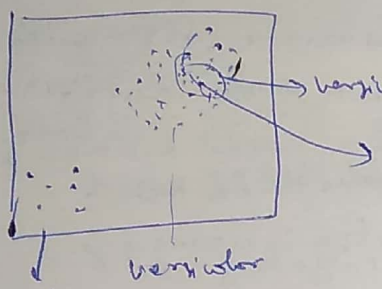
11 Probability and Statistics introduction.

Random variable → two types → discrete → die is Rnum = {1,2,3,4,5,6}
→ continuous → height of student in cla.
↳ suppose (120 cm - 190 cm)
Continue any real number.

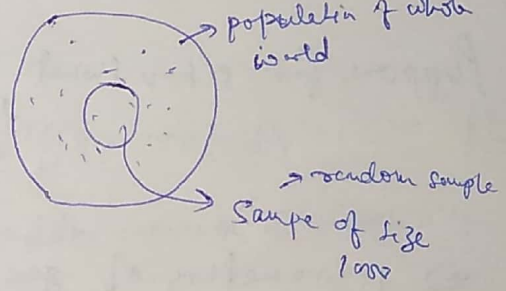
Outlier can also affect the probability.

height → {120.1, 130.2, 146.8, 12.6, 184.8, 90.6}
↳ human error.
↳ short student



There is 80% of chance of versicolor and 20% virginica

Population and Sample.



Like we want to calculate ~~population~~ avg height of people in the world, which is not possible to calculate the mean with normal process.

$$\mu = \frac{1}{7B} * \sum_{i=1}^{7B} h_i$$ (7B → billion)

So to overcome this we take some random sample from the whole population say 1000 random sample [which has 1 in 6 from india]

So $$\bar{h} = \bar{x} = \frac{1}{1000} \times \sum_{i=1}^{1000} h_i$$ → this estimation is more easy and nearly accurate

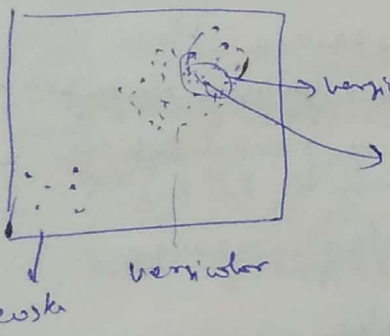
Note: Size of sample increases
↳ $\bar{x} = \mu$
estimation will be equal to population mean of sample

Probability and Statistics introduction.

Random variable → two types → discrete → die is Rnow = {1,2,3,4,5,6}
→ continuous → height of student in class.
↳ suppose (120 cm - 190 cm)
Assume any real realworld.

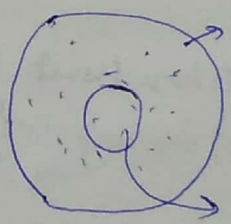
Outliers can also affect the probability.

height → {120.1, 130.2, 146.7, 12.6, 184.8, 90.6}
↳ human error.
↳ short student



there is 80% of chance of versicolor and 20% virginica

Population and Sample.



Here we want to calculate ~~population~~ avg height of people in the world, which is not possible to calculate the mean with normal process.

$$\mu = \frac{1}{7B} * \sum_{i=1}^{7B} h_i$$ (7B → Billion)

So to overcome this we take taken random sample from the whole population say 1000 random sample [which has 1 in 6 from india]

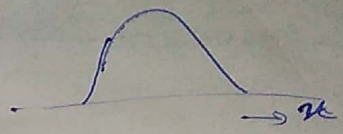
$$\bar{h} = \bar{x} = \frac{1}{1000} * \sum_{i=1}^{1000} h_i$$ → this estimation is more easy and really accurate

When size of sample increases

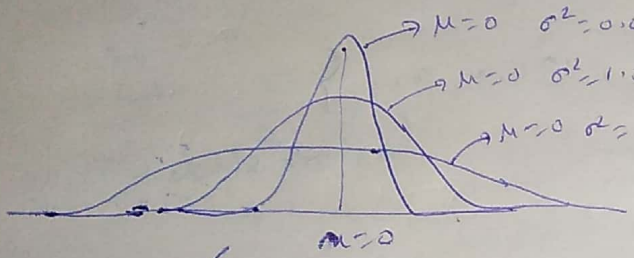
→ $\bar{x} = \mu$ estimation will be equal to population mean of sample

Gaussian and Normal Distribution Function.

if x has probability density function then x has also gaussian distribution function.

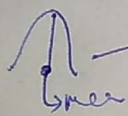


we learn about distribution because this is a simple model and easy to read, it can easily tell us how things are distributed randomly.



⇒ when we have idea of μ → mean and σ² → variance, then with this we can draw gaussian distribution curve.
↳ given normal distribution it is possible to that.

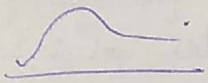
Here in this figure μ=0 → same with increase in σ² → spread is also increases.



→ Generally peak is always at the mean value.

