Probability & Statistics

① Random variable

1.1 Definition:

Let us understand with an example

eg:

dice: six sides → {1, 2, 3, 4, 5, 6}

↓

rolling dice can have random outcomes

Let's denote random variable by.

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

↓ random variable.

↓ random outcomes

Eg: Tossing of a coin.

$Y = \{H, T\}$

↑

random variable.

Dice roll :-

- $P(X=1) = \frac{1}{6} = P(X=2)$

where

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

All are equally likely to occur.

- $P(X \text{ is even}) = P(X=2) + P(X=4) + P(X=6)$

$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$

$= \frac{1}{2}$

General $P(X=x_1)$

read as 'probability that random variable X is x_1'.

Concise way: $P(x_1)$. [random variable implicitly]

1.2 Types of random variable

(i) Discrete random variable

can take a value from set of finite value.

eg: dice roll

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

(ii) Continuous random variable:

can take any real value. eg

Height: 120cm — 140cm

$Y = 162.45$
$Y = 132.65$] Continuous random variable.

② Outlier

Eg:

Y: weight of a student

{122.2, 146.4, 132.5, ... 12.26, 156.23, ...}

↓

Outlier

Outlier may be:

(i) Genuine outlier

(ii) OR caused by

- Human error
- Observation error

③ Population & Sample

* population [set of all population in the world]

sample [subset of the population

Eg: sample of 10 size 120 kgs

Mean of population — average height of

$$\mu = \frac{1}{73} \sum_{i=1}^{73} h_i$$

[Since we cannot take population as whole,

Mean of average height of sample (1000) — sample is taken for analysis]

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

[called random-sample]

Typically, $\bar{x}$ used to represent mean of the sample.

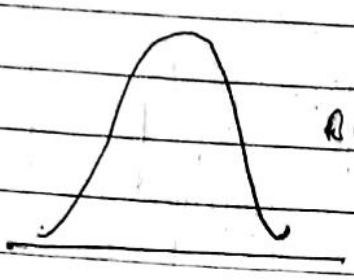
As sample size increases

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

↑ sample mean ↑ population mean.

[Sample is taken to estimate statistics of some property (like height) of population]

(5) Gaussian Distribution / Normal Distribution



Bell shaped curve.

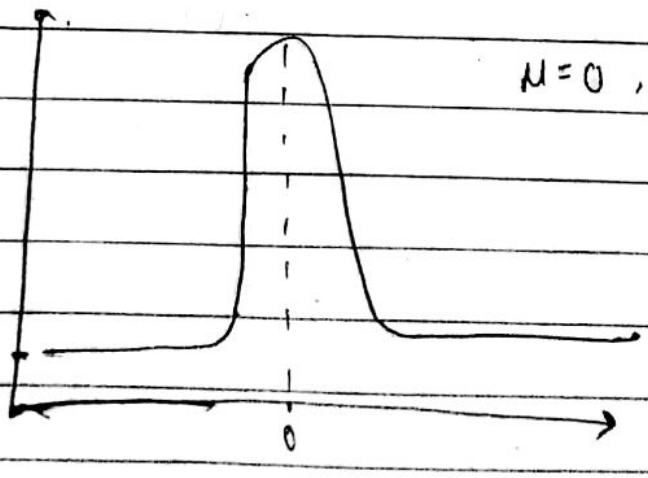
If

Probability density function (PDF) of a gaussian distributed random variable.

Why learn distributions?

↳ Helps in formulation of simple models giving details on various features and attributes

In General



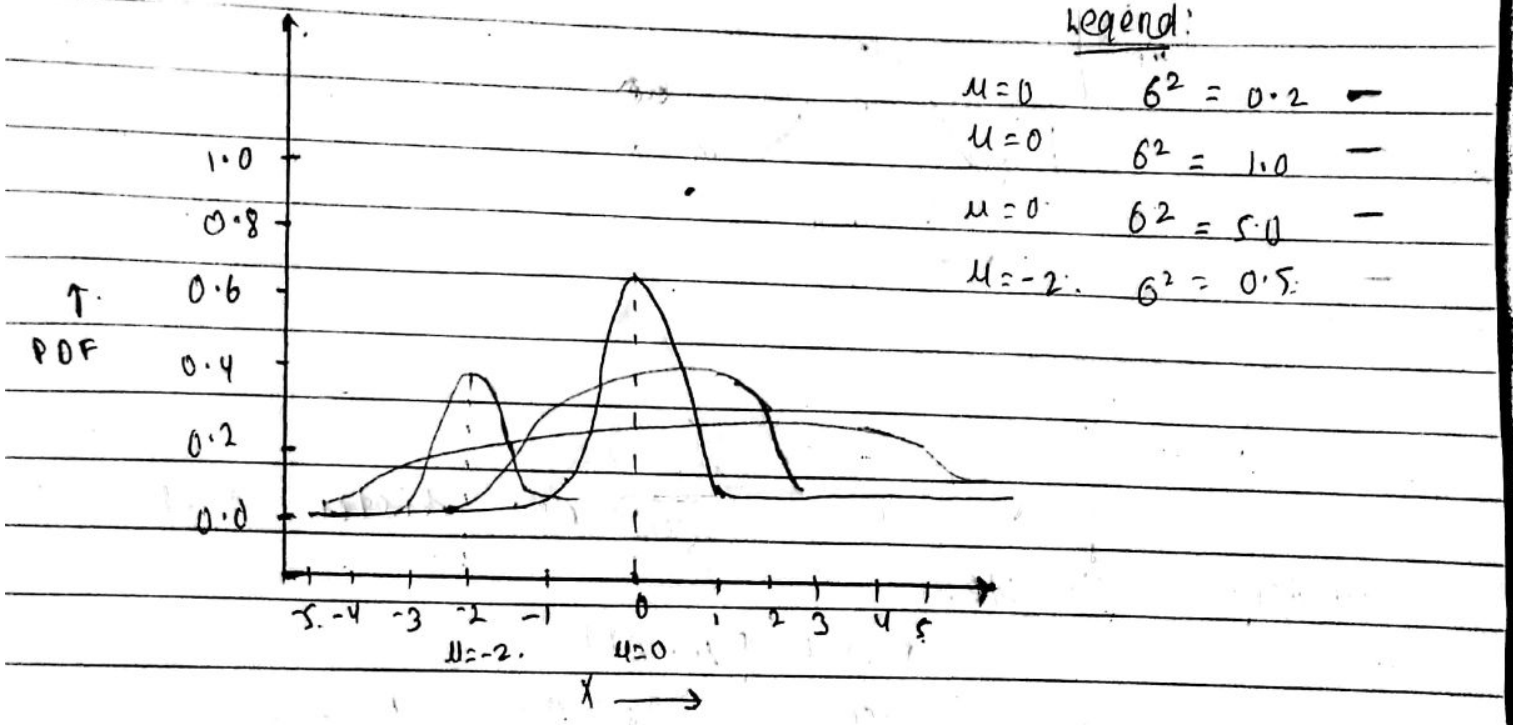
$\mu = 0, \sigma^2 = 0.2$

μ = mean

σ^2 = variance.

i.e. just by knowing σ, μ and σ^2, we can find out PDF.

[μ and σ^2 are parameters of Gaussian distribution]



Here, variance represents the spread.
Greater the spread, more pointer the PDF.

normal / Gaussian distribution.

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

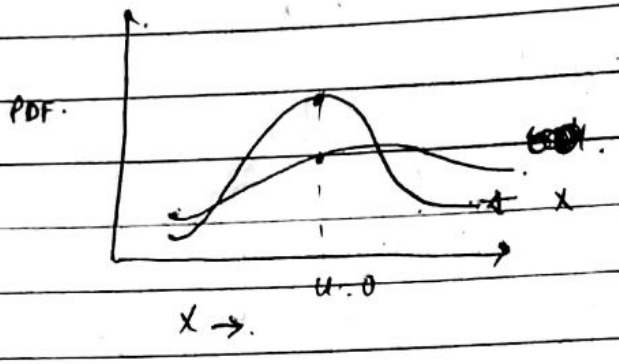
- X: random variable
- follows: $\sim$
- N: parameters of Gaussian distribution
- μ: mean
- σ^2: variance

Eg. $X \sim N(0, 2)$

Scenario:

$x \sim N(0,2)$

$y \sim N(0,4)$



Conclusion

If x is a gaussian distributed random variable!

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

Just by knowing μ, σ^2, we can calculate PDF.

And, 'u' also lies at the peak of bell shaped curve.

Mathematically,

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

↑ random variable ↑ value

To simplify and understand,

let's assume,

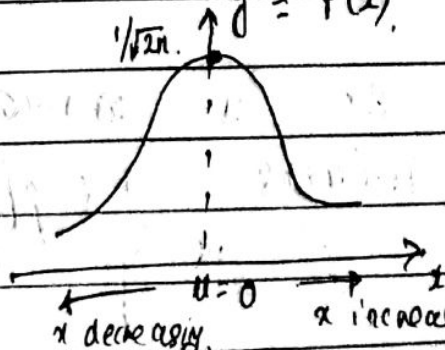
$\mu = 0$, $\sigma^2 = 1$; $\sigma = 1$

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

Ignoring constants

$$y = \exp(-x^2)$$

Hence, when we plot for $y = \exp(-x^2)$, we get a bell shaped curve.



$x \uparrow$ or $x \downarrow$ → $-(x^2) \downarrow$ (always reduces)

↓

And hence $y \downarrow$ (also reduces)

ie: $y = \exp(-x^2) \downarrow$ "

Conclusion from Gaussian distribution / normal distribution.

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

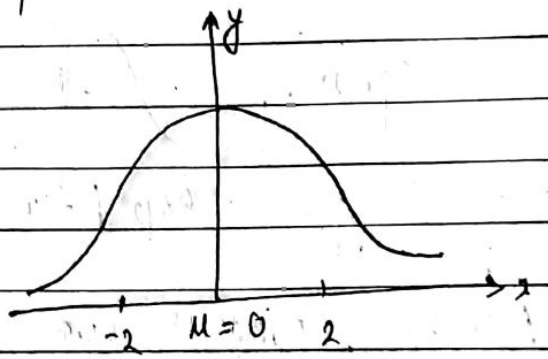
↓

Simplified

$$P(x) = y = \exp(-x^2)$$

↓

Graphical



Here,

(1) PDF is symmetric.

(2) As x moves away from μ, y reduces exp(-x²) i.e. y reduces very fast

ie:

$y = \exp(-x^2)$

$x = 0$ $\quad y = 1$

$x = 1$ } 1x $\quad y = \exp(-1) = 0.3679$ }

$x = 2$ } 1.5x. $\quad y = \exp(-4) = 0.018$ } 10x decrease

$y = \exp(-9) = 0.000123$ } 100x. decrease

It's clear here.

for x 1x increase, y 10x. decrease.

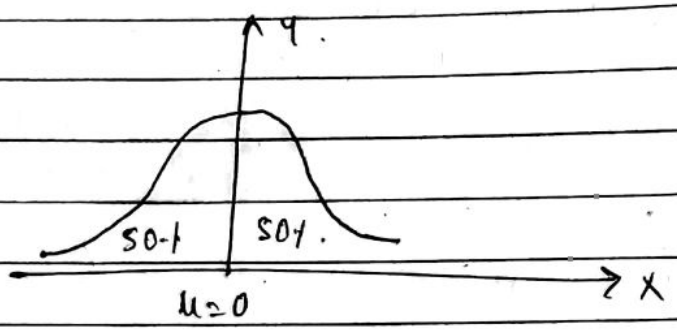
for x 1.5x increase, y 100x decrease.

