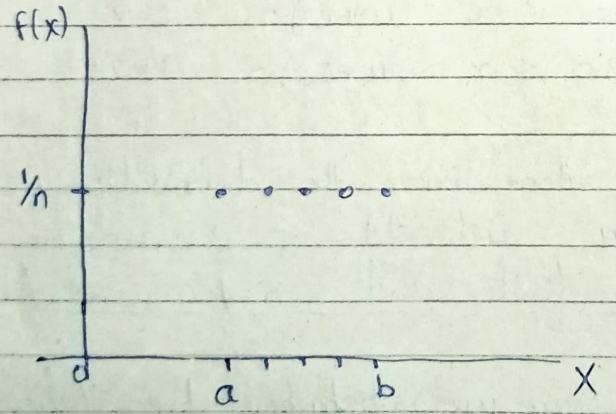
a and b are integer s.t $b \geq a$

$$n = b - a + 1$$

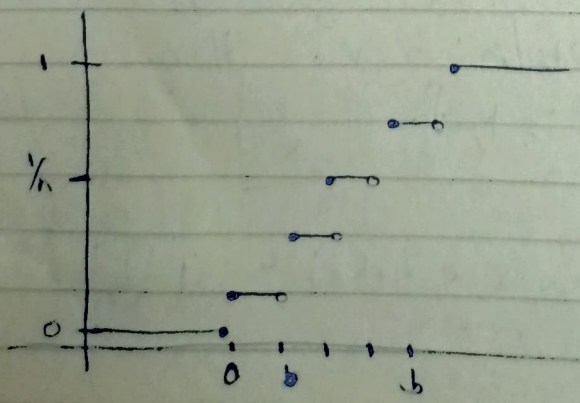
Notation → unif(a, b)

In a uniform distribution all the value are equal probable, i.e. each value have prob = $\frac{1}{n}$

PMF



CDF



Notation : $\mathcal{U}(a,b)$ or unif $\{a,b\}$

Parameter : $a \in \{\dots,-2,-1,0,1,2,\dots\}$

$b \in \{\dots,-2,-1,0,1,2,\dots\}$, $b \geq a$

$n = b-a+1$

Support : $\{a, a+1, a+2, \dots, b-1, b\}$

PMF = $1/n$

CDF = $\frac{\lfloor k \rfloor - a + 1}{n}$

mean = $(a+b)/2$

Media = $(a+b)/2$

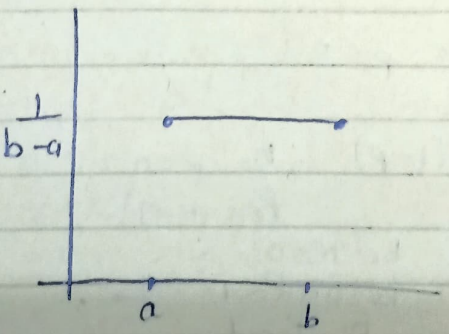
mode = N/A

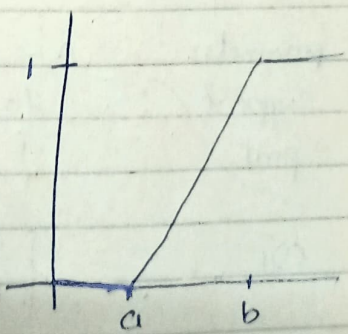
Variance = $\frac{(b-a+1)^2 - 1}{12}$

(11) Continuous UD: the continous uniform distri or ~~eg~~ rectangular distribution is a family of symmetric probability distribution s.t for each member of the family, all intervals of same length are equal probable.

parameter : a, b

As it ~~have~~ have continous R.V, it have PDF





PDF

Notation: $\mathcal{U}(a,b)$ or unif (a,b)

Parameter: $-\infty < a < b < \infty$

Support : $x \in [a,b]$

PDF : $\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$

CDF : $\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b) \\ 1 & \text{for } x \geq b \end{cases}$

Mean: $\frac{1}{2}(a+b)$

Media: $\frac{1}{2}(a+b)$

Varian: $\frac{1}{12}(b-a)^2$

✡ Random Numbes generator : UD

C++ , rand() or in any language rand() function uses uniform distribution to generate a random numbers.