Suppose for setosa flower. we have to make Gaussian distribution

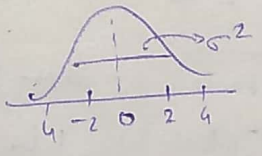
μ=1.5
σ²=0.25



⇒ parameters of gaussian distribution function (μ, σ²)

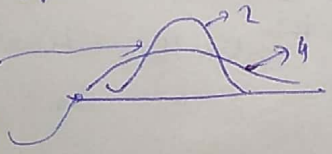
$X \sim N(\mu, \sigma^2)$

X follows Normal distribution given mean and variance



$X \sim N(0,2)$

$X \sim N(0,2)$
$Y \sim N(0,4)$



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

probability density ← P(X=x) ← random variable

Simplify this as.

let $\mu = 0$; $\sigma^2 = 1$ ∴ $\sigma = 1$ → $P(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$

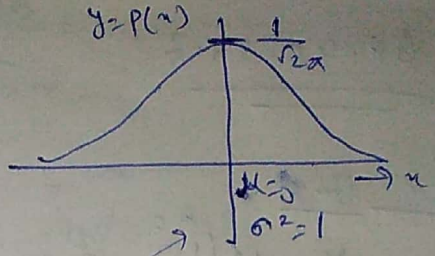
$\frac{1}{\sqrt{2\pi}}$ ← constant ; $\frac{x^2}{2}$ ← constant.

Now, further simplifying it

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

$x \uparrow \quad -(x^2) \downarrow$

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



from this we can see

⇒ As x moves away from μ, $y \downarrow$

⇒ Symmetric

⇒ x move away from μ, y reduces $\exp(-x^2)$

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

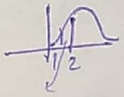
$x=0 \quad y=1$

$x=1 \quad y = \exp(-1) = \frac{1}{e^1} = 0.368$

$x=2 \quad y = \exp(-4) = \frac{1}{e^4} = 0.018$ } 2σ

$x=3 \quad y = \exp(-9) = \frac{1}{e^9} = 0.000123$ } 100σ ↓ drastic decrease in value.

Q/A ⇒ pdf is used to specify the probability of the random variable falling within particular range of value



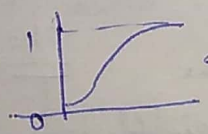
↳ if → P(1.1) = 0.594 from std (1-2)

else P(1.1) = 0.54

⇒ cdf will be continuous because there will not be sudden increase of probability at a single point.

CDF of Gaussian Distribution.

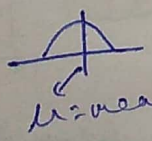
↳ cdf always lie b/w 0 and 1



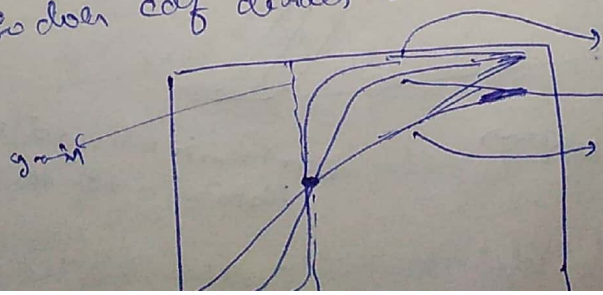
when variance is small → more close to y-axis

when variance is large → far away from y-axis.

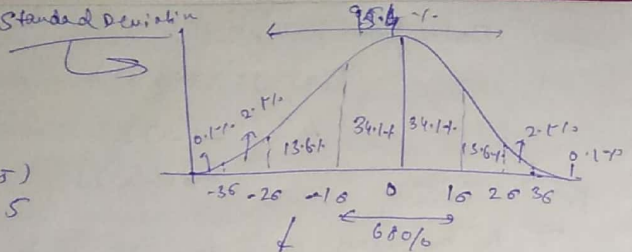
↳ mean divides all data into two parts



so does cdf divides it into two equal percentage (probability)



68-95-99.7



This shows 68% of data lies b/w first Standard deviation and 95.4% of data lies b/w 2nd Std. and 99.7% of data lie b/w 3rd std.

Suppose Given

$X \sim N(\mu = 150, \sigma^2 = 25)$

$\sigma = 5$

from this

$\mu = 150$

$\sigma^2 = 25$

from $(150-25)$ to $(150+25)$

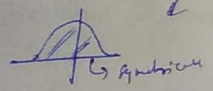
125 to 175 cm

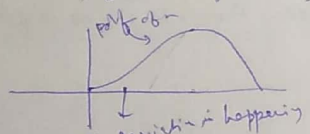
↳ 68% of people live in world with height (125 cm to 175 cm)

Q/A ⇒ Cdf is obtained by cumulative adding of probabilities. So it can't be symmetrical. Pdf usually is skewed.

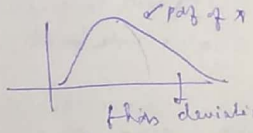
Symmetric distribution, Skewness and kurtosis

↳ deviation from symmetric normal distribution.



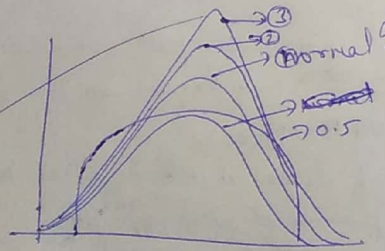


This deviation is happening on left side so, known as -ve skew



this deviation is happening on +ve side ∴ known as +ve skew

Excess kurtosis



distribution curve → kurtosis for normal distribution curve is 0

when kurtosis is ③ there is sharpness at the tip and plot is higher than normal curve.

when kurtosis is 0.5 plot is lower and → sharpness decreases

kurtosis is measured as tailedness of distribution not the flatness.

when kurtosis > 3 → heavily tailed distribution.

normal distribution whose kurtosis = 3 → mesokurtic (known as)

kurtosis > 3 → leptokurtic

kurtosis < 3 → platykurtic

Standardization and Standard normal Variate (z)

① $z \sim N(0,1)$

↳ the standard normal Variate we have mean as 0 and variance as 1.