(6) Cumulative Distribution function (CDF) of Gaussian / Normal distribution :-

(7) Symmetric distribution, Skewness and Kurtosis :-

- Symmetric distribution :-

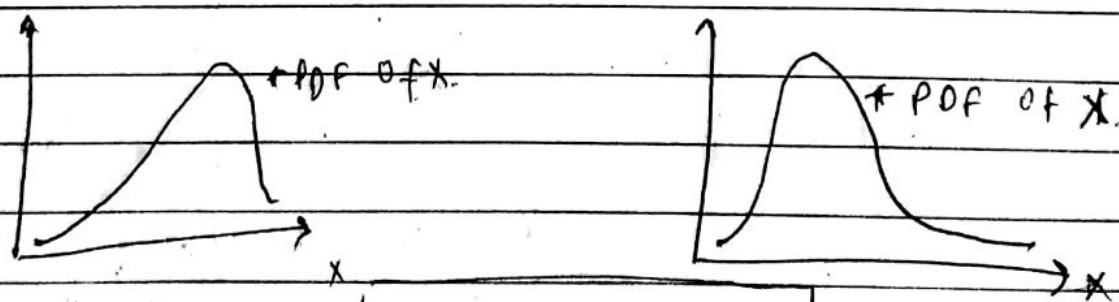
Ex: Gaussian distribution



to quantify above non symmetric distribution

(1) - Skewness :-

- negative skew
- positive skew

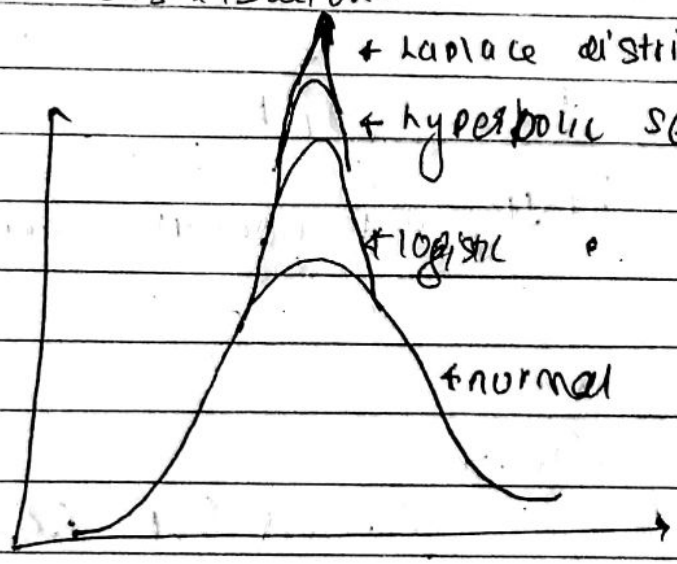


- left tail is longer | X: Random variable | - right tail is long

Here, skewness tells us how different the distribution is from symmetric distribution.

· KURTOSIS

Measures peakness / sharpness of symmetric distribution.



(8) Standard normal variate (z) and standardization:

- Z (SNV)

$Z \sim N(0, 1)$

i.e $\mu = 0$

$\sigma^2 = 1$

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

↳ $[x_1, x_2 \ldots x_{50}]$

Hence

Standardization:

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

$x_i' \sim N(0, 1)$.

↳ Standard normal variate (Z)

Here,

Conclusion:

Given

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

Then, we can get standard normal variate (z) just by substracting every value of x from mean and then dividing by σ.

ie:

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

Hence

$z \sim N(0, 1)$

The process is called standardization.

Why standardization?

If we standardize, we get to know that:

If we have any normal distribution:

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

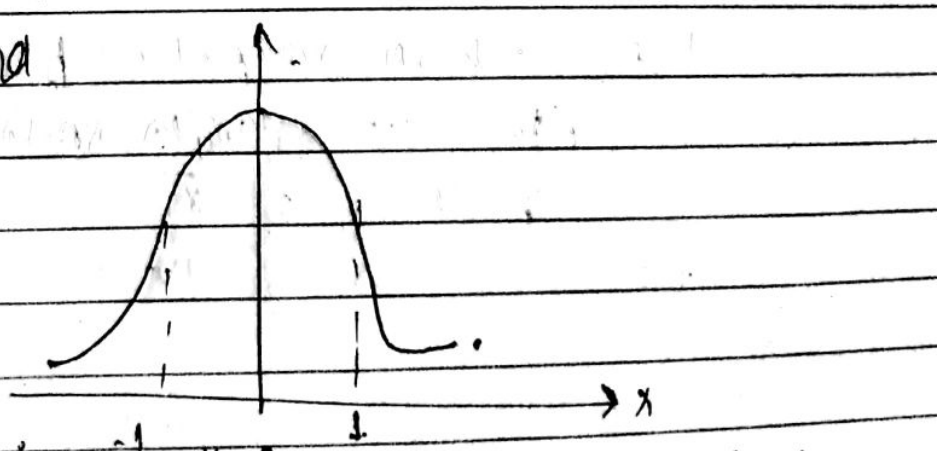
Then on standardization via:

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

$$\therefore Z \sim N(0,1)$$

So on standardization, we get to have the normally distributed curve with $\mu = 0$ and $\sigma^2 = 1$.

and



Also 34.1% of x_i lies between -1 and 1

68.1% x_i lies between -1 and 1.

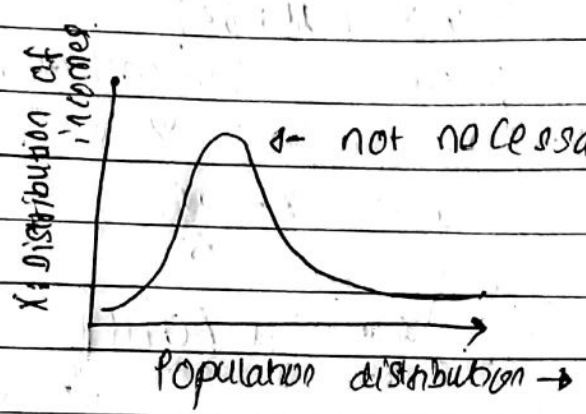
95.1% x_i lies between -2 and 2

(9) Kernel density estimation.

(10) Sample distribution and central limit theorem.

Sample distribution:-

Let's consider:



Now, let us take random samples data from given population distribution.

let

(i) Random sample of size $(n = 30) \rightarrow S_1$
Let's the sample mean (of income) of $S_1 = \bar{x}_1$

(ii) " " " " " $\rightarrow S_2$
" $S_2 = \bar{x}_2$

(iii) " " " " $(n = m) \rightarrow S_m$
" $S_m = \bar{x}_m$

P.S: All random sample : $S_1, S_2, \ldots S_m$ taken are independent of each other.

Here,

$\bar{x}_1, \bar{x}_2, \ldots \bar{x}_m$: sample means.

$\bar{x}_i$

All these $\bar{x}_i$ where $i = 1$ to m will have distribution :

FOI

Distribution of $\bar{x}_i$ = Sampling distribution of sample means

We have talked about it so as to understand central limit theorem.

Central Limit Theorem

need not to be gaussian.

X = popn. distribution with finite μ and σ^2.

In case if has infinite μ and σ^2 : pareto distribution

↓

'm' samples of size 'n' :

s_1 s_2 ..... s_m

↓ ↓ ↓

sample means : $\bar{x}_1$ $\bar{x}_2$ ..... $\bar{x}_m$

Here,

distribution of $\bar{x}_i$ = sampling distribution of sample mean.

So,

CLT :-

$\bar{x}_i$ will be a gaussian distribution with μ and $\frac{\sigma^2}{n}$ as $n \to \infty$.

i.e.

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

8166 1088 3103

Conclusion :-

Any distribution [need not to be gaussian]

$X : \mu, \sigma^2$

Let $n = 30$: $S_1, S_2, \ldots S_m$.

(sample size)

Say $m = 1000$

So here, we need to compute $30 \times 1000 = 30K$ people's income

ie:

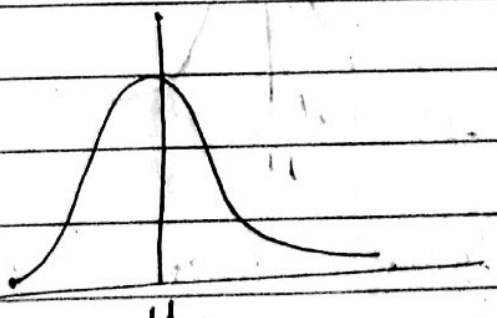
$S_1 \quad S_2, \ldots \quad S_m$

↓ ↓ ↓

Mean income → $\bar{x_1} \quad \bar{x_2} \ldots \bar{x_m}$

in each sample. ↳ $m = 1000$.

If I plot a distribution of it I will get gaussian distribution as



will have μ.

mean of $\bar{x_i} \approx \mu$ (pop. mean)

variation of $\bar{x_i} \approx \frac{\sigma^2}{n}$.

such that $n \to \infty$.

But

Pratically

According to rule of thumb:

We will get gaussian distribution for $n \geq 30$.

where

30 = size of a sample.

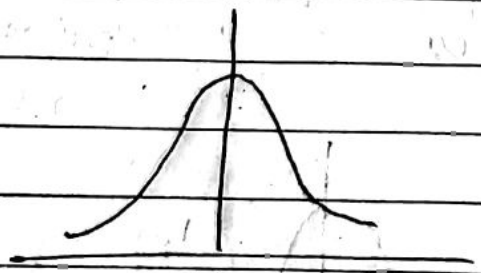
Hence

for any distribution, we can estimate μ and σ^2

↓

just by doing simple sampling distribution of sample mean

And getting distribution as



And

CLT says that sampling distribution of sample means represent gaussian/normal distribution with

μ = popn mean

variance = $\frac{\sigma^2}{n}$ as $n \to \infty$.

OR $n \geq$ sample size.

(11) Quantile - Quantile (Q-Q) plot

Let us consider random variable 'x' with 500 observations.

i.e.

$X: x_1, x_2, x_3, \ldots x_{500}$

Now, question is:

Is 'x' Gaussian distribution?

To find out the answer we have.

'Q Q plot' [graphical method]

Steps for QQ plot:

(1) sort x_i's in ascending order and compute percentile.

i.e.

$x_1, x_2 \ldots x_{500}$

↓ sorting

$x_1', x_2', \ldots x_{500}'$

↓ Percentile computation

$x_5', x_{10}', x_{15}' \ldots x_{500}'$

↓ ↓ ↓ ↓

Representation $x^{(1)}$ $x^{(2)}$ $x^{(3)}$ $\ldots$ $x^{(100)}$

↓ ↓ ↓

1st percentile value of x_i's | 2nd percentile value of x_i' | 100 percentile " " "

② Now, let's create a random variable y.

such that

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

y: y_1 y_2 ..... y_{1000} → 1000 observations from $N(0,1)$.

↓ sort

y_1', y_2', ... y_{1000}'

↓ percentiles

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

③ Plotting of Q-Q plot using.

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

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

Pairing like:

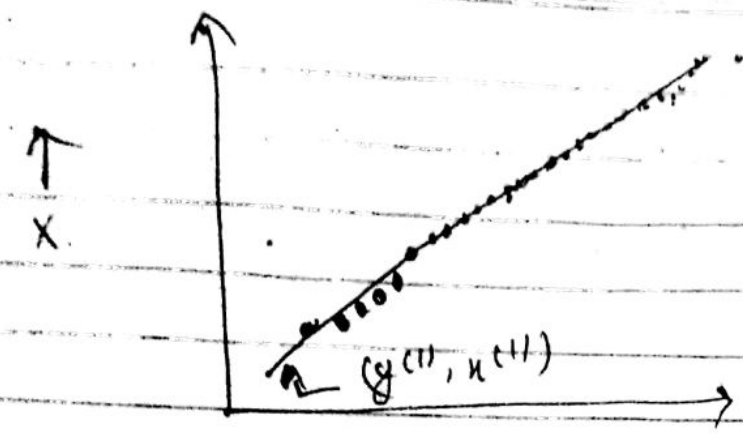
$(y^{(1)}, x^{(1)})$

$(y^{(2)}, x^{(2)})$

$\vdots \quad \vdots$

$(y^{(100)}, x^{(100)})$

QQ plot



containing 100 pair points

$y = N(0,1)$ → (Theoretical quantiles)

Now, now (Q,Q) plot helps to find whether X is gaussian distribution or not:

Here each point correspond to $(y^{(i)}, x^{(i)})$ for $\forall i \to 1$ to 100.

if $(y^{(i)}, x^{(i)})$ $\forall i \to 1$ to 100 lie on a same straight line, then random variable X and Y have similar distribution

[Since 'y' is gaussian dist. hence 'x' is also gaussian distributed since $(y^{(i)}, x^{(i)})$ lies on some straight line.]

Some code.