UD can also be used to uniformly sample the data. suppose our iris data set which have 150 samples. from 150 we used a sample of 30 data.

```
n = 150
m = 30
p = m/n
sampled_data = []
for i in range(0,n):
    if random.random() <= p:
        sampled_data.append(d[i,:])
```

rand() used UD and gives a no. b/w (0-1) → treat that no. as probab. and check if prob ~~is equal~~ less than eq to eq prob.

✡ Bernoulli distribution

Benoulli distribution is used when R.V can take only 2 values. Like a coin toss experiment we have 2 outcome probability of one outcome is p and other outcom is 1-p

parameter : $0 < p < 1$, $p \in R$

support : {0, 1}

pmf : $\begin{cases} q = (1-p) & \text{for } K=0 \\ p & \text{for } K=1 \end{cases}$

CDF : $\begin{cases} 0 & \text{for } K<0 \\ 1-p & \text{for } 0 \le K < 1 \\ 1 & \text{for } K \ge 1 \end{cases}$

mean : p

media : $\begin{cases} 0 & \text{if } q>p \\ 0.5 & \text{if } q=p \\ 1 & \text{if } q<p \end{cases}$

variance : pq

* Binomial Distribution

Binomial Distribution with parameter n and p is the discrete prob distribution of the no: of successes in a seq: of n independent experiments, each asking a yes-no question.

when n=1, Binomial distribution = Bernauli distribution.

Notation : $B(n, p)$

parameter: $n \in N_0$ - no: of trials.

$p \in [0,1]$ - success probability in each trial

support : $k = \{0, 1, \ldots, n\}$

pmf : $\binom{n}{k} p^k \cdot (1-p)^{n-k}$

CDF : $I_{1-p}(n-k, 1+k)$

Mean : np

Variance : $np(1-p)$

* log-normal distribution

- continous probab: distn.. of a R.V whose logarithm is normally distributed. Thus if the R.V x is log-normally distributed then, $Y = \log(x)$ has a normal distn

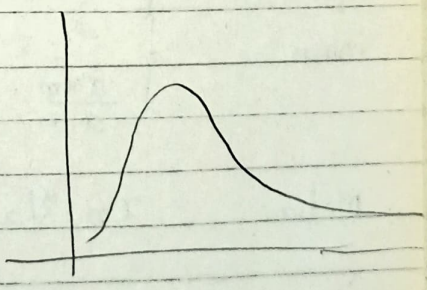
Notation : $\text{Lognomal}(\mu, \sigma^2)$

Parameter : $\mu \in (-\infty, +\infty)$

$\sigma > 0$

Support : $x \in (0, +\infty)$

PDF : $\frac{1}{x\sigma\sqrt{2\pi}} \cdot e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}$



mean : $\exp(\mu + \sigma^2/2)$

median : $\exp(\mu)$

Variance : $\cdot [\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2)$

- * The length of comments posted in Internet discussion forums follow a log-normal distribution.
- * The user's dwell time on online article follow a log-normal distn
- * Income of 97%-99% population is distributed log-normally.

* Power law distribution :-

Assume we have 2 variables, a power law is a functional relationship b/w two quantities, where a relative change in one quantity results in a proportional change in other quantity, relative change in independent of the initial size of those quantities : one quantity varies as a power of another.

For ex :- area of square in terms of length of its side, if the length is doubles, area is multiplied by factor of 4.