② Let $X \sim N(\mu, \sigma^2)$

where X has random samples $[x_1, x_2, x_3, \cdots x_{50}]$

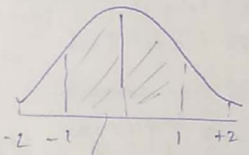
Now, with standardization :-

$$x_i' = \frac{x_i - \mu}{\sigma} \quad \forall i = 1, 2, \cdots 50,$$

↳ this will make every value b/w 0 to 1.

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

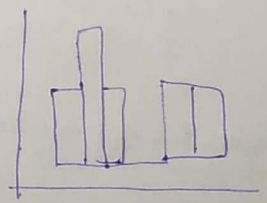
this will be very useful while solving problems later



with standardization we would know that 68% of value lies b/w -1 to 1 and 95% ——— -2 to 2.

Kernel Density Estimation (KDE)

↳ using this we can know how a pdf can be created using histogram



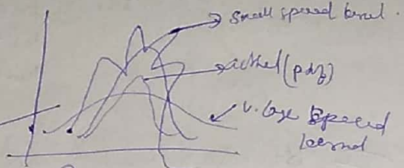
this is based on density function. Here, at each point (or a near) our pdf is created

Suppose this reason is taken where, density of point 0 is the sums of all the kernels of pdf below it.

At point P, mean of higher kernels are present, so it has highest peaks.

Band width selection.

↳ how bandwidth (spread) is affected by variance.



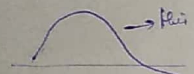
So spread will be more as

if distance b/w the kernels are more than the will be more spread.

if distance b/w the kernels will be small (near set of points in one region) then spread will be ← zigzaged.

Central Limit Theorem (CLT)

X :



↓ distribution of income

Suppose we have the population of

from this population, we are take 30 independent values on n different sample.

(1) → random sample of size n(=30) → s_1 → $\bar{x}_1$ → sample mean

(2) → sample of size n(=30) → s_2 → $\bar{x}_2$

⋮

(n) —— $s_m \rightarrow \bar{x}_n$

So, here, $\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_n$ are n sample mean.

$\bar{x}_i$ ~ distribution of sample mean

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

(CLT) if X has finite μ and σ^2 then $\bar{x}_i = N(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$

↑ Gaussian distribution.

sample of size n: $S_1 \ldots S_2 \ldots S_n$

↓ ↓ ↓

sample mean ← $\bar{x}_1$ $\bar{x}_2$ $\bar{x}_m$

↳ Suppose n = 1000

∴ $\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_m$ → 1000

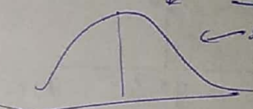
↓ sampling distribution for this mean. large n ≥ 30 → Gaussian.

Let n = 30

total sample data = 30 × 1000 = 30000

μ = mean of $\bar{x}_i$

$\frac{\sigma^2}{n}$ = Var of $\bar{x}_i$



this mean will be similar to population mean.

Q/A ↳ So, CLT state that if you have population of mean μ and std σ^2 and (var → σ) take sufficient large no. of samples from the population with replacement. then distribution of sample mean will be approx normally distributed.

Q-Q plot : How to check if a Random Variable is normally distributed or not.

Suppose we have some random variable X

Step 1 ↳ X : $x_1\ x_2\ x_3 \ldots x_{500}$ → we can check if it is normally distributed or not by two ways → Q-Q-plot → statistical tests (KS, AD)

Now sort X in ascending order first

$x_1\ x_2\ x_3 \ldots x_{500}$

↓ sort (asc)

$x_1', x_2', \ldots\ldots x_{500}'$ → Now compute percentile → $x_5', x_{10}', x_{15}' \ldots x_{500}'$

$x^{(1)}$ $x^{(2)}$ ... $x^{(100)}$ (100 percentile)

Suppose x_5' is 1st percentile, 2nd percentile, 100th percentile

Step 2 Take a standard normal distribution

$Y \sim N(0,1)$

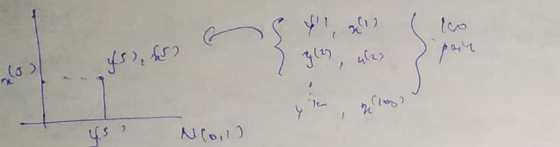
↳ $Y_1, Y_2, Y_3 \ldots Y_{1000}$ → (1000 observations from N(0,1))

↓ sort (ascending)

$Y_1', Y_2', Y_3' \ldots Y_{1000}'$

percentile → $Y^{(1)}, Y^{(2)}, \ldots Y^{(100)}$

Step (3) plot QQ-QQ plot using $x^{(1)}, x^{(2)} \ldots x^{(100)}$ (percentile of) $y^{(1)}, y^{(2)} \ldots y^{(100)}$

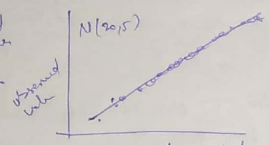


$y^{(1)}, x^{(1)}$
$y^{(2)}, x^{(2)}$
$y^{(100)}, x^{(100)}$
100 pairs

→ Theoretical Quantiles

if they follows a straight line

Then, Random Variable X is also Normally distributed as Y.