(1)
```
import numpy as np
import scipy.stats as pylab
std_normal = np.random.normal(loc=0, scale=1, size=1000)
# It gives random observation from normal distribution with mean (0) and variance (1)
std_normal ~ N(0,1)
# Calculating 0 to 100th percentile of std_normal
for i in range(0,101):
    print(i, np.percentile(std_normal, i))
```

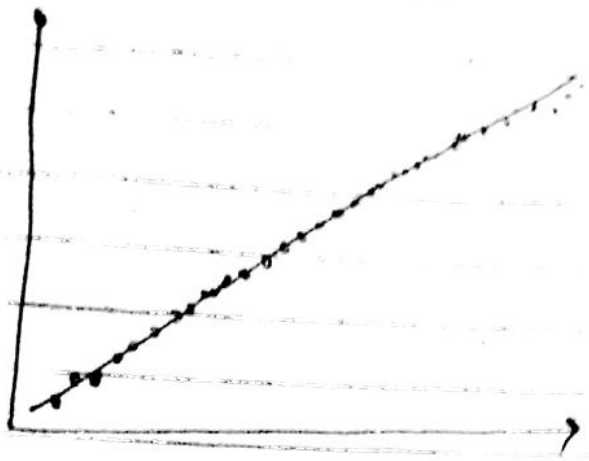
(2)
```
# Some import.
# For QQ plot

# generate 100 samples from N(20,5)
measurements = np.random.normal(loc=20, scale=5, size=100)

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

x 4.

o/p :



Y : N(0,1)

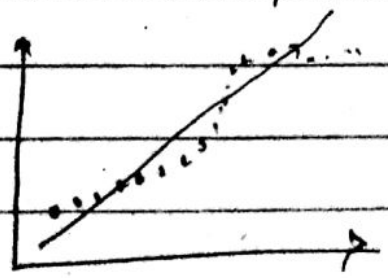
Here, x and y are gaussian (previously said)

As my number of samples increases, , QQ plot is more a condensed straight line.

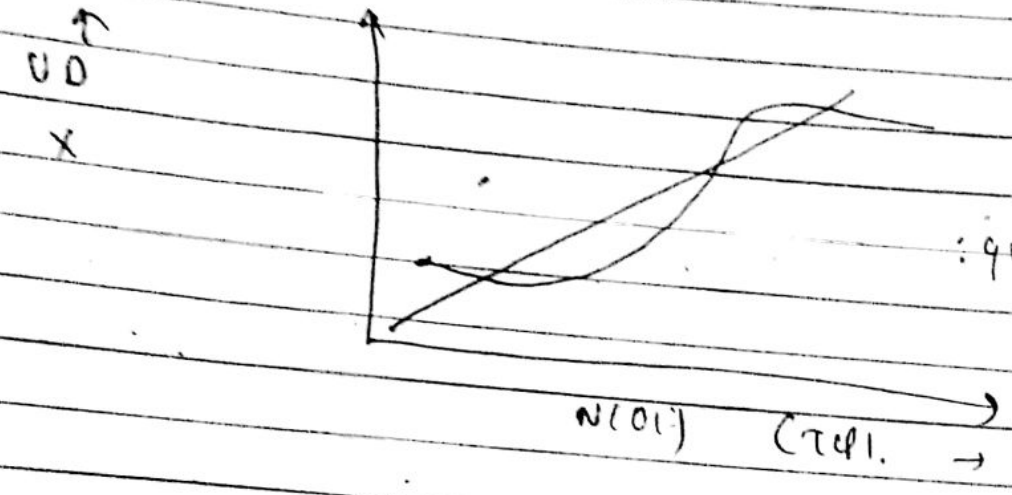
Limitation : If # obs is small, it is hard to interpret QQ plot.

③ # Let us take uniform distribution along y-axis

#generate 100 samples from N(0,1)
measurements = np.random.uniform(low=-1, high=1, size=100)
stats.probplot(measurements, dist="norm", plot=pyl
pylab.show()



As we increase size of uniform dist from 100 to 10000, we get o/p as



hence X and 4 are from different family.

Conclusion:-

(i) Is $X \sim N(4, 6^2)$?

(ii) Given any two random distributions. X & 4

Does X and 4 belong to same family?

↑ broader sense.

(12) How to distribution are used?

Till now we learnt a lot like rv, pdf, gaussian, cdfs, chi sq and soon.

Question is where we can use all these mathematical concepts.

① # Data analysis : Answering questions about data.

Q. company → x42.
task : order T-shirts for all employees (100K)
(S, M, L, XL)
Q. How many XLT-shirts should you order?

✗ Solution 1 : Collect data for all 100K employees. But cost more time & money.

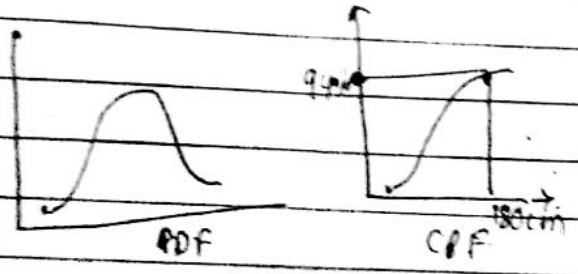
✓ Solution 2 : From domain knowledge, we know
height ≥ 180cm → wear XL T-shirt.
160 ≤ height ≤ 180cm → " L T-shirt
what we will do is:
① collect height of 500 random employees.

(2) Calculate μ and σ of the sample.

(3) From doctor report, we get that height of people is normal distributed.

i.e. height $\sim N(\mu, \sigma)$

Now we can easily compute PDF and CDF.



$\therefore P(h \geq 180cm) = 2\%$

Hence, 2% employee will need XL size tshirt.

Hence, gaussian dist. gives mathematical model that is observed in many natural phenomenon.

q salaries of SVNV(416)

↓

?How we know salaries are gaussian distributed?

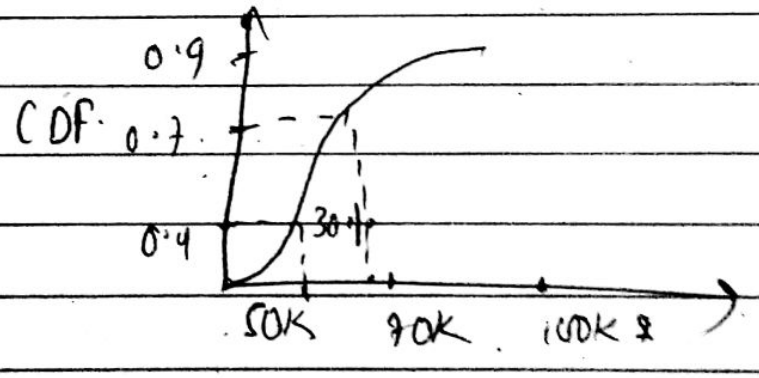
Ans : QQ plot

Then, we can easily predict how many employees make a salary

(i) > 100K $

(ii) 50K $ ≤ salary ≤ 70K $

using CDF



It means 30% have salary between 50K $ and 70K $.

(13) Chebyshev's inequality :

gaussian distribution :

→ $X \sim N(\mu, \sigma)$

→ follow 68% - 95% - 99.7% rule

Assume $\mu = 150$ cm

$\sigma = 10$ cm.

so P (height of students = 95%)

$= P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$

$= P[150 - 20 \leq x \leq 150 + 20)$

$= P[130, 170]$.

Hence, 95% student height lie in [130, 170]

Similarly

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

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

Why chebyshev's inequality ??

In case we Don't know the type of distribution ie whether: gaussian / uniform / ...

But we know:

X : random variable.

↳ Salaries of individuals in a city with.

$\mu = 40k\$$

$\sigma = 10k\$$

Q. What percentage of individual in range of

26. (a) $[20K, 60K\$]$?

36. (b) $[10K, 70K]$?

In case we knew it is gaussian distribution, ans will be:

(a) : 95%.

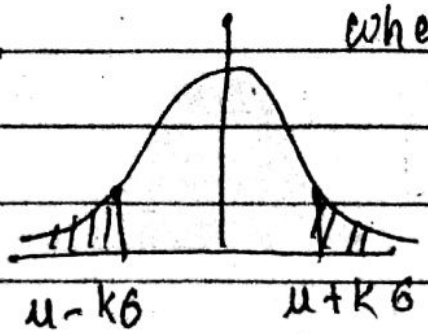
(b) 99.7%.

But, here, we don't know the type of distribution incorporated.

Chebyshev's inequality:

Usage: when distribution is unknown

o/gram



where: μ = finite

σ = non zero and finite.

$$\therefore P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

↑ This is $k\sigma$ range around μ

What % of points lies in ↑

Now answering prev question.

Q salaries :-

$\mu = 40K\$$

$\sigma = 10K\$$

(a) What % of individuals have salary in range of [20K, 60K] ?

$40 - 2\times10$ $\quad$ $40 + 2\times10$

i.e $\mu - 2\sigma$, $\mu + 2\sigma$

$$\therefore P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \geq 1 - \frac{1}{2^2}$$

$$\geq \frac{1-1}{4}$$

$$\geq 3/4 = .75\%$$

$\therefore \geq 75\%$ have salary [20K, 60K].

likewise.

(b) What if ..... [10K, 2017]

Sol: $P(\mu + 3\sigma < X < \mu + 3\sigma)$

$\geq 1 - \frac{1}{9} \quad \geq \frac{8}{9} \quad \% \geq \boxed{\frac{800\%}{9}}$

Answer

Here, chebsev inequality helps us to answer such question knowing not what the distribution is just by knowing its μ and σ.

(14) Discrete and continuous uniform distribution

Uniform distribution types

(i) If random variable (X) is discrete,
↳ Discrete uniform distribution.

(ii) " " " " continuous
↳ continuous " "

(i) Discrete uniform distribution.

For discrete random variable : PMF
(Probability Mass fn)

" continuous " " : PDF
(" density fn)

Here

Let us consider dice throws

outcomes = {1, 2, 3, 4, 5, 6} $p = 1/6$

Parameter used in discrete uniform dist's

a : low

b : high

$n = b - a + 1$

In case of dice throw

$a = 1$

$b = 6$

$n = b - a + 1 = 6 - 1 + 1 = 6$

Basically n = no of outcomes

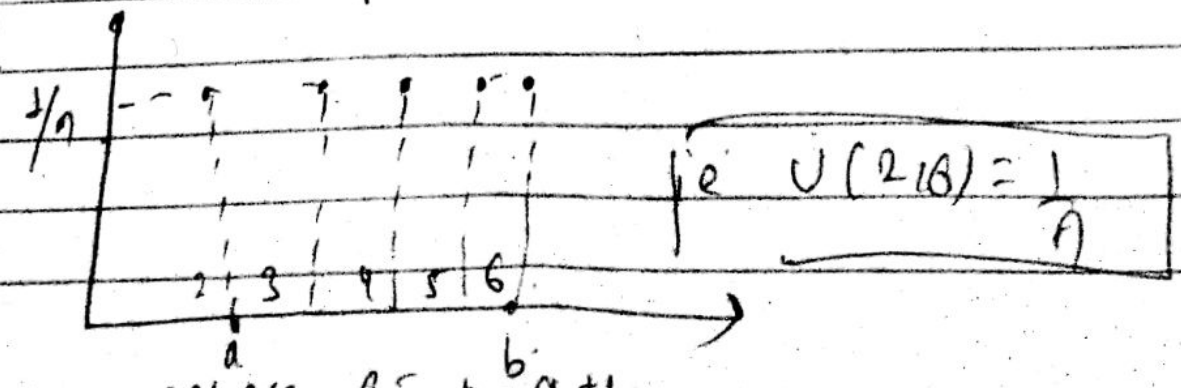
Like

$N(\mu, \sigma)$	→	$U(a, b)$
↑		↑
Gaussian dist		uniform dist

In $U(a, b)$

All cases are equiprobable $= \frac{1}{n}$

Eg.



where $n = b - a + 1$

(ii) Continuous uniform distribution:

$U(a, b)$.

Here, we take in account of points from the interval as well.

ie:

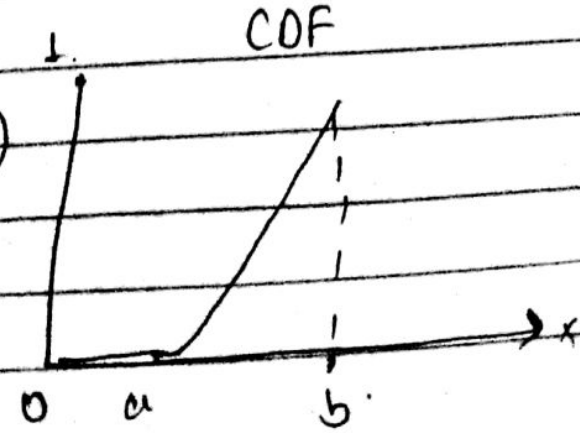
PDF : continuous random variable

$$PDF = \frac{1}{b-a} \text{ for } x \in [a, b]$$

$$mean = \frac{a+b}{2}$$

$$median = \frac{a+b}{2}$$

Skewness = 0.



(1st) How to randomly sample data points.
(uniform() distributions)

we have dataset of 'n' points

↓

D : 'n' (rows)

↓

D : m (sample)

we are trying to uniformly sample 'n' datapoints in 'm' datapoints where all 'n' datapoints selection is equi probable.

- random.random() generates random value (any)

Ex: import random.
print(random.random())

O/P: 0.79.

any random value ranging from 0

• Loading iris dataset with iso point