→ Pareto distribution :- when a distribution follows power law it is called pareto distribution

Parameter : $x_m > 0$ scale (real)
$\alpha > 0$ shape (real)

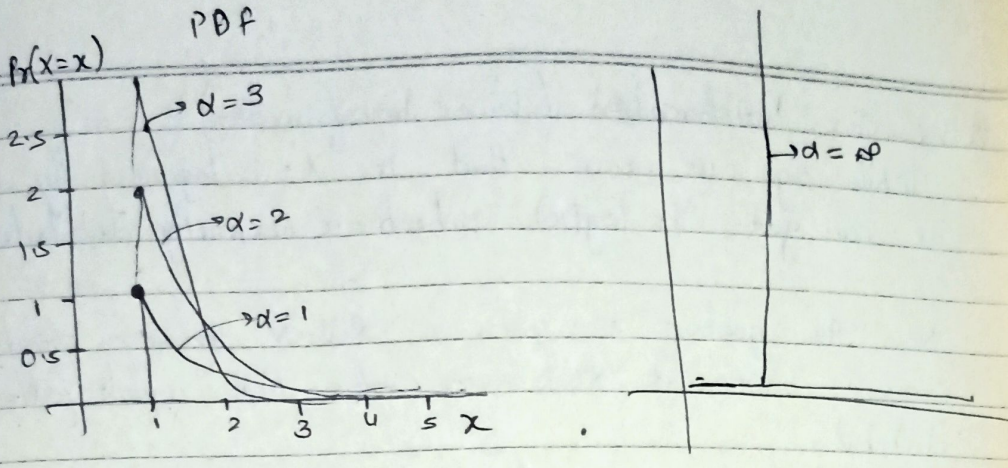
Support : $x \in [x_m, +\infty]$

PDF : $\frac{\alpha x_m^\alpha}{x^{\alpha+1}}$. for $x \geq x_m$

CDF : $1 - \left(\frac{x_m}{x}\right)^\alpha$. for $x \geq x_m$

mean : $\begin{cases} \infty & \text{for } \alpha \leq 1 \\ \frac{\alpha x_m}{\alpha - 1} & \text{for } \alpha > 1 \end{cases}$

Median : $x_m \sqrt[\alpha]{2}$

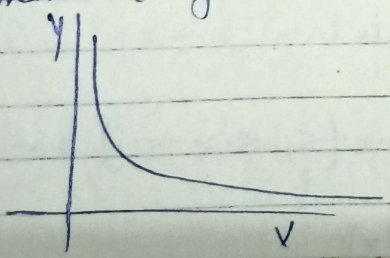


$\alpha = \infty$ · give a Dirac delta func, when $\alpha = \infty$ there is only one value and give a peak value.

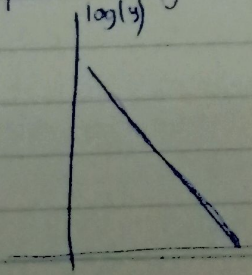
In pareto distri we can see there is a peak value and then distribution start falling.

- File size distr of Internet traffic use the TCP prot follow pareto distri
→ Hard disk drive error rates follow pareto distribution.

* log-log plot is used to determine if the given plot is power law plot or not.
· Let assume we have 2 vari x & y :, when we plot them we get something like this



To check if the above plot is power law or not · ,"plot another plot of log(y) and log(x)



If plot comes out like this a straight line, then the two variable follow power law.

☆ Box-Cox transformation (Power transformation)

while ago we saw that if X is log-normally distributed we can get $Y = \log(X)$ which is normally distributed.

Now the question is given a R.V X having pareto distribution can we convert X to some other R.V which follows gaussian distribution.

Ans is Yes, we can do that using Box-Cox distribution.