Code ⇒ # QQ-QQ-plot

```
import numpy as np
import pylab
import scipy.stats as stats
```

Adding Std. Normal distribution N(0,1)

⇒ std_normal = np.random.normal(loc=0, scale=1, size=1000)
(loc → mean, scale → variance, size → 1000 samples)

0 to forth percentiles of std-normal

```
for i in range(0,101):
    print(i, np.percentile(std_normal, i))
```

O/p
0 −3.76
1 −2.42
⋮
100 2.779

Random variable X ↳ # Generate 100 sample from N(20,5)

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

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

pylab.show()

↓ checks for normal distribution

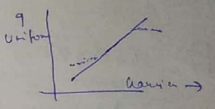
if size is small Then we can see deviation but when we increase the size sample the the plot will be linear.

generate 100 sample for N(20,5)

measurements = np.random.uniform(low=1, high=1, size=100)
↳ try (100

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

pylab.show()



How/where to use distributions?

till now we have learned random variable, pdf, cdf, normal → 68-95-99.7 rule.

Distributions are generally use to ~~the~~ data analysis

Suppose: In an company we have to order tshirts for ~~look~~ employs. range of tshirts can be → XL, L, M, S

So, here we can't just go to every employ to know the tshirt size. this would be time consuming.

So, to do this task in a simple way.

We will take the height of 500 people out of 1000 from our ~~office~~ Then, if height > 180cm → XL

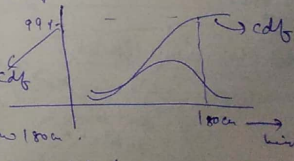
height 160 to 180cm → M ↳ there are general assumptions.

So, now, when we have collected 500 height we will calculate mean and std.

$N(\mu, \sigma)$.

We will use QQ plot N(0,1) to check if both are normally distributed or not..

if yes,



∴ 1% of 1000 = 1000 tshirts are wastly required.

This can be done for Calculating Salaries range also.

Gaussian distribution are theoretical model
↳ this happen is most of natural phenomenon that's why we take it as reference.

Chebyshev's Inequality.

Till now, we studied Gaussian distribution → 68 - 95 - 99.7 rule (1σ, 2σ, 3σ)

Suppose X → Height of Students

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

X is normal distributed having finite mean and variance.

Then we can say,
how much students lies b/w
$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 95\%$ with Gaussian distribution
↳ [130, 170]

But suppose, we don't know the distribution, all we know that we have finite mean and non zero Variance

Can we Calculate
how much
x% of data lies within $\mu - 2\sigma$ & $\mu + 2\sigma$
y% of data lies ____ $\mu - 1\sigma$ & $\mu + 1\sigma$

Suppose we have to Calculate Salaries of individuals and we don't know the distribution and we have (μ) 40k, (σ) 10k

(Q1) what % of individual has salary in the range of [20k, 60k] ? (40 − 10×2, 40 + 2×10)

(Q2) [10k, 70k] ? (40 − 3×10, 40 + 3×10)

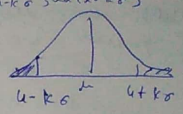
⇒ Chebyshev's inequality.

when we have X → r.v
don't know the distribution
finite mean = μ
non-finite standard-dev = σ

probability that X lies outside region of $(\mu - k\sigma)$ and $(\mu + k\sigma)$

Eg: $P(|X - \mu| \ge k\sigma) \le \frac{1}{k^2}$

$X \ge \mu + k\sigma$
$X \le \mu - k\sigma$



probability that X lies b/w $(\mu - k\sigma)$ and $(\mu + k\sigma)$ is

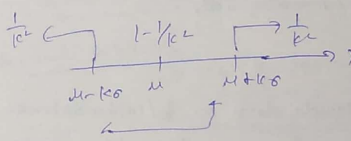
$$P(\mu - k\sigma < X < \mu + k\sigma) \ge 1 - \frac{1}{k^2}$$

(Q1) [20k, 60k]

20k = μ − 2σ ↳ $P(20k < X < 60k) > 1 - \frac{1}{2^2} = \frac{3}{4} = 75\%$

$P(10k < X < 70k) > 1 - \frac{1}{(3)^2} = \frac{8}{9} = 90\%$ approx

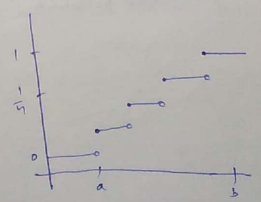
(40 − 3×10, 40 + 3×10)



Uniform Distribution

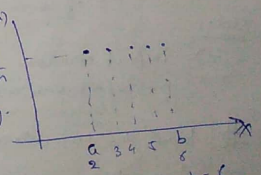
↳ Here are two types of uniform distribution → discrete → discrete random variable
→ uniform → uniform random variable

probability mass function is also called as probability density function.



each point has the probability $\frac{1}{5}$

Here in discrete nothing is defined b/w 2 and 3.



Here n = 5, a = 2, b = 6

$\therefore n = b - a + 1 = 5$

Notation used $U\{a, b\}$

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

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

$n = b - a + 1$

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

median

skewness $= 0$

Prob $\rightarrow \frac{1}{n}$

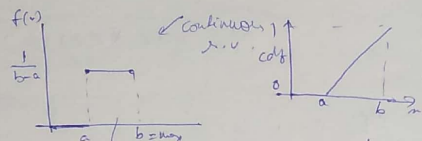
CDF $= \frac{[k] - a + 1}{n}$

↳ Uniform (continuous)

Notation $= U(a, b)$

pdf $= \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{elsewhere} \end{cases}$

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



Continuous r.v

area should be 1 $\rightarrow \frac{1}{b-a} \times b - a = 1$

mean $= \frac{1}{2}$ (ans)

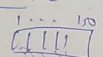
median $= \frac{1}{2}$

Skewness $= 0$

Qn The uniform distribution all outcome is equally likely.