```
from sklearn import datasets
iris = datasets.load_iris()
d = iris.data
d.shape
```

O/P : (150, 4)

• Now to check internal working of random number generation (uniform distribution)

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

One of the greatest application of uniform distributed random variable

Here p = 0.2
Till random value <= 0.2 we are going to append dataset.
Finaly print length.
So, if result is 32 it means 20% equi chance value for 0 from 0

O/P 0.2
32 / 28 / 29 / 27 (always changing)

(18) Bernoulli and binomial distribution

(1) Bernoulli distribution:

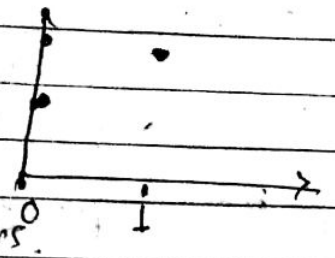
- discrete
- Parameters: $0 < p < 1$.
- pmf: p for $K = 1$, $q = 1-p$ for $K = 0$
- mean $= p$

used on event has 2 outcomes

like: coin toss

1st outcome prob $= p = 0.5$

2nd outcome prob $= q = 1-p = 0.5$



$p + q = 1$.

(2) Binomial distribution

Let us consider coin toss :- where

$X \sim$ Bernoulli ($p = 0.5$)

↓

when 'n' trails are conducted.

Then.

Let Y = no of times I get a head when I toss my fair coin 'n' times. (say $n = 10$)

I may get 10 heads on 10 trails

$\therefore Y = \{0, 1, 2, \ldots, 10\}$

I may get '0' heads on 10 trails

Hence,

$Y \sim Bin(n, p)$

↑ ↑

prob of success on getting head

no of trials performed

This is called binomial distribution.

- Parameters

n = no of trials

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

- PMF (discrete)

$$P(K) = {}^{n}C_{K}\, p^{K} (1-p)^{n-K}$$

(17) Log normal distribution

If a random variable is log normally distributed:

$X \sim$ log-normal (μ, σ)

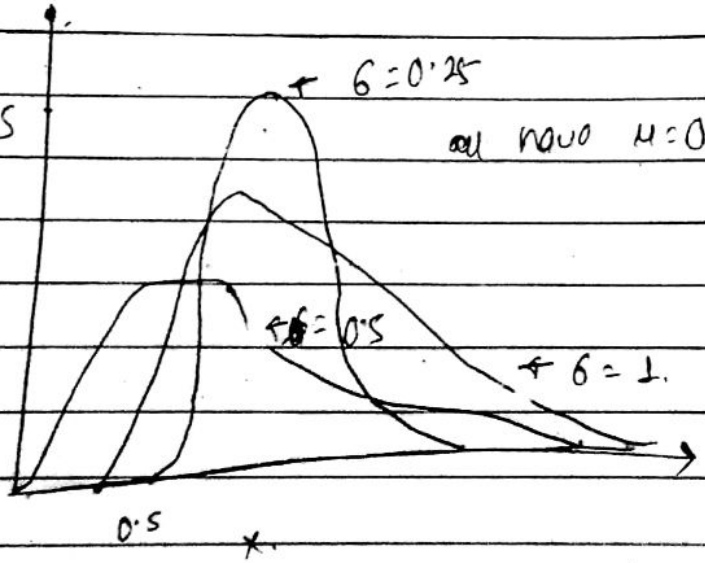
Then, if

$Y = \ln(X)$ has a normal distribution

$\log(n) =$ natural logarithm

Log normal

PDF

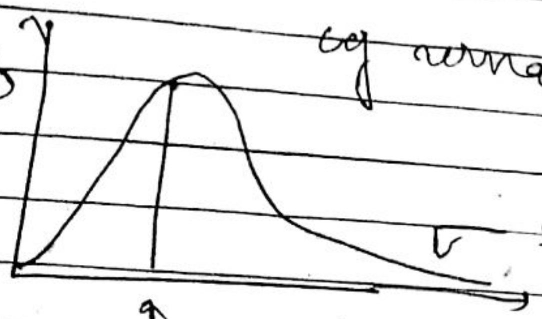


- All are skewed
- And as σ ↑, spread has increased

* Occurrence and application:

Internet company

(i) The length of comments posted in internet discussion like: quora, reddit etc.) follows log normal distribution.



(ii) Users dwell time on online articles follows a log normal distribution.

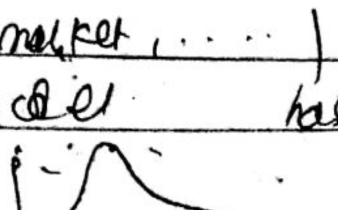
[Alternative: Pareto distribution]

Economics

(i) In economics, there is evidence that income of (97 - 99) % popn is distributed log normally. The distribution of higher income individuals follows a Pareto distribution.

(ii) In finance (stock market, ...) follows Black Scholes model has ~~log~~ heavy tail.

Eg normal dist.



* X ~ log-normal (μ, σ)

Q. Is X ~ log normal?

To find ans.

X: x_1 x_2 ... x_n

↓ ↓ ↓

$\ln(x_1)$ $\ln(x_2)$ $\ln(x_n)$

↓ ↓ ↓

y_1 y_2 ... y_n

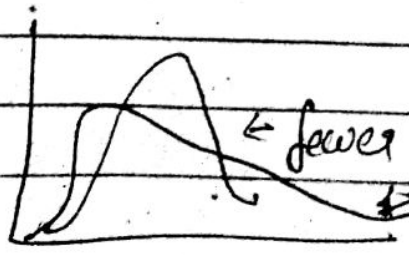
Check if y_i's are gaussian using QQ plot

↓

If yes

X is log normal. (μ, σ).

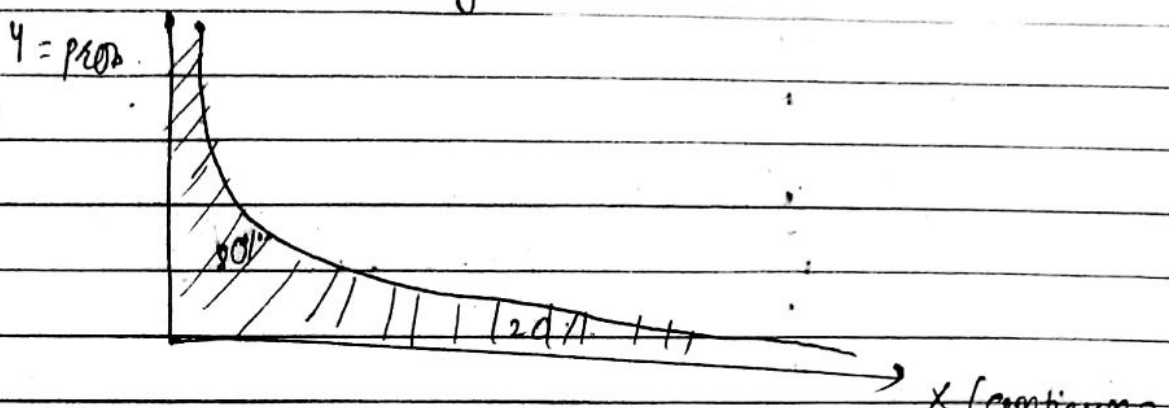
In most internet companies, we will get log normal distribution with skewed tail



(18) Power law distribution : (Pareto distribution)

Before discussing on pareto distribution, let's focus on power law :-

Eg of power law graph



To the right: long tail. and to the left are the few that dominate 80-20 rule.

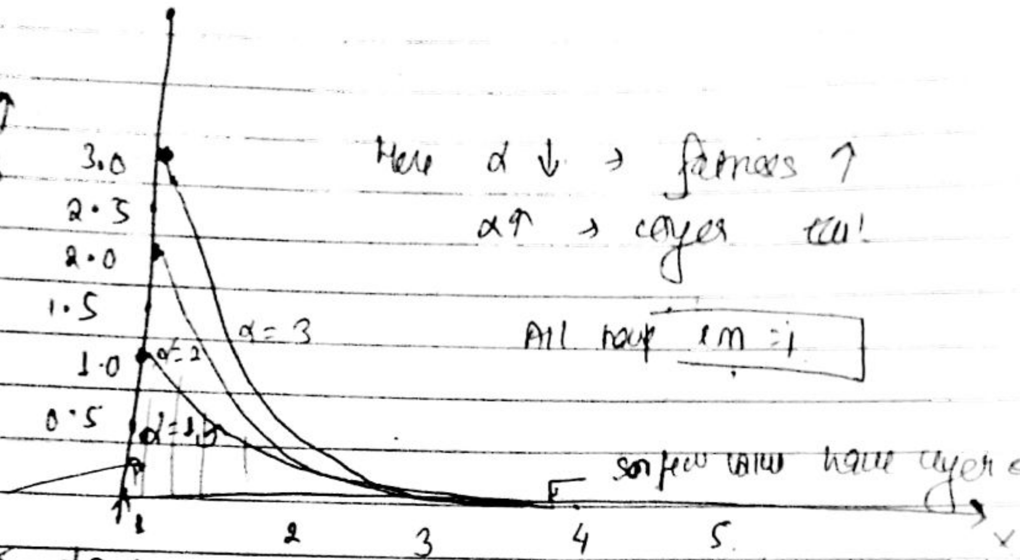
↓

: Pareto distribution :

Parameters : $x_m > 0$: scale. = similar to μ in gaussian distribution

$\alpha > 0$: shape = similar to σ

Example PDF



Here α ↓ → fatness ↑

α↑ → larger tail

All have $x_m = 1$

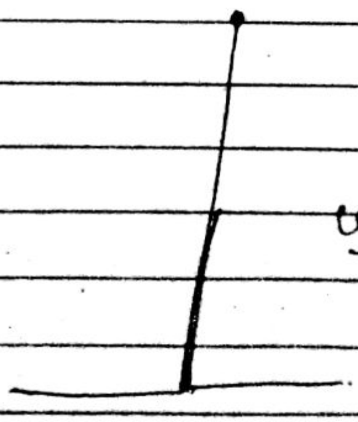
so few values have larger range

Shape ... α = ∞.

As α → ∞, distribution approaches $\delta(x - x_m)$ where δ is known as

Dirac delta function

ie α → ∞

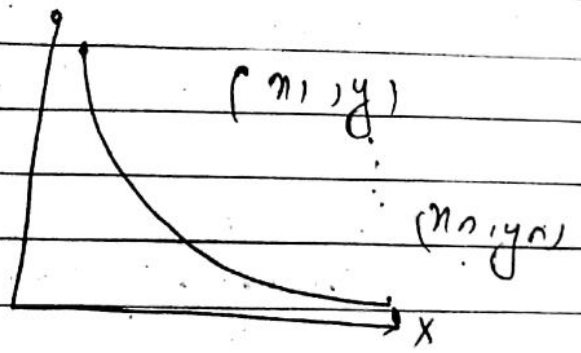


like 'δ' just a peak value at a given value.

* Application.

* To check for power law?

we have data

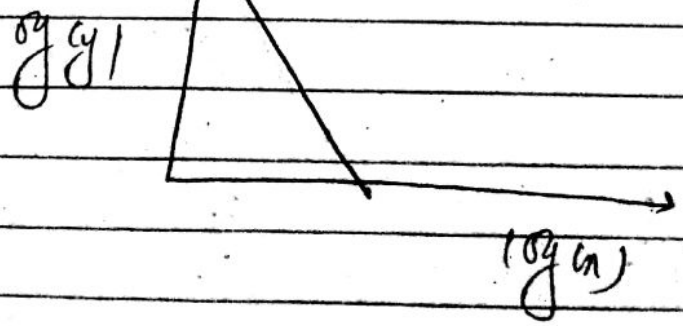


for power law check y

$x_1 \rightarrow \log(x_1)$

$y_1 \rightarrow \log(y_1)$

make log-log plot (if we get a straight line like this, then this follows power law)



To check for pareto distribution

- X → obs
- Y → pareto distribution.

If X, Y give straight line (QQ plot)
then, X : pareto dist.

Power transform

(1) Box cox transform.

Case 1: $(x_1, x_2, \ldots x_n)$ $\xrightarrow{\ln}$ $(\ln x_1, \ln x_2, \ldots \ln x_n)$ → we can study new since we know a lot on Gaussian dist

X (exponential) → Y (gaussian)

✗ ML (assume that a feature is gaussian dist) :

Case II: X (pareto dist) → ? (How we do it.) → Y (Gaussian dist)

powerlaw

Here comes:

Power Transform

↓

We use Box Cox transform under it.

Pareto ~ X: $[x_1, x_2 \dots x_n]$

↓

we need to transform it to gaussian

to Gaussian y $[y_1, y_2 \dots y_n]$

Tot we have to y_i's from x_i's

Step 1: Apply box-cox (x) function

$[x_1, x_2 \dots x_n]$

Generates lambda [d]

↓

How it generates ? No need to understand.

Step 2: Now $y_i = \begin{cases} \frac{x_i^d - 1}{d} & \text{if } d \neq 0 \\ \ln(x_i) & \text{if } d = 0 \end{cases}$ $\forall i = 1 \text{ to } n$

i.e.:

In case if $d = 0$

$\Lambda \sim$ log-normal

because $y_i = \ln(x_i)$ if $d = 0$

* Look at also.

scipy.stats.boxcox

boxcox (x, lmbda = None, alpha = None)

where:

x: nd. array.

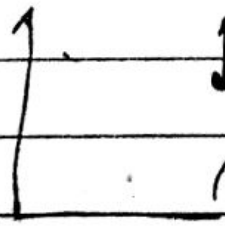
Returns

boxcox : nd array

i.e. box-cox gives transformed array.

i.e.

y = box cox (x)



automatically returns gaussian distribution

Let's look at one case.