Pareto ~ X : $x_1, x_2, \ldots, x_n$ → conversion.
Gaussian ~ Y : $y_1, y_2, \ldots, y_n$

① box-cox(x) = lambda(λ)

$x_1, x_2, x_3 \ldots x_n$

② $$y_i = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \quad \forall i : 1 \to n \\ \log(x_i) & \text{if } \lambda = 0 \end{cases}$$

→ This case suggest that X is lognormal when $\lambda = 0$

#code:

```
from scipy import stats

x = stats.loggamma.rvs(5, size=500) + 5   # R.V of gamma distn. It is one of pareto dist.

xt, _ = stats.boxcox(x)
```

* Application of non-gaussian distribution

(i) Uniform → Random generate
(ii) Bernulli, Binomial
(iii) log-normal, pareto
and many more non-gaussian dist are there.

* Covariance :- Lets assume we have 2 R.V X and Y
X = heights of student
Y = weights of student

Q: is there a relationship b/w X & Y
There are 3 measure to check if 2 R.V are related
(i) Co-variance
(ii) Pearson correlation coeffs
(iii) Spearman rank corr. coeffs.

Let's understand co-variance :-

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

(μ_x → mean of x, μ_y → mean of Y)

$Cov(X,Y) = +ve$, if X↑ then Y↑
$Cov(X,Y) = -ve$ if X↑ then Y↓

* One problem with cov is that for same data set if we change the unit the cov may not be same. Suppose we find value cov(X↓ fixed, Y↓ fixed) then cov(X↓ incm kg, Y↓ kg) both of them will not be equal.

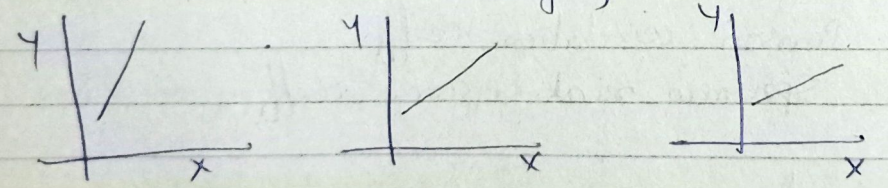
(11) Pearson correlation coeff (PCC)

$$\rho_{x,y} = cov(x,y) / (\sigma_x * \sigma_y)$$

Problem with cov was it was not taking in use of variability that std.
In cov we knew that if x↑ and y↑ the cov is +ve but we don't know how much positive.

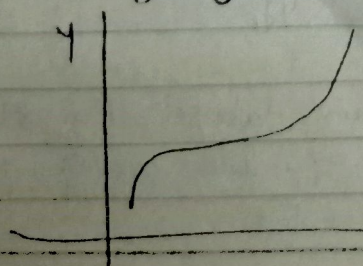
value of ρ is always $-1 \leq \rho \leq 1$
when $\rho = -1$ that means as x↑ y↓ perfectly
when $\rho = 1$ " " " x↑ y↑ perfectly on straight line.
$\rho = 0$ if there is no relation b/w X and Y

But PCC also fails, PCC works well only if X and Y have linear relationship.
The other thing is if X any Y are lieing in straight line then slope of line is not considered that means it doesnot tell any thing about how the R.V are changing.



$\rho = 1$ $\rho = 1$ $\rho = 1$

(11) Spearman rank-corr-coeff : This works for some slightly linear relationship.



$\rho_{x,y} = 0.88$

$\rho_{spearm} = 1$

↑ rank of x_i in sorted data

	x = height	y = weight	r_x	r_y
s_1	160	52	4	3
s_2	150	66	2	4
s_3	170	68	5	5
s_4	140	46	1	1
s_5	158	51	3	2

$$r = \rho_{r_x, r_y}$$

Rank-corr: doesn't care about linearity, it just check if $x\uparrow$ $y\uparrow$ or $x\uparrow$ $y\downarrow$ using ranks of x & y..

→ Correlation doesnot imply causation
Correlation says only if $x\uparrow$ $y\uparrow$ but we cannot conclude that y happens due to x. Just becau 2 r.v are correlated that doesn't mean that x causes y or y causes x.