How do randomly sample data point (uniform distribution)

⇒ Suppose we have 150 data points



we have to select 30 data point out of 150

select → $x_1, x_2, x_3 \dots x_{30}$

So here we are doing this sampling in mind that probability of each data should remain same. to fall in Set 1.

Now,

⇒ import random

print(random.random()) this will print value b/w [0 1) with equal probability.

Load the iris dataset with 150 point

⇒ from sklearn import datasets

iris = datasets.load_iris()

d = iris.data

d.shape

↳ o/p → (150, 4)

Sample 30 points randomly from 150 point

⇒ n = 150

n = 30 → $\frac{30}{150} = 0.2$.

p = 30/n

print(p)

sampled_data = [ ];

for i in range(0, n):

a = random.random() this will print probability of occuring of data whose probability is <= p → 0.2.

if a <= p;

sampled_data.append(d[i, :])

len(sampled_data)

o/p ↳ 0.2

37 → it can vary around 30.

Bernoulli & Binomial distribution.

Bernoulli distribution

↳ when we have only two outcome then bernoulli distribution is used.

like tossing a coin (T, H)

if occurrence of one outcome is p

then occurrence of another outcome is $(1 - p = q)$.

Parameter → $0 < p < 1$, $p \in R$

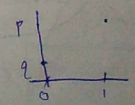
pmf $= \begin{cases} 1-p & \text{for } k = 0 \\ p & k = 1 \end{cases}$

cdf $= \begin{cases} 0 \rightarrow k < 0 \\ 1-p \rightarrow 0 \leq k < 1 \\ 1 \rightarrow k \geq 1 \end{cases}$

mean → p

median $= \begin{cases} 0 & q > p \\ 0.5 & q = p \\ 1 & q < p \end{cases}$

Variance → pq.



Binomial distribution.

↳ Number of time I get a head when I toss a fair coin n times

$Y \sim Bin(n, p)$

notation → n → no of toss; p → prob of getting head ($\frac{1}{2}$)

* pmf → $\binom{n}{k} p^k (1-p)^{n-k}$

mean = np

Variance = $np(1-p)$

Log Normal Distribution

Suppose we have to check if $Y \sim$ log-normally distributed or not.

So in a log normal distribution $Y \to y_1, y_2, y_3 \cdots$

take $\log_e Y \to \log y_1, \log y_2 \cdots \log y_n$

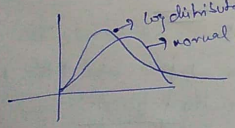
$\to y_1', y_2', \cdots y_n'$

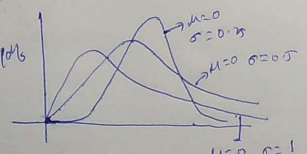
Then this will be normally distributed.

To check this: plot QQ plot b/w y' and $N(0,1)$ →

if this follows linear plot then it is log normally distributed.

In most of the internet companies they work on log-normally distributed value, because very few times they get normally distributed value.



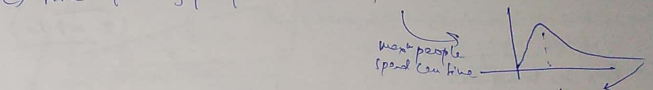


Notation → $\text{lognormal}(\mu, \sigma^2)$

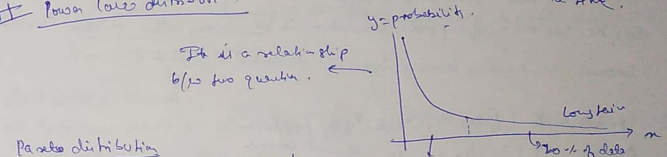
pdf → $\frac{1}{x\sigma\sqrt{2\pi}} e^{-\left[\frac{(\ln x - \mu)^2}{2\sigma}\right]}$

↳ Comments on internet discussion forum follows a log normal distribution.

↳ time spend by people in internet follows log normal distribution.

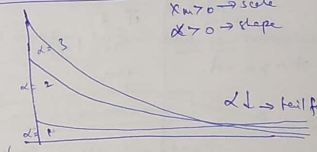


Power laws distribution



It is a relationship b/w two quantities.

Pareto distribution



$x_m > 0 \to$ scale

$\alpha > 0 \to$ shape

$\alpha \downarrow \to$ tail features ?

when $\alpha \to \infty$ then → a particular value probability is high.

↳ dirac delta function.

known as

as $\alpha \downarrow$ tail will reach to 0 for very high value.

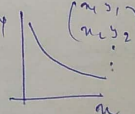
pareto distribution is the distribution where, most population is having small value and small population having a very large value.

Application ↳ Hard disk drive error.