```
fig = plt.figure()
ax1 = fig.add_subplot(211)
from scipy import stats
import matplotlib.pyplot as plt
X = stats.loggamma.rvs(5, size=500) + 5
```

Scenario ↓

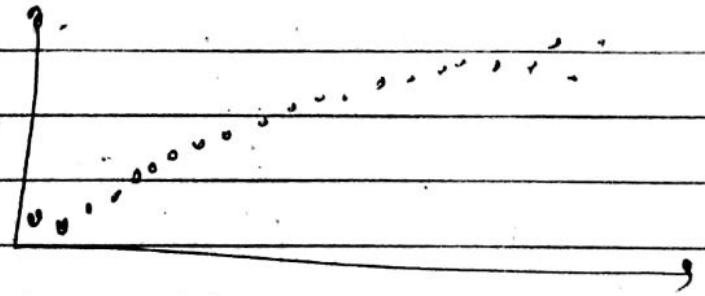
```
# X ~ loggamma (power law/pareto dist)
# rarely used.
```

QQ plot

```
prob = stats.probplot(X, dist=stats.norm,
plot = ax1)
plt.show()
```

QQ plot ↵

gives



QQ plot

* points deviates

X and Y don't have same distribution.

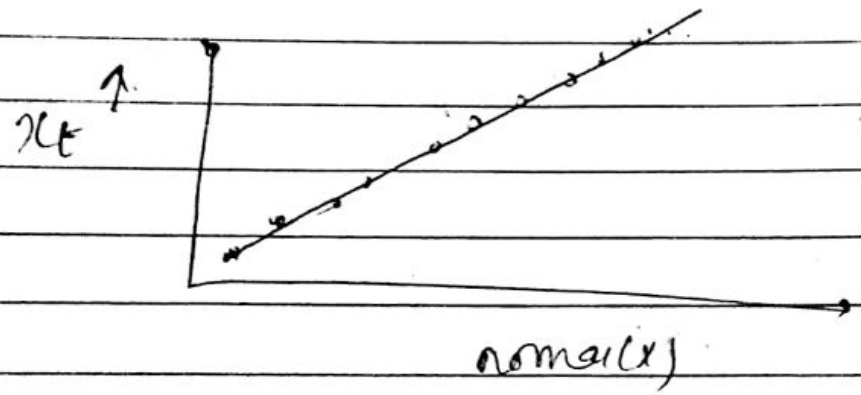
i.e. not gaussian.

b

Best scenario

Box cox.