Application of Correlation:-

☆ C.I for mean(μ) of a R.V (any distn not necessarily gaussian)

let a R.V $X \sim F$ with pop-mean of μ & std-dev of σ

$\{x_1, x_2, \ldots x_{10}\}$ → sample of size, $n = 10$

Q) What is the 95% confidence Interval (C.I) of μ.

Case 1:- Assume somebody told us the popula std-deviation.

☆ Confidence Interval :- Assume we have R.V X we don't know distribution

Sol: $\{x_1, x_2, \ldots x_{10}\}$ → sample of size 10 from X

we want to estimate population mean of X.

One way is we can say populatio mean will approx Eq. to sample mean.

$\mu \approx \bar{x}$

μ ↓ pop mean, $\bar{x}$ ↳ sample mean

But This is not but approximation:-

suppose $\{x_1, x_2, \ldots x_{10}\}$

$\{180, 162, 158, 172, 168, 150, 171, 183, 165, 176\}$ ← heights of population

POINT ESTIMATE of $\mu = \frac{1}{10}\sum_{i=1}^{10} x_i = 168.5$ cm

other thing we can say that

$\mu \in [162.1, 174.9]$ with 95% probability.

↳ Confidence Interval

☆ Computing C.I given the underlying distribution:-

$X \sim N(\mu, \sigma)$; $\mu = 168$cm, $\sigma = 5$cm

heights ↓ Gaussian distn.

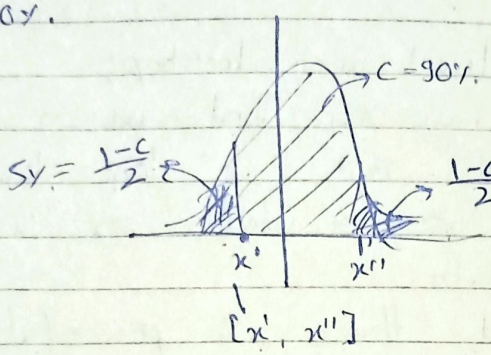
Q) C.I of hight ?

for a gaussian distn we know that 95% of data lies in $[\mu-2\sigma, \mu+2\sigma]$

↳ [158, 178]

hence this is CI with 95% probability that height lies in interval [158, 178]

Suppose someone ask a interval with confidence = 90% or 80% of 70%.



We can use Normal distn table and find the value of x', x''.

C.I. for mean(μ) of a R.V

Let a R.V X with some distr (Not known) with pop-mean μ & X std-dev σ.

we have a sample of size 10 of $\{x_1, x_2, x_3, \cdots x_{10}\}$

$\{180, 162, 158, 172, 168, 150, 171, 183, 165, 176\}$

Q) what is the 95% C.I. of μ.

There are 2 cases for this

Case 1: Assume somebody told us the pop-pop std-devi

$\sigma = 5cm$.

Acc. to CLT: $\bar{x}$ = sample mean $= \frac{1}{10}\sum_{i=1}^{10} x_i$

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Now, we can say μ lies in $\left[\bar{x}-\frac{2\sigma}{\sqrt{n}}, \bar{x}+\frac{2\sigma}{\sqrt{n}}\right]$ with 95%.

Case II : If we don't know μ & σ
we use t-distribution

Acc. to t-distr $\bar{x} \sim t(n-1)$
($\bar{x}$ ↓ sample mean; $n-1$ ↳ degrees of freedom)

→ t-distribution : This distr was developed to calculate C.I when pop std. mean is unknown.

* Confidence Interval using bootstrap!-
previously we have seen that we can calculate C.I. for μ of R.V But we don't know how to calculate C.I of $\bar{x}$, σ, media .or 90^{th} percen or any other statistical data.

for that there is a powerful technique called bootstrap which can be used to calculate C.I for any of the above term (σ, median etc)

→ $X \sim$ Any distn

task : estimate C.I of median of X.
let's assume we have a sample of size n
$\{x_1, x_2, \ldots x_n\}$
using only this we need to compute C.I of median of x. we don't know type of distr.