↳ sizes of human settlement (few cities, many villages)

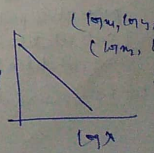
Log-log plot

To check power law.



take log on both sides

$\log y = \log - \log$ both sides it should follow straight line



Or make QQ plot of (x, y)
x has ↳ as pareto distributed

↳ distribution of wealth in society
↳ pareto distribution.
↳ type portion of wealth is own by small portion of people.
(80-20 Rule)

Box Cox transformation

we know that with gaussian distribution we can calculate a lot of things.
So, the conversion of pareto distribution is necessary into gaussian distribution.
because in machine learning Gaussian distribution is most often used.

Here, pareto ~ X : [x1, x2, x3 ... xn] → we have to convert
Gaussian ~ Y : y1, y2 ... yn

So, to convert this box-Cox is used

① box-cox(x) = (xi^λ − 1)/λ → this give = λ

② yi = { (xi^λ − 1)/λ if λ ≠ 0
       { log(xi)   if λ = 0
∀ i : 1 → n

Gaussian distributed

if λ = 0
xi = log(xi)
else

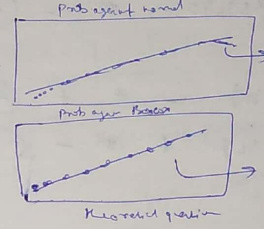
the value we get after this conversion will be normally distributed.

Code

⇒
```
from scipy import stats
import matplotlib.pyplot as plt
```
we will generate some random variables from a non-normal distribution and make a probability plot for it, to show tail is not normally distributed.

⇒
```
fig = plt.figure()
ax1 = fig.add_subplot(211)
x = stats.loggamma.rvs(5, size=500) + 5
prob = stats.probplot(x, dist = stats.norm, plot = ax1)
ax1.set_xlabel('')
ax1.set_title("prob against normal distribution")
```
we now change data using sub plot so its close to normal.

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



→ Here data we derived in first cell bet not so not normally distributed

→ data follows straight line, so normally distributed.

⇒ Box Cox work on every data, its working or not can be checked using QQ plot.

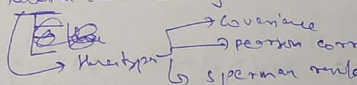
⇒ Box Cox transform power law transform into normal distribution.

Co-variance

Suppose we = height
y = weight

	x = h	y = w
s1	160	62
s2	170	54
.	.	.
.	.	.
sn	140	49

So, here we need to find what relation does x and y hold.



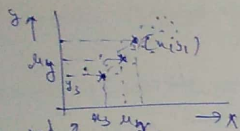
Use law,

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

$$Var(x) = Cov(x,x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)$$

if Cov(x,y) = +ve x↑, y↑
Cov(x,y) = −ve x↑, y↓

For



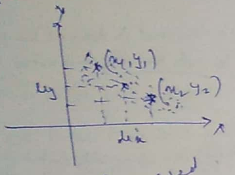
+vely correlated

$(x_1 y_1)$ +

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

$(x_1 y_1) = +\ *\ +$ (+ve)

$(x_3 y_3) = -\ *\ -$ (+ve)



→ $(x_1 y_1)$ $-\ *\ +$ (−ve)

$(x_2 y_2)$ $+\ *\ -$ (−ve)

−vely correlated

again suppose $\text{Cov}(\overset{cm}{x}, \overset{kg}{y}) \neq \text{Cov}(\overset{ft}{x}, \overset{pound}{y})$

if height in cm is correlated with weight in kg, then height in ft and ~~weight~~ weight in pound may not be correlated.

Pearson-Correlation Coefficient (PCC)

$$\rho_{x,y} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} \qquad \sigma_x = \sqrt{\text{Var}(x)}$$

The covariance

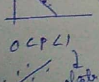
if $x\uparrow$ and $y\uparrow$, $\text{Cov}(x,y) \to$ +ve

if $x\downarrow$ and $y\downarrow$, $\text{Cov}(x,y) \to$ −ve

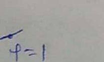
but how much +ve and −ve it will be it is never said.

So, to overcome this pearson-correlation coeff is used.

$-1 \leq \rho \leq 1$ → pcc is always less than 1 and greater than −1.

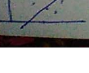


$\rho = -1$ when all data point are linear and −ve ($\rho = -1$)

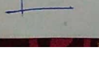


$-1 < \rho < 0$ → dataset decreasing but not linear

$0 < \rho < 1$

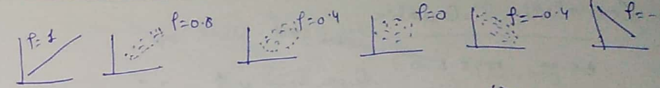


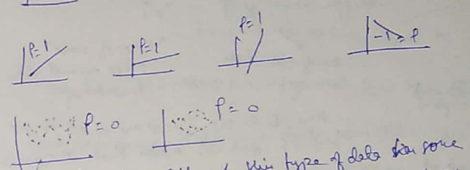
↓ dataset ↑ but non linearly



$\rho = 1$

$\rho = 0$ pearson cannot read any relation other than linear.





↳ linear relationship doesnot depends on slope.

non linear relationship b/w some other relationship exist.

↳ although this type of data has some shape but pearson can't read it

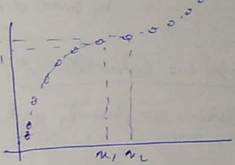
Here graph is monotonically ↑

$x_2 > x_1$

$y_2 > y_1$

∴ spearman correlation = 1

pearson correlation = 0.88.