```
ax2 = fig.add_subplot(212)
xt = stats.boxcox(X)
prob = stats.probplot(xt, dist = stats.norm
                      plot = ax2)
plt.show()
```



BoxCox() best fm to convert pareto dist to gaussian distribution.

(20) Application of non gaussian distributions?

- Dist.
  - Gaussian
  - Non gaussian
    - Uniform → used to generate random numbers
    - Bernoulli, binomial
    - log-normal, pareto

There are 100's of distributions

↓

Q. Why?

If we have a random variable (well studied gives us a X. v. D dist) → theoritical model

↓

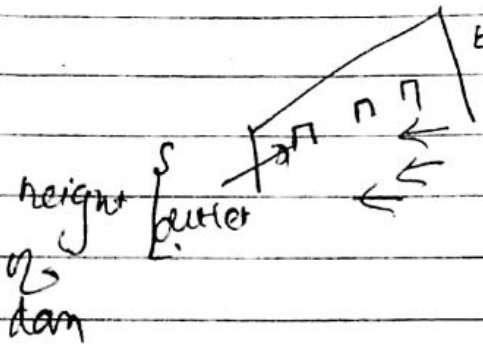
Many with we can ~~have~~ have a lot of inferences / analysis / understanding.

12:40

558 Appeal no. 4.4 AB → 0.0

* let us talk about 'weibull distribution' non gaussian

personal story



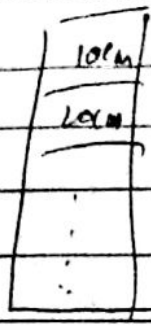
To mitigate accidents, CE have to determine max. one day rainfall @ for dam height and no. of outlet supposition

let X = max. one day rainfall

Now P(X ≥ 20cm) = ?

In traditional time (pre computer era)

Let say, CE cooks few 200. data points



$P(X \geq 20cm)$

↓ percentile ↓ distribution

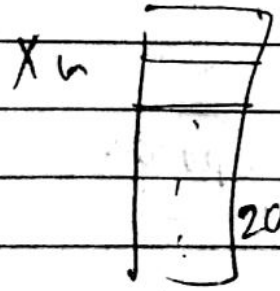
* Percentile :

Let say only 1 obs meet $P(X \geq 20cm)$

$\therefore P(X \geq 20cm) = \frac{1}{200} = 0.5\%$

(not trustworthy (can't trust)

* Theoritical model



→ fit a known and well studied dist / theoritical model

↓

Let say it fit weibull distribution (originally used for particle physics)

* Now X = max-daily rainfall

(1) → weibull (params)

(2)

(3) pdf and cdf

(4) $P(X \geq 20cm) = 1 - P(x < 20cm)$

↓

from cdf of X can be done

Why theoritical model is more trustworthy?

Because → in model all 200 data points fits the model.

while in percentage only 4 datapoint fit.

So $P(x \geq 20cm)$ is more trustworthy here.

Today

Weibull → Previous physics (originally)

→ Wind Eng (now)

Trend distribution has greater application

[Non gaussian dist appear in lots of applications.]

(21) D Finding relationship between random variables :-

Ex:

X: heights Y: weights

	X = h.	Y = w.
S_1	160	62
S_2	150	54
:	:	:
S_n	140	48

Q relationship betn X & Y.

when X↑, does Y↑ ?

when X↑, does Y↓ ?

↓

Measurement of relationship.

(A) Covariance

(B) Pearson correlation coeff

(C) Spearman rank correlation coeff

(4) Covariance

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)*(y_i - \mu_y)$$

↓

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

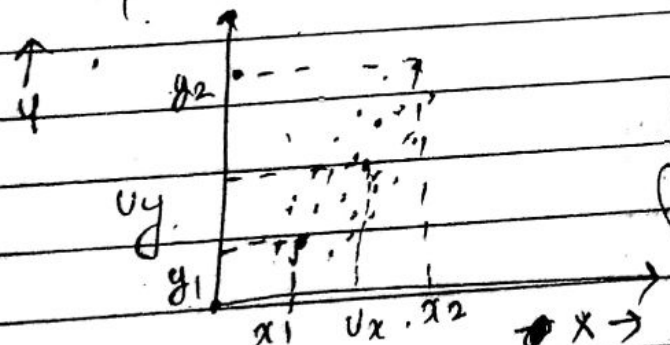
i.e. $cov(x,x) = var(x)$

Note

cov (x,y) = +ve if x↑, y↑.
cov (x,y) = -ve if x↑, y↓.

1st scenario

cov (x,y) = +ve if x↑, y↑



$(x_i - \mu_x) * (y_i - \mu_y)$

(x_1, y_1) : - -

(x_2, y_2) + +

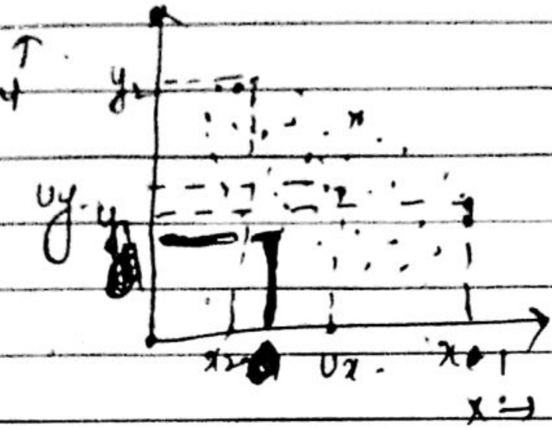
x

cov (x,y) = +ve

Because the points from mean like (x_2, y_2) more positive the covariance

2nd scenario

$cov(x,y) = -ve$ if $x\uparrow$ $y\downarrow$



$(x_i - \mu_x)$	$y_i - \mu_y$	
(x_1,y_1) +	−	(−)
(x_2,y_2) −	+	(−)

$cov(x,y) = -ve \rightarrow x\uparrow, y\downarrow$

Farther the point from mean (μ_x, μ_y), more negative the covariance.

+ (less positive)

+ (more positive)

ie: Tells the relationship b/w x and y

Drawbacks

↓ weight(lb), weight(kg)

cov (x,y) → weight (kilo) ≠ cov(x,y) ↑ lbs.

↓ height (cm)

ie: Just by changing the metrics of measurement of same set of data, covariance may not match.

① Pearson correlation coefficient: (ρ)

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

where $\sigma_x = \sqrt{\text{var}(x)}$

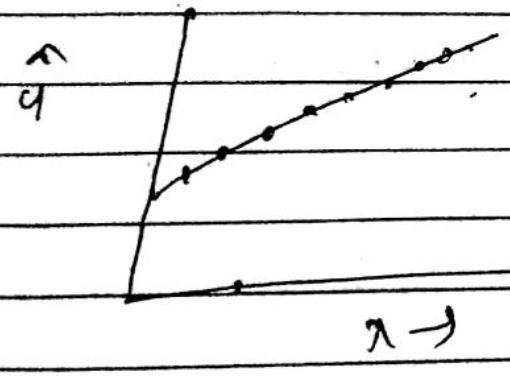
- cov(x, y)

$x \uparrow, y \uparrow$, = +ve. } Let covariance

$x \uparrow, y \downarrow$, = −ve }

correlation coefficient shows how much +ve or how much −ve.

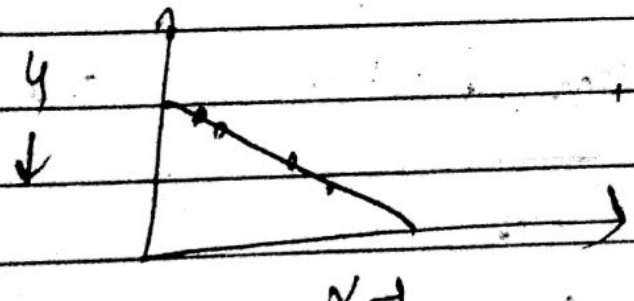
Graphs



All points lies on straight line.

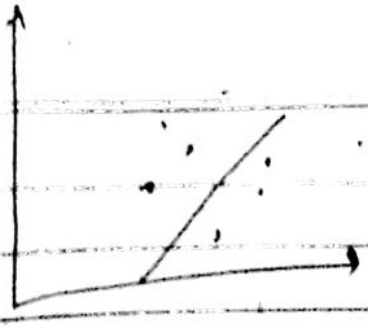
$\rho = +1$

(since $x\uparrow, y\uparrow$)



$\rho = -1$.

Since ($x\uparrow, y\downarrow$)



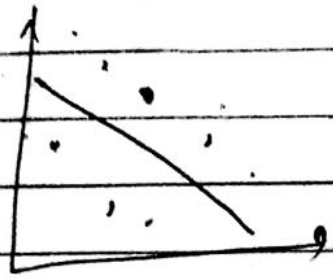
$0 < \rho < +1$.

(~~All~~ ~~point~~ No linear relationship.

$X \rightarrow$ bcm x and y) so,

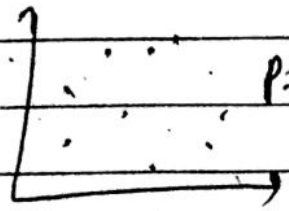
And.

$X \uparrow, Y \uparrow$.



$-1 < \rho < 0$.

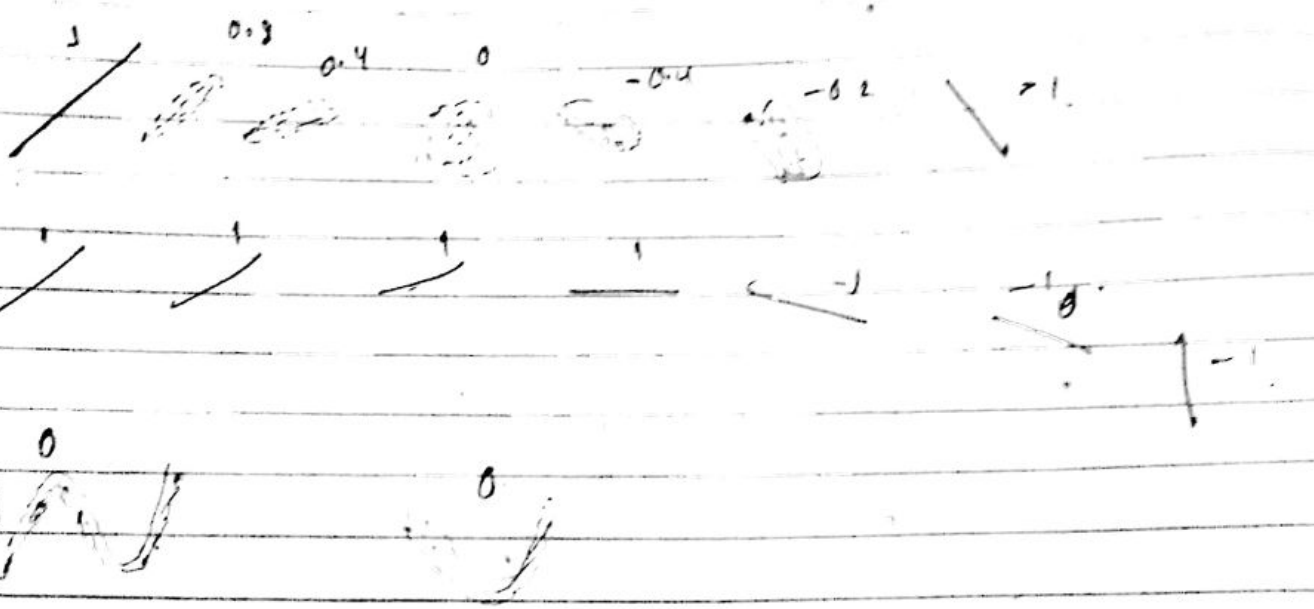
$X \uparrow Y \downarrow$.



$\rho = 0$ No relationship at- all.

{ Hence pcc only perform well in terms of linear relationship bcm x and y. Else not. }

Some examples values of P.



(y = sin(x) + 0.1v)

ie. PCC only useful in case of linear relationship between x and y.

So we introduce our next concept.

(C) Spearman Rank correlation coeff (r)

↳

here, we calculate ranks

Sl sg	X	Y	r_x	r_y
S_1	160	52	4	3
S_2	150	78	2	4
S_3	170	68	5	5
S_4	140	46	1	1
S_5	158	51	3	2

~~Rank~~ r_x calculated as

$r_x = 1$ (smallest x)

$r_x = 2$ (bigger than r_x value)

and so on.

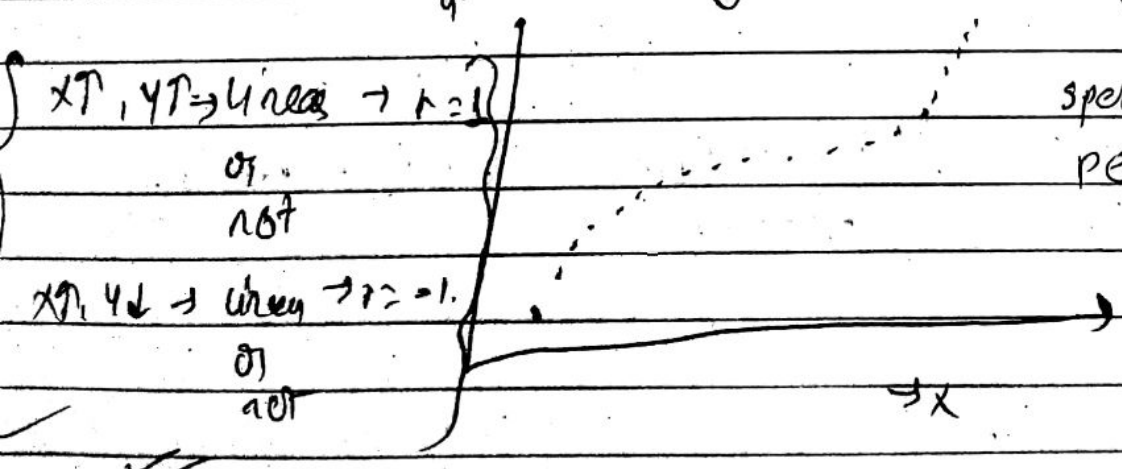
Similarly r_y [assigned $r_y = 1 \ldots$] on the basis of ascending order of y.

Now,

$$r = Pcc(r_x, r_y)$$

let see what happens when we apply

Consider strictly non decreasing graph



spearman (r) = 1

pearson (P_{cc}) = 0.88

io: Regardless of linear or not

r = 1 if [x1, y1]

r = -1 [x1, y1]

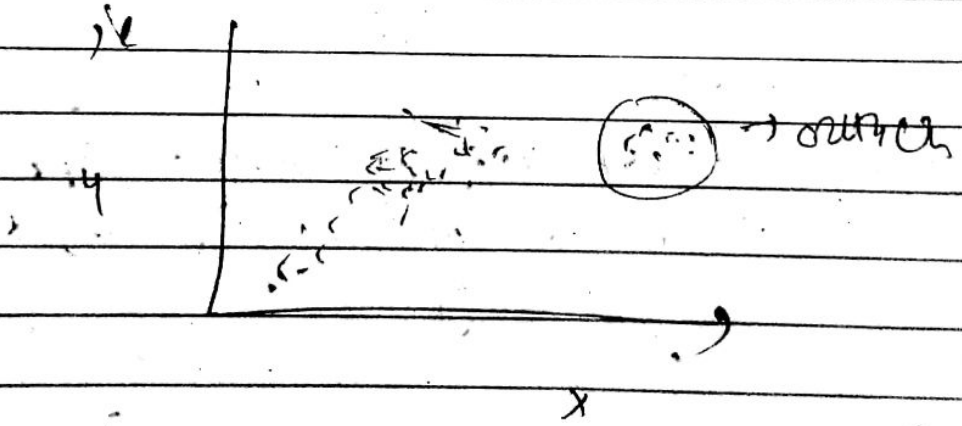
↓

But

Pcc always is exact..

* In case of outlier

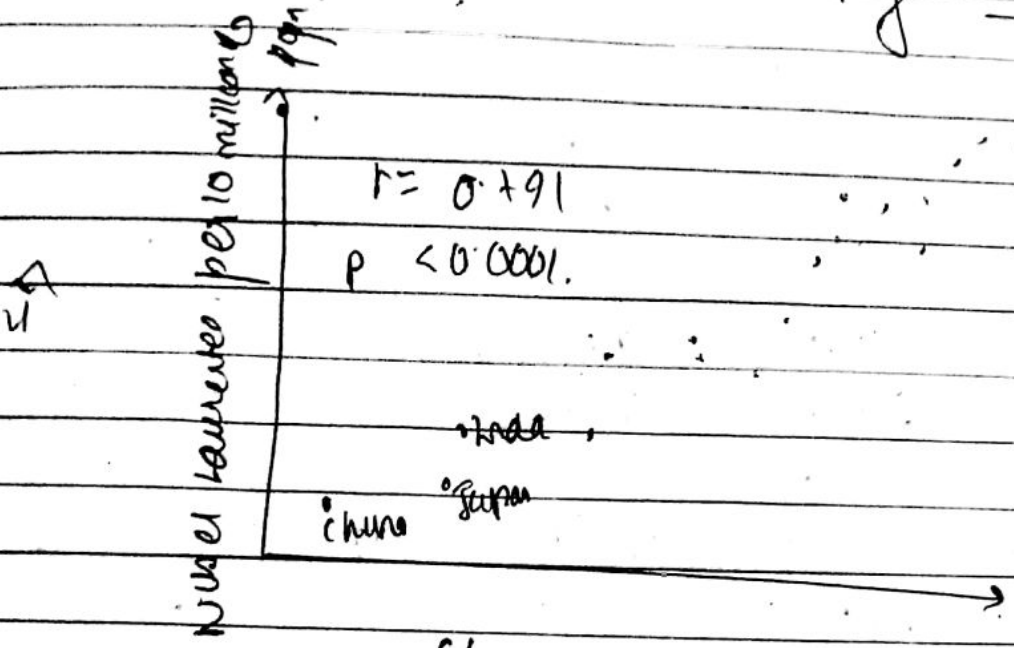
r' is more robust than 'Pcc'



* r = 0.84 → more robust.

Pcc = 0.67

(22) "Correlation" doesn't imply "causation"



x↑ → y↑ → good correlation

r = 0.791
(Spearman rank)

But x causes y / y causes x
i.e. more chocolate consumption produces more Nobel laureates. That's absurd

i.e. if x and y are not necessarily correlated, x causes y or y causes x.

(23) Applications of correlation.

pcc srcc.

[just correlation; not causation]

(1) Is salary correlated with sq. footage of your home?

ie if salary ↑ → sq footage ↑

So, Brokerage will show house according to salary.

That doesn't mean salary causes sq. feet of home or vice versa.

(2) Is # years of education correlated with income?

x y

if x ↑ → y ↑

If compared in this way, education minority will focus on good education.

(3) Commerce organism

(4) Time spent in 24 hrs : → money spent in 24 hrs

↑ ↑

If ↑ → ↑

(4) # unique visitors in a day : $ scales in a day

↑ ↑

If ↑ → ↑

100 K → $ 1 M

120 K → $ 1.6 M } good for company.

(4) Medicine

dosage of a drug/medicine ↓ reduction in blood sugar

↑ ↓

If ↑ → ↓ → good for medical practioner

(24) Confidence Interval Intro : →

C.C.I)

μ = popn. mean
$\bar{x}$ = sample mean

Let us consider a sample 'x' of size 10 such that:

$x = \{x_1, x_2, \ldots x_{10}\}$ → size of people

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

∴ POINT ESTIMATE of $\mu = \bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$

" " " " $= \bar{x} = 166.5$ cm

Point estimation of μ as $\bar{x}$ is less richer.

Despite saying $\mu \in [162.1, 174.9]$ with 95% prob

interval is more richer

confidence

Called confidence interval (C.C.I)

G:32.

(2) Compute CI given underlying distribution.

Let

height in cm.

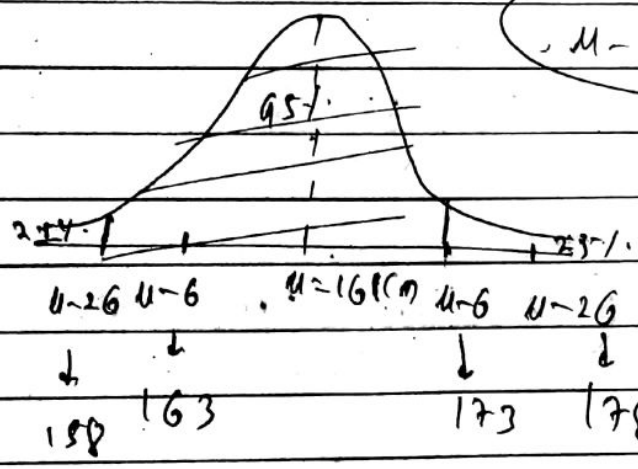
$X \sim N(\mu, \sigma)$

Let

$\mu = 168$ cm

$\sigma = 5$ cm

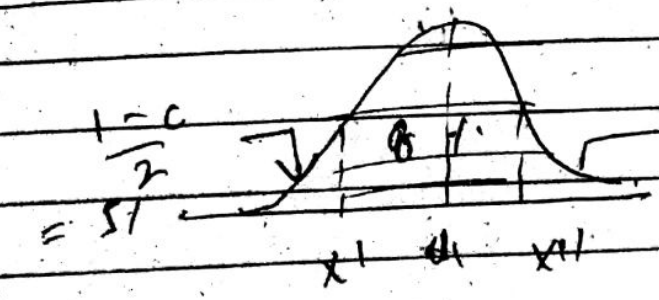
$\mu - 2\sigma, \mu + 2\sigma$: 95%.



∴ 95% of my obs. [158, 178] with 95% prob.

~~95%, 68%, 68.1%~~ are okay.

What about $C = 90\%$.



where x', x'' are find out using normal dist. table.

∴ μ lie in $[x', x'']$ with 90% confidence

lower bound → upper bound

(3) • CI for mean of a R.V

Let X in F. with pop mean μ and std dev of σ.

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

{180, 150, 171, 162 Given the sample
150, 171, 183, 161, 171}

(4) What is 95% CI of μ?

Case 1 σ = Sem (We know pop. std. dev)

As per CLT

$\bar{x}$ = sample mean = $\frac{1}{10}\sum_{i=1}^{10} x_i$ = 168.7 cm

So, for CLT

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

$\bar{X}$ → sample mean; μ → pop. mean; $\frac{\sigma}{\sqrt{n}}$ → pop. std dev / sample std dev

Gaussian dist

$\therefore \mu \in \left[\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right]$ $\{\mu - 2\sigma, \mu + 2\sigma\}$

at 95% confidence

∴ using sample of size 10

∴ pop mean (μ) $\in$ [167.30, 171.60] with 95% confidence.

Case I: If we don't know σ (pop. std. dev)

To find CI of μ

↓

Consider sample of size 'n'

In this case, we use t-distribution also called Student's t-dist

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

↑ sample mean; ↑ t-dist; ↑ degrees of freedom

- Student's t-dist was created to study CI of mean of RV. when 'σ' is not given.

Conclusion

~~Estimate CI~~

Estimate CI of μ of a r.v.

- Case-I: σ is known, → CLT → $N(\mu, \frac{\sigma}{\sqrt{n}})$
- Case II: σ is unknown → t dist (n-1)

Estimate CI of σ of a r.v.

" " " median " "

" " " 90th percentile " "

↓

using

bootstrap CI → modern computers

9:3 11:18 3:18

(4) Confidence interval (CI) using bootstrapping

Let

$X \sim F$

↑

any distribution

Let S^1 = sample of size $(n) = \{x_1 \ldots \ldots x_n\}$ ← discrete

↓ processed by uniform random number generator $U(1,n)$

↓

random sample of size 'm' generated from 'S' using $U(1,n)$

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

such that $m \leq n$

may be sampling with repetition

Hence

Conclusion

~~Estimate CI~~

estimate CI of μ of a r.v.

- Case-I: σ is known → CI → $N(\mu, \frac{\sigma}{\sqrt{n}})$
- Case-II: σ is unknown → t dist (n-1)

Estimate CI of σ of a r.v.

" " " median " "

" " " 90th percentile " "

↓

using

bootstrap CI → modern computers

(9:3) (11:18) (2:18)

(4) Confidence Interval (CI) using bootstrapping

let

$X \sim F$

↑

any distribution

let S^1 = sample of size $(n) = n$

$\{ x_1 \ldots \ldots x_n \}$

↑ discrete

↓ processed by uniform random number generator $U(1,n)$

↓

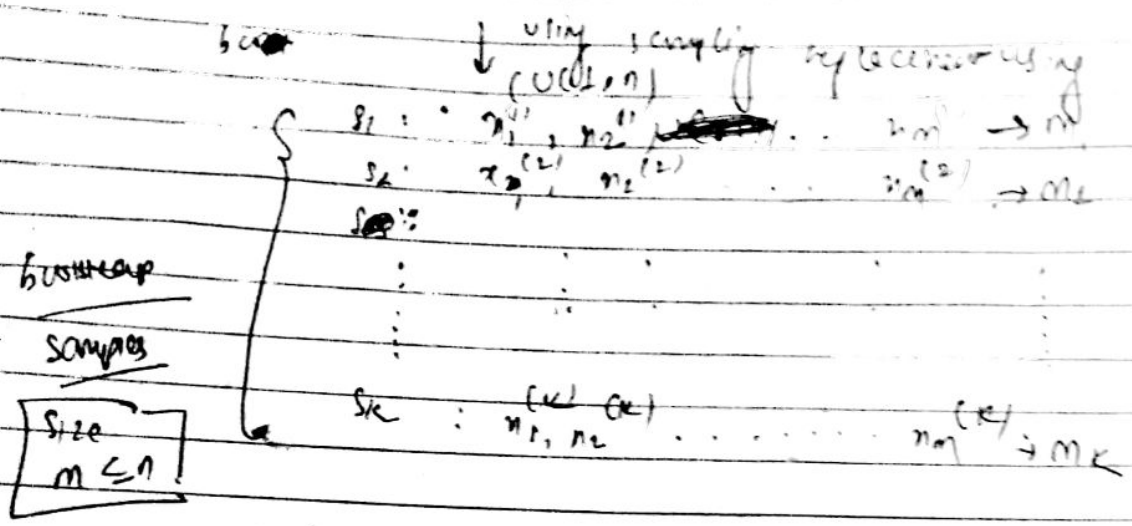
random sample of size 'n' generated from S using $U(1,n)$

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

may be sampling with repetition

Hence.

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



where

m_1 = median of sample 1

m_2 = " " " 2

m_k = " " " k

Now,

Let two observations from median using bootstrap samples

$m_1, m_2, \ldots, m_{100}$

↓ sort

$m'_1 \leq m'_2 \leq m'_3 \ldots \leq m'_{100}$ (increasing)

25 ≤ m'_{25} ↓ CI (95%) ↓ m'_{975} ← 975 observations

15:47

∴ 95% CI of mean of X is

$\{x'_{2.5}, x'_{97.5}\}$

for 6 col/variable, just calculate mean from bootstrap sample and find the CI from sorted one.

Hence, Bootstrap technique is non parametric technique where we don't make any assumption about dist of data.

Code

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

#load dataset
x = numpy.array([135, 162 ..... 176])
# configure bootstrap
n_iterations = 1000          (K)
n_size = int(len(x))  → len of sample (n)
```