① generate new sample using give sample.

$S_1 : x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots x_m^{(1)}$ s.t $m \leq n$

↳ random sample of size generate from S.

Repeation is allowed i.e. some values can be repeated.

One way to do this is generate a uniform R.V $U(1, n)$ and generate Random sample : m times with U R

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

↓ using sampling with replacement

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

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

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

.

$S_K : x_1^{(K)}, x_2^{(K)}, x_3^{(K)}, \ldots, x_m^{(K)}$

} bootstrap samples

for each sample compute medians

$S_1 \rightarrow m_1$

$S_2 \rightarrow m_2$

$S_3 \rightarrow m_3$

.

$S_K \rightarrow m_K$

let say $K = 1000$

$m_1, m_2, m_3 \ldots m_{1000}$

↓ sort

$m_1', m_2', m_3', \ldots m_{1000}'$ (increasing order)

↓

$[m_{25}', m_{975}']$

95% value → 95% of 1000

sorted

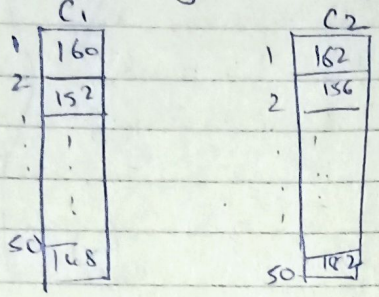
This process can be done for anything i.e. μ, σ, σ^2, possible.

This called a non-parametric technique.

↳ doesn't make any assumption about distribution.

Hypothesis testing :-

Assume we have 2 classes C1 & C2 and assum we taken ht height of 50 students from both classes.



a) Is there a diff in heights of students in C1 & C2 students.

To ans this ques we use Hypothesis testing :

① choosing a test-statistics :- choose a statistical parameter

$\mu_2 - \mu_1$

mean of C1 ↳ mean of C2

② Null hypothesis (H_0)

H_0 : no-difference in μ_1 & μ_2

Alternatin hypothes (H_1) :- Inverse of H_0

H_1 : there is diff in μ_1 & μ_2

Assume H_0 is true.

③ compute p-value :- p value says what is the probability of observing your test statistic If H_0 is true assume we got $\mu_2 - \mu_1 = 10$, then p-value says what is the prob to get $\mu_2 - \mu_1 = 10$ if H_0 is true.

If p-value = 0.9 it means that probability of $\mu_2 \mu_1$ is 0.9 if H_0 is true

if p-value is High then we accept H_0 is true.

Ex:- Given a coin, determine if the coin is biased towards heads or not.

~~biased toward~~.

we can answer questions like this ~~you~~ using Hypothesis testing.

Initially, we try to prove that ~~it~~ using basic probability.

→ Using Basic prob:-

→ design a expt: flip a coin 5 time and count No. of heads.

let No. of Head = X ← Test statistics

→ perform the experiment :- flip, flip, flip, flip, flip

H H H H H

X = 5 ← observation by experimenter.

$P(X=5 \mid \underbrace{\text{coin is not biased towards heads}}_{H_0}) =$

$P(X=5 \mid H_0) \rightarrow \frac{1}{(2)^5} = (0.5)^5 = 3\%$

There is 3% chance of getting 5 Heads in 5 tosses if the coin is not biased towards Head.

Hypothesis testing works like this:-

$P(\text{obs by exp} \mid \text{assumption}) = 3\%$ ← small

P-value.