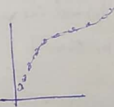
Spearman Rank Correlation Coeff. (r)

Suppose we have dataset of heights and weights

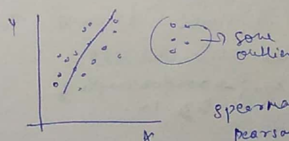
Here is spearman rank ←

we arrange the dataset A/T its rank.

	X	Y	r_x (rank of x)	r_y (rank of y)
r_1	160	52	4	3
r_2	150	66	2	4
r_3	170	68	5	5
r_4	140	46	1	1
r_5	158	57	3	2



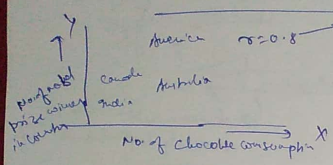
↳ so as long as there is monotonicity (increasing or decreasing) in dataset, if it is linear or not rank can be 1 or −1 ($r = 1$, $r = -1$)



spearman correlation = 0.84

pearson correlation = 0.67

Correlation Vs Causation



$r = 0.8 \rightarrow$ Correlation $= 0.8$.

Suppose here X ↑ Y ↑

X → no. of chocolate consumption is increasing

Y → no. of nobel prize winner also increasing.

Here both are increasing and $r = 0.8$ but it doesnot mean that X causes Y

Both of these are increasing by some different ways → no one is affecting other.

How to use Correlations?

Correlation are heavily used in many areas, like, ~~ecommerce~~ ecommerce, etc.

Suppose In amazon website

if (unique visitors in a day) Vs ($ sales in a day)

↳ if this increases then sales in a day will ~~directly~~ also increase. (correlated)

So, Companies will try ~~to make their website such that~~ to get more no of visitors per day.

Suppose in education

10th →
12th →
engg →
Mtech →
phd →

Salary ↓

if we increase the study Salary might correlated increases. So education minister should focus on this.

Confidence ~~Int~~ Interval (CI)

Suppose any distribution ← X: height

↳ a random sample $\{x_1\ x_2\ x_3 \cdots x_{10}\}$ → random sample of size 10.

estimated population mean of $X = \mu$

for small sample set.

$$\mu \approx \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow \text{Simple avg.}$$

this will be more equal to $\mu = \bar{x}$ when sample size increase.

So, here this $\bar{x} = \mu$ is giving some exact (point) mean.

But for some different data samples mean could change.

So, we use Confidence Interval for this.

exp $\{x_1\ x_2 \cdots x_n\}$

$\{180, 162, 157, 172, 168,$
$150, 171, 183, 165, 176\}$ → height of people in cm

$\hookrightarrow \mu = \bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 168.5$ cm.

But C.I ⇒ $\mu \in [162.1, 174.91]$ with 95% probability.

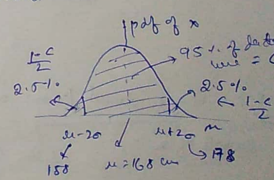
it gives a range of value with confidence of 95% probability in which ~~dataset~~ mean would lie.

Computing Confidence-interval given the underlying distribution.

Suppose we have heights → that is gaussian distributed

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

Let $\begin{pmatrix} \mu = 168\text{ cm} \\ \sigma = 5\text{ cm} \end{pmatrix}$



Q? Value of ~~the~~ range of height in the range of 95% probability.

⇒ Since it is Gaussian distributed.

∴ 95% data will lie b/w $(\mu - 2\sigma, \mu + 2\sigma)$

$[158, 178]$

∴ $C = 95\%$ So, Value of 90%, 80%, 70% probability can be known by Chart of normal distribution.

C.I for mean of a normal random variable

Suppose $X \sim F$ (not normally distributed) with population mean of μ and std. deviation of σ.
(follows)

↳ $\{x_1, x_2, x_3 \cdots x_{10}\}$ → having sample of size $n = 10$

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

Q. what is the 95% CI of μ.

Case 1 :- Suppose when $\sigma = 5$ cm is known {we know pop. std. dev}.

Here, in this case, we will use CLT, since CLT is used in a lot of mean value, but since we have one sample CLT is used.

CLT :- $\bar{x}$ = sampled mean $= \frac{1}{10}\sum_{i=1}^{10} x_i$ $(n=10)$

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

(Sample mean, follows, Gaussian dist., with pop. mean μ, and pop. std. deviation $\frac{\sigma}{\sqrt{n}}$)

$\therefore \mu \in \left[\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + \frac{2\sigma}{\sqrt{n}}\right]$ with 95% confidence

(2 → 1.96 ≈ 2; $\mu - 2\sigma$; $\mu + 2\sigma$; $\bar{x} = 168.5$ cm, $\sqrt{n} = 10$, $\sigma = 5$)

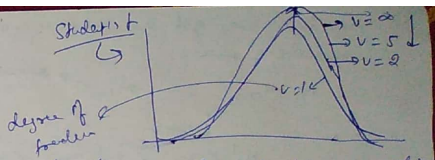
Case 2 :- if we don't know σ (pop std-dev).

Sample of n → in this case t-distribution is used. → Student's t-dist.

↳ C.I of mean when σ is unknown

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

(Sample mean, follows, t-dist, with degree of freedom $(n-1)$)



↳ this student's t looks like some what gaussian, so we can calculate from this.

Confidence Interval using Bootstrapping -

↳ using this we can calculate any C.I for median, mean, std etc.

Suppose we only have sample size and don't know the distribution or anything.