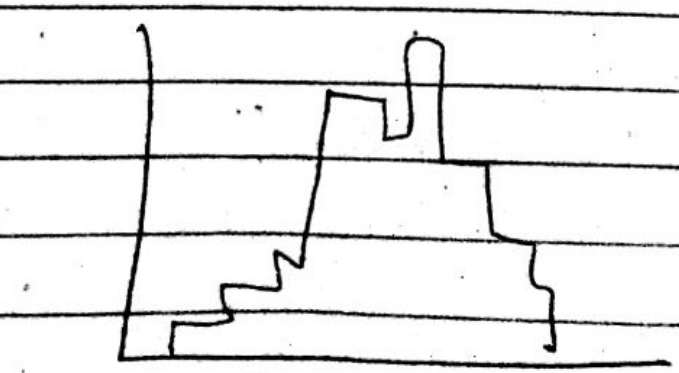
```
# num bootstrap
median = list()
for i in range(n_iterations):
    # prepare train and test sets
    s = resample(x, n_samples=n_size);
    m = numpy.median(s);
    medians.append(m)

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

# confidence interval
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))
```

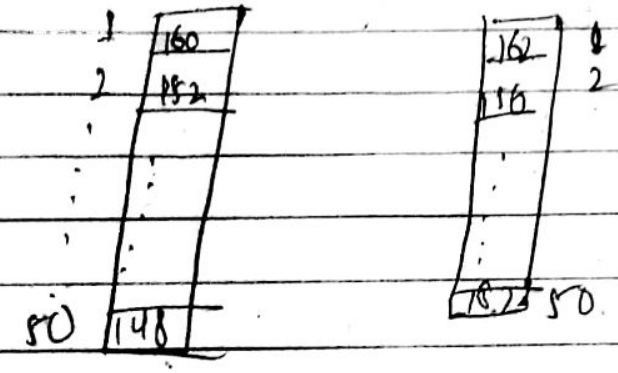


95.0 confidence interval 162.0 and 186.0

(22) **Hypothesis Testing**

(i) **Intro**

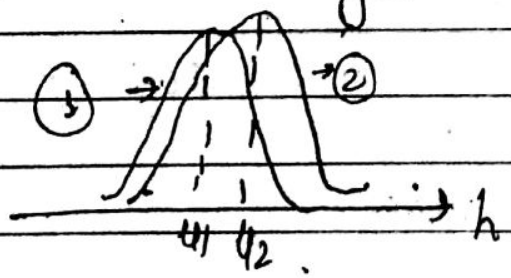
Let us assume heights of 50 students in Cl 1 Cl 2 classrooms 1 & 2.



Q) Is there a diff in heights of students in Cl1 and Cl2?

→

Let us draw histograms



$\mu_2 > \mu_1$: height Cl2 > height Cl1

∴ But μ_2 is slightly greater than μ_1

How we can be so sure that height Cl2 > height Cl1 ???

To accurately estimate such conclusions.

Hypothesis Testing.

Step1 Choosing a test-statistics

μ_2 = mean height of class 12 students
μ_1 = " " " " class 11 "

∴ Test statistics $\bar{x} = (\mu_2 - \mu_1)$

Step2: Null Hypothesis (H_0). {Proof by contradiction}

H_0 = No difference in μ_1 and μ_2.

Alternative hypothesis (H_1):

H_1 = difference in μ_1 and μ_2

Contradiction
Fast step

Step3: p-value:

prob of obs $(\mu_2 - \mu_1)$ provided if null hyp is true.

Assume H0 is true

i.e. $\mu_0 - \mu_1 = 0$

Case-I If p-value = 0.9 $\mu_0 - \mu_1$

Prob of 10cm is 0.9

provided that H0 is true.

It means that assuming $\mu_0 - \mu_1 = 0$,

there is not probability that $\mu_0 - \mu_1 = 10cm$.

Hence, H0 is accepted.

Case-II If p-value = 0.05 It means

Prob of 10cm is 5%.

if H0 is true.

In this case, if H0 = true. (assume)

But since prob of $\mu_0 - \mu_1 = 5\%$. (small)

H0 is rejected.

Instead

H1 is accepted (taken into account)

(11) Hypothesis testing with intuition
coin toss example

Q: Given a coin, determine if coin is biased towards head or not

Basic probability: { biased towards head $P(H) > 0.5$
not " " " $P(H) = 0.5$

(i) design expt:
flip a coin 5 times and count #heads = X. [Test statistics]
[1.0]

(ii) perform expt: f f f f f
↓ ↓ ↓ ↓ ↓
H H H H H

[X = 5 → Observation by expt]

So,
$P(X = 5 \mid \text{coin is not biased towards head}) = f(\text{Obs} \mid H_0)$

Obs
assumption
Null hypothesis (H_0)

H_0: coin is not biased towards head.

$P(X=5 \mid H_0) = \frac{1}{2^5} = \frac{1}{32} \approx 0.03 \approx 3\%$

five heads in five tosses

coin is not biased towards head.

{ f f f f f
$\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$.

$P(H) = 1/2 = 0.5$

$\therefore P(X=5/H_0) = 3\%$

It means, null is 3% chance of getting 5 heads in 5 flips if the coin is not biased towards head

pvalue → $P(\text{obs} / \text{assumption } H_0) = 3\%$

↑ performed

↓ small

According to rule of thumb if $< 5\%$: H_0 not considered

if $P(\text{obs} / H_0) < 5\%$

then

H_0 may be incorrect

↓

assumption is not true

↓

reject H_0 → reject coin is not biased towards head

↓

accept coin is biased towards head.

i.e we accept alternative hypothesis

exp 2 : flip coin 3 times.

count # heads = X ← test statistic.

Perform exp

H H H $\quad X = 3$ observation.

$P(\text{obs} / \text{assumption } H_0) = \frac{1}{2^3} = \frac{1}{8} = 12.5\% < 5\%$.

↳ coin is not biased towards head.

p-value =

Here, we accept H_0

Careful conclusion:

we must be careful in

(i) design of expt

(ii) H_0 → Null hypothesis.

(iii) design ×.

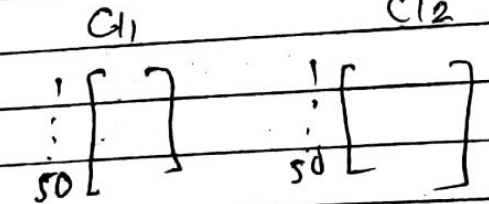
↓

so that

$p(obs|H_0)$ is easy and feasible.

4.10 11:00

(iii) Resampling and permutation testing:



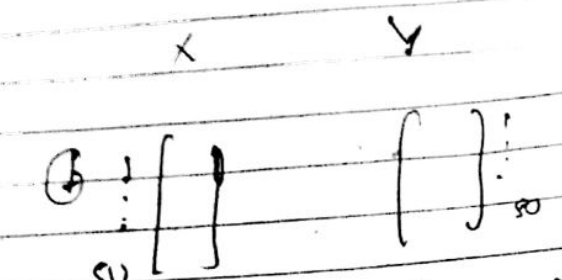
$\mu_1 - \mu_2 = \Delta$

↓ all datas merged into Jumble

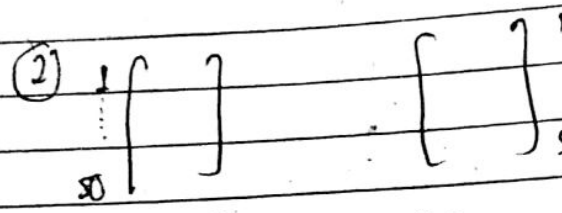
↓

i.e (//)

Now, we randomly sample



$\mu_1 - \mu_2 \rightarrow \delta_1$



$\mu_1 - \mu_2 \rightarrow \delta_2$

(10K) $\mu_1 - \mu_2 \rightarrow \delta_{10K}$

Sort Δ_1 δ's in increasing order

we get.

$\delta_1', \delta_2', \delta_3' \ldots\ldots \Delta \ldots \delta'_{10K}$

n.l.

~~How much n.l.~~

By how much n.l. δ's greater than Δ.

2.00

Area that n% gives p value.

A most ogives

n% = 5%

p = 0.05

Or

p = 0.01 : medical uses

23

(10) KS test for similarity of two distributions

Aim: To find out whether two random variable have same distribution or not.