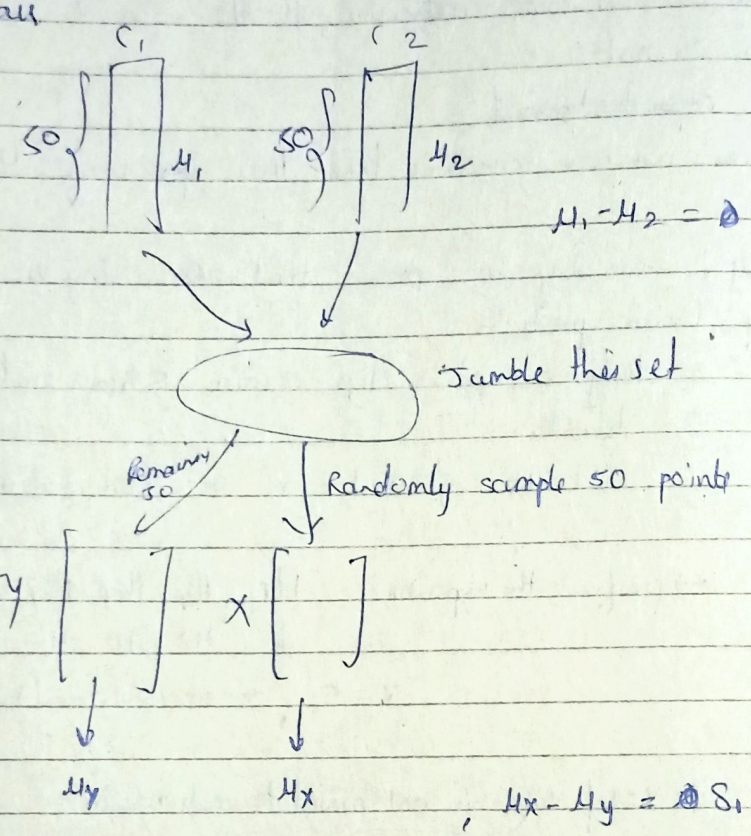
If $P(\text{Obs} \mid H_0) < 5\%$ then H_0 is not true.

Hence coin is biased.

* We need to be careful about (i) sample size or ~~desi~~ experiment size
(ii) Defining Null Hypothesis.
(iii) design X.

☆ p-value calculation (permutation testing):-

Assume we have C_1 and C_2 and 50 height from both class



Again resample the jumble (mixed set) and again calculate $H_y \times H_x$ and Repeat this process say 10K time.

we get 10K δ's

$\delta_1, \delta_2, \ldots, \delta_{10K}$

sort

$\delta_1', \delta_2', \delta_3', \ldots, \delta_{10K}'$

find correct position of Δ in sorted δ's

If there are 5% value greater the Δ in sorted δ's then p-value = 0.05%.

If x% value greater than Δ in sorted δ's then p-value = x/100

Kolmogorov – Smirnov

↑

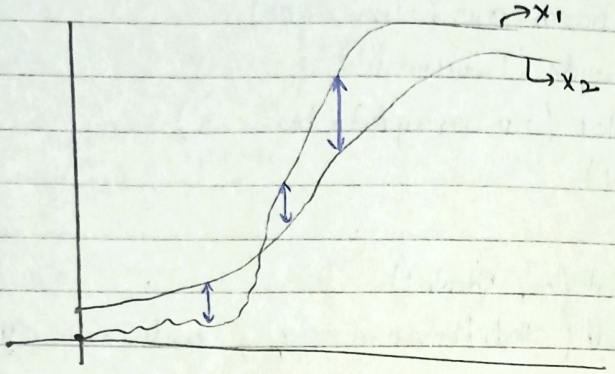
A K-S test for similarity of 2 distributions

2 R.V X_1 : $[x_1^1, x_2^1, \ldots x_n^1]$

X_2 : $[x_1^2, x_2^2, \ldots x_m^2]$

is distribution X_1 is same as X_2

First we plot CDF of both X_1 & X_2



Ho in K-S test = X_1 & X_2 have same distribution

Test statistic :- $D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$

$F_{1,n}(x)$ → CDF, $F_{2,m}(x)$ → CDF

↓ maximum difference b/w CDF of X_1 & X_2

If n & m are large · Then $D_{n,m} = 0$ and here Ho = true.