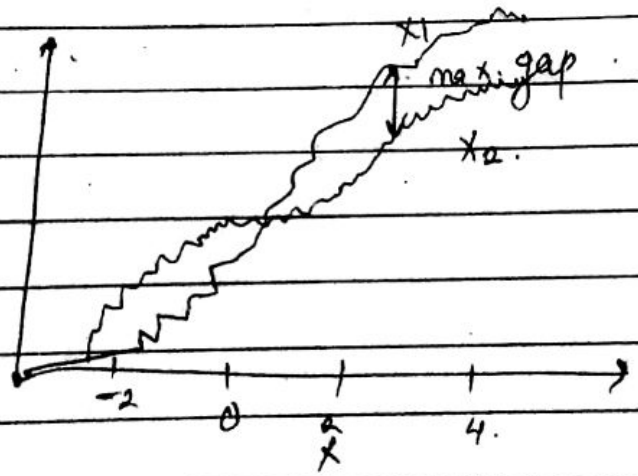
Let X_1 and X_2 are two random variables with sample size 'n' and 'm' respectively.

Hypothesis testing :-

(i) Let the null hypothesis be

H_0: X_1 and X_2 have same distributions

Plotting CDF of X_1 and X_2



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

Where, $F_{1,n}(x)$ = cdf of random variable X_1 with n' sample

$f_{2,m}(x)$ = cdf of random variable X_2 with m sample

$D_{n,m} = \sup |\quad|$

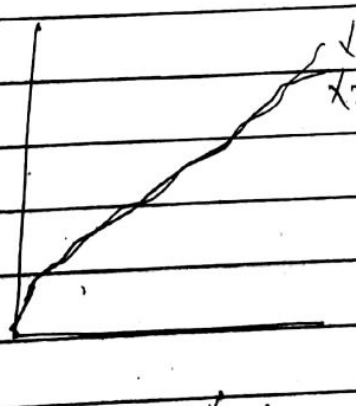
= supremum function representing max. gap bet cdf of X_1 and X_2.

In case

If $n >>>$

$m >>>$

then plot will be



then $D_{n,m} = 0$

ie. X_1 and X_2 same distribution

So. H0 is true.

(iii) Consider α i/p representing significance level

α	0.10	0.05	0.025	0.01
$c(\alpha)$	1.073	1.224	1.358	1.514 .....

where $c(\alpha) \sqrt{-\frac{1}{2}\ln\alpha}$.

(a)

(b) If

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

(Calculated from graph)

then H_0 is rejected.

else accepted.

Eg.

$n = 1000$

$m = 8000$

LOS $\alpha = 0.05$

$c(\alpha) = 1.224$

$\Rightarrow D_{max} > 0.047$ (after calculation)

True.

Hence H_0 is rejected at 0.05 significance level.

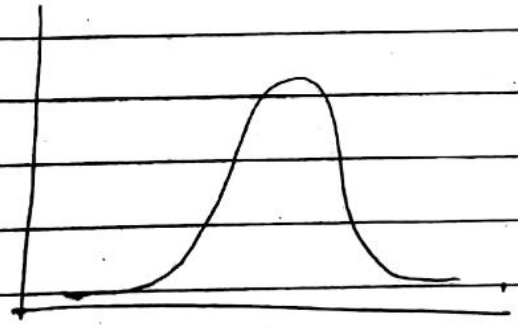
i.e. X_1 and X_2 don't have same distribution.

Code for KS testing

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

```
x = stats.norm.rvs(size=1000)
# generate a normal distribution with some value
sns.set_style('whitegrid')
sns.kdeplot(np.array(x), bw=0.5)
plt.show()
```

Op:



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

Op:

KStest Result (statistic = 0.021.... p-value = 0.75....)

↑ Norm

↑ $c(\alpha)\sqrt{\frac{n+m}{nm}}$

Since $D(n,m) < c(\alpha)$,

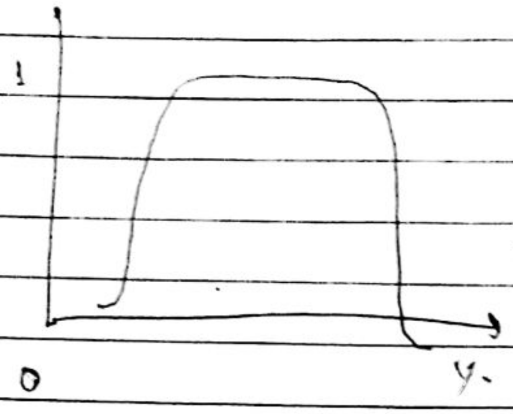
∴ X is a normal distribution

Rule 2: Continuous uniform dist^n (0,1)

y = np.random.uniform(0,1,1000)

sns.kdeplot(np.array(y), bw=0.5)

plt.show()



stats.kstest(y, 'norm')

o/p:

KstestResult(statistic = 0.5....., pvalue = 0.0

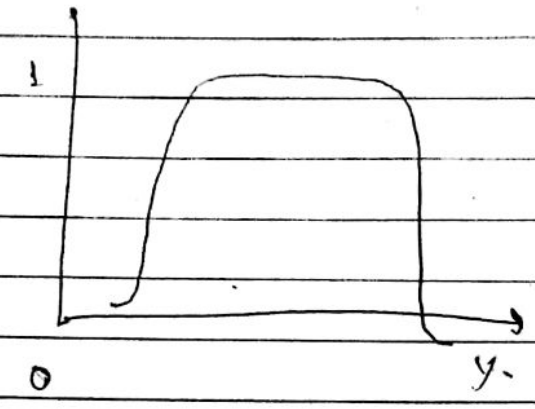
ie. $D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$

ie H0 rejected

ie y is not a normal distribution

Case 2: Continuous uniform dist (0,1)

```
y = np.random.uniform(0,1,1000)
sns.kdeplot(np.array(y), bw=0.5)
plt.show()
```




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

o/p:

KstestResult(statistic=0.5, ..., pvalue=0.0)

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

i.e. H0 rejected

i.e. y is not a normal distribution.

3:41 6:59 9:38 11:58 15:00 7:28

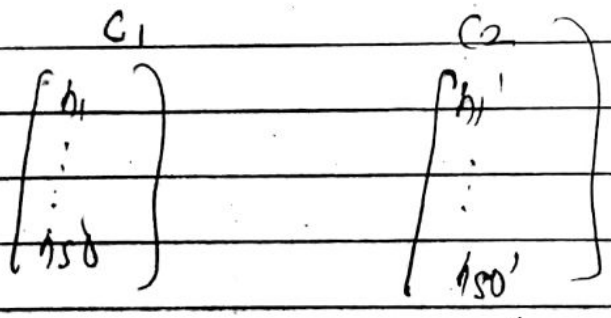
(24) Hypothesis Testing continued.

Another example :-

Q Ask:
Determine if the pop. mean of heights of people in city C_1 and C_2 is same or not.

Step 1: Designing experiment:

Let us randomly take 50 sample from each city as.



sample height of 50 people in C_1 and C_2

Observed value

$\therefore u_1$ = sample height in C_1 = 162cm (let say)
$\therefore u_2$ = " " " C_2 = 167cm (")

Step 2 Test statistic $\bar{x}_2 - \bar{x}_1 = d = 167 - 160 = 5cm$

c_2 c_1

Step 3: Null hypothesis

H_0: There is no difference in population means.

Step 4: Computation

$P(X = 5cm \mid H_0)$

We have to compute it

Means prob. of observing a difference in sample mean height of c_1 and c_2 if there is no pop. diff. in mean height

We will compute next

Let's assume first to make conclusion.

Step 5: Assumption & conclusion.

Case 1: $P(X = 5 \mid H_0) = 0.2 = 20\% > 5\%$

i.e There is 20% chance of observing diff of 5cm in sample mean height of c_1 and c_2 (of size 50) if there is

no popⁿ mean difference.

: P (obs | assumption) = 20% = significant

means: assumption must be true
ie accept H0

Case I . P (x = 5 | H0) = 0.03 = 3% < 5%

p value ie P (obs | assumption) = 3% (very small)

Here,
→ Assumption must be incorrect
→ reject H0
→ Accept H1

Now, for actual computation of p-value

We use something known as resampling

& permutation testing.

Scanned by CamScanner

0:29 2:10 5:47 8:21 11:05 14:36 16:50

24.1. Resampling and Permutation Testing

$P(X = 5/H_0)$ → compute it.

i.e

P (Data / assumption)

What we have

① C_1 C_2

[ ] [ ]

50 sample 50 sample.

② $X = \mu_2 - \mu_1 = 5cm$

Step 1 :

C_1 C_2

jumble.

$S: \{h_1\ h_2 \ldots\ h_{50}, h_1', h_2' \ldots h_{50}'\}$

Step 2 randomly → $S_1 \{h_1, h_1', h_2', h_4 \ldots\} = \mu_1$

Resampling ① S

→ $S_2 \{h_3', h_4', h_2 \ldots\} = \mu_2$

↓

To find $\mu_2 - \mu_1 = 3cm > d_i$

?

Resampling Repeated K times

(1) $\mu_2 - \mu_1 = 3cm = d_1$

Repeat (2) $\mu_2 - \mu_1 = -2cm = d_2$

repeat (3) $\mu_2 - \mu_1 = 1cm = d_3$

⋮ ⋮

(K) $\mu_2 - \mu_1 = 6cm = d_K$

Step 3: Sort d's

$d_1' \leq d_2' \leq d_3' \leq \cdots \leq d_K'$

increasing difference → increasing order

Case 2: Obs-diff $= x = 5cm$

$P(d_{80} \geq 5cm \mid H_0) = \alpha$?

Let K = 1000

Let say

$d_1' \leq d_2' \leq d_3' \leq \cdots \leq 5cm \leq d_{801}' \leq \cdots \leq d_{1000}'$

80% of sim. differences $\leq 5cm$ (obs diff)

20% of sim. diff $\geq 5cm$ (obs. d

P(Obs. dup / assuming H0) = 20% → Significant
10 x 5cm H0 → > 5%

∴ Assumption must be true
∴ Accept H0

Case 6

$d_1' \leq d_2'$ ... $d'_{H0} < 5cm \leq d'_{HA} < d'_{max}$

97% 3%

∴ P(Obs dup > 5cm | H0) = 3%
∴ may ≤ 5%

∴ Assumption (H0) must be incorrect
∴ Reject H0

(25) How to use hypothesis testing ??

(1) KS test used hypothesis testing to check if two random variables have same distribution or not.

Application.

(1) designing drug / medicine.

Comp C_1 — Comp C_2

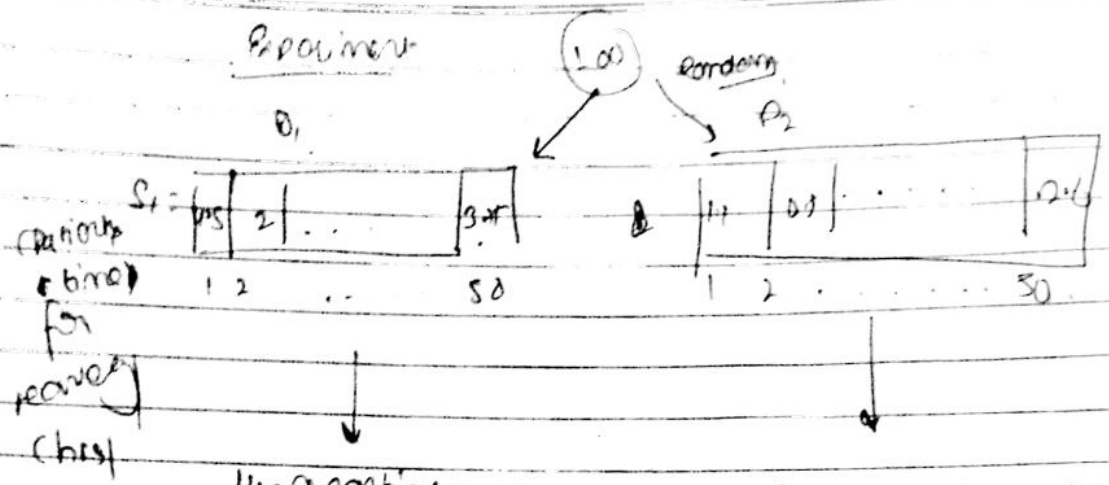
↓

(old) drug D_1 — drug D_2 (new)

↓ ↓

reduces fever in 4 hrs — claim: reduces fever faster than D_1

Task: experiment → to determine if claim is true or not.



μ_1 = mean time = 4 hrs μ_2 = mean time = 2 hrs

To be sure about observation, we apply

Hypothesis testing

(i) H_0 : D_2 and D_1 take same time to reduce fever

(ii) test statistic: $X = \mu_0 - \mu_1 = 2$ hrs (obs)

(iii) $P(X \geq 2 / H_0) = 0.1$ max $= 1\%$ (assume)

↓

H_0 rejected.

means D_2 for faster recovery than D_1

↓ for actual computation

Then resampling and permutation testing

In case of medicine
we must take in account the
significance level (α)
Since it is name of life and death.
If p-val. $< \alpha$
reject H_0
else accept H_0

For medicine
$\alpha = 1\%$ or even 0.1%.
Normally
$\alpha = 5\%$.

(11) Ecommerce
similar web application
like here $\alpha = 3\%$ or 5%.