The null hypothesis is rejected at level α if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$$

→ when n & m are small we get higher cutoff to reject Ho

→ If n & m are large we get a small cutoff to reject Ho.

where n & m as size of X_1 & X_2

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

$$c(\alpha) = \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)}$$

Code

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

x = stats.norm.rvs(size=1000)
sns.set_style('whitegrid')
sns.kdeplot(np.array(x), bw=0.5)
plt.show()

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

o/p : KstestResult(statistics=0.021308, pvalue=0.75424)

```
y = np.random.uniform(0,1,10000),
stats.kstest(y, 'norm')
```

KstestResult(statistic=0.501596, pvalue=0.0).

Another example :- Hypothesis testing

Task :- two cities C_1 & C_2 determine if population mean of heights of people living in these cities is same or not.

C_1 C_2

problem is there are lot of people we can't calculate by going to each people in city and calculate height.

① Design experiment :- Randomly pick 50 peoples from each city

$$50 \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{50} \end{bmatrix} \quad C_1 \qquad 50 \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{50} \end{bmatrix} \quad C_2 \longrightarrow \text{samples of size 50}$$

Sample mean $\mu_1 = \frac{h_1 + h_2 + h_3 \cdots + h_{50}}{50}$ Sample mean $\mu_2 = \frac{h_1 + h_2 + \cdots h_{50}}{50}$

(iii) Test statistics : $\mu_1 - \mu_2 = x$

let $\mu_1 = 162$ $\mu_2 = 167$

$|\mu_1 - \mu_2| = 5$

(i) Null Hypothesis :- There is no difference in population means of height of people of 2 cities.

(iv) p-value :- $p(x = 5cm \mid H_0)$

Diff in sample mean of sample size of 50

How to calculate $p(x = 5cm / H_0)$:- using Resampling & Permutation Test

* proportional sampling :-

let we have a matrix d with 5 element

$d =$

d_1	d_2	d_3	d_4	d_5
2.0	6.0	1.2	5.8	20.0
1	2	3	4	5 = n

Task :- I want to pick an element amongst these n-elements s.t the probability of picking an element is proportional to d_i's

If we randomly pick a value, each of the value is equally prob But we don't want to pick randomly. But we pick element s.t picking the 5th element is very large because 5th element is maximum.

Similar, the prob of picking the 3rd element must be extremely small because 3rd element is least.

So, how do we do it.

Step 1a : $S = \sum_{i=1}^{n} d_i$ { sum of all values }

$S = 2 + 6 + 1.2 + 5.8 + 20 = 35$

Step 1b : $d_i' = d_i / S$ { Normalizing these values }

$d_1' = 0.0571$, $d_2' = 0.17148$, $d_3' = 0.034$, $d_4' = 0.165$, $d_5' = 0.571$

Step 1c: Commulative Normalized sum.

$d_1 = 0.0571$	$\tilde{d}_1 = d_1 = 0.0571$
$d_2' = 0.171428$	$\tilde{d}_2 = \tilde{d}_1 + d_2' = 0.228528$
$d_3' = 0.0343$	$\tilde{d}_3 = \tilde{d}_2 + d_3' = 0.262828$
$d_4' = 0.1657$	$\tilde{d}_4 = \tilde{d}_3 + d_4' = 0.428528$
$d_5' = 0.5714$	$\tilde{d}_5 = \tilde{d}_4 + d_5' = 1.00$

Step 2: sample one value from uniform Random vale b/w 0

```
r = numpy.random.uniform(0.0, 0.1, 1)
```

let $r = 0.6$

Step3: If $r \leq \tilde{d}_1$

return 1

else if $r \leq \tilde{d}_2$

return 2

else if $r \leq \tilde{d}_3$

return 3

Proof: prob of picking up d_i

= prob. of r lying b/w $\tilde{d}_{i-1}$ & $\tilde{d}_i$

$= d_i'$

$d_i' \propto d_i$ $\quad \because d_i' = \frac{d_i}{S}$