↳ sample of size 'n' $S \rightarrow \{x_1, x_2, \cdots x_n\}$ fix $n = 10$

once we have the sample using discrete uniform random variable we create different sample size of length $m \leq n$

$\therefore u(1, n)$ (1st parameter, 2nd parameter)

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

↳ Sampling with repetition ← random sample of size m generated from S.

(this is bootstrap samples)

$S_2 = \{x_4^{(2)}, x_2^{(2)}, \cdots x_m^{(2)}\}$

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

$\vdots$

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

Here we generate k different samples from sample S with repetition then calculate median for all sample k for CI of median.

Suppose median of sample $S_1 \rightarrow m_1$,

$S_2 \rightarrow m_2$

$S_3 \rightarrow m_3$

$\vdots$

$S_k \rightarrow m_k$

Now once we got the medians, we arrange the medians in ascending order.

$m_1, m_2, \ldots, m_{1000}$ → suppose k = 1000

↓ sort

$m_1' \le m_2' \le m_3' \ldots \le m_{1000}'$

↓ CI (95%)

2.5 m'_{25} ... m'_{975} ... 2.5

95% value

95% → 950 / 1000

Now to get 95% CI. out of 1000 sample medians we take median from m'_{25} — m'_{975}.

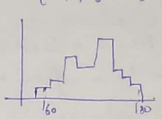
Similarly we can calculate for variance also.

Code

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

# load dataset
⇒ X = numpy.array([180, 162, 158, 172, 163, 150, 171, 183, 165, 176])

# Configure bootstrap
⇒ n_iterations = 1000
  n_size = int(len(X))

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

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

Confidence Intervals

```
alpha = 0.95
p = ((1.0 - alpha)/2.0) * 100        → 2.5
lower = numpy.percentile(medians, p)
p = (alpha + ((1.0 - alpha)/2.0)) * 100
upper = numpy.percentile(medians, p)
print('%.1f, if confidence interval %.1f %%: %.1f and %.1f' % (alpha*100, lower, upper))
```



95.0 confidence interval 161.5 and 176.0

Hypothesis testing methodology, Null-hypothesis, p-value.

Suppose we have 500 students in two classes each.

So, out of 500 students in each class, we take 50-50 samples from two class. and noted down the heights of each students.

	Cl1	Cl2
1	160	162
2	152	156
⋮	⋮	⋮
50	142	182

So, the Question here is

Is there any difference b/w the heights of students b/w these two classes.

So, to know this mean would do a better work.

Suppose the mean of Cl1 is x_1

Cl2 — x_2.

So to do hypothesis testing we do the following steps.

① Choosing the test-statistic

Here we have chosen mean. ($x_2 - x_1$)

↳ if there is difference in mean, then this shows height of Cl2 student is greater than Cl1 in most of the case.

(2) Null Hypothesis (H_0) (Here we will follow proof by contradiction)

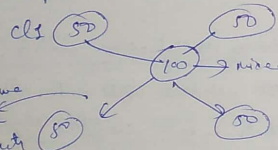
H_0 → no difference in μ_1 and μ_2.

Alternate hypothesis → (H_1) → difference in μ_1 and μ_2.

(3) p-value :- Check if prob. of observation ($\mu_2 - \mu_1$) $\geq$ 10cm (mean) iff null hypothesis is true.

Now, from the observation we know there is a difference b/w mean of two class is greater than equal to 10 cm. from all the 10000 students of two class.

Now to check for 100 students → 50 from one class and 50 from other.



[← using permutation and Resampling]

after mixing we again resample if 50-50 students then calculate the mean.

Suppose $\mu_1 - \mu_2 = \delta_1$ → from first resample.

We repeat again and again resampling suppose we took 10000 times.

$\mu_1 - \mu_2 = \delta_2$ → 2nd resample

⋮

$\mu_1 - \mu_2 = \delta_{10000}$ → 10000th resample

We get $\delta_1, \delta_2, \delta_3, \ldots \delta_{10000}$

↓ sorting in ascending order.

$\delta_1^{(1)}, \delta_2^{(1)}, \delta_3^{(1)} \ldots \delta_{10000}^{(1)}$ (1 for 1 sample)

Suppose $\chi = \mu_1 - \mu_2 = 10$ cm lies at 8000th place

So, prob. that ($\mu_1 - \mu_2 = 10$) greater than 8000 $= \frac{2000}{10000} \times 100$

$= 20\%$.

↳ So here null hypothesis is right.

Let just say. if 10 cm lies at 200 place.

Then prob that $\chi > 10$ cm $= \frac{200}{10000} \times 100$

$= 2\% < 5\%$

∴ Null hypothesis is wrong ∴ check for alternate hypothesis.

So. if p.value ($\chi \geq 10$ cm (true) | H_0 (true)) $= 0.9$

↳ ∴ H_0 → accepted

if p.value ($\chi \geq 10$ cm | H_0) $= 0.01$

↓ this is from observation ∴ true

↓ false

∴ H_0 → rejected.

Hypothesis testing :- Coin intuition.

Ex:- Here, we have given a coin and we have to check if the coin is biased towards heads or not,

if it is biased towards heads then $P_{head} > 0.5$

if it is not biased towards heads then, $P_{head} \geq 0.5$

for this we flip the coin 5 times and count heads ⇒ χ → Test-statistics

after performing the expt suppose we get flip, flip, flip, flip, flip

H H H H H $\chi = 5$

↳ this is observational expt.

$P(\chi = 5 \mid \text{coin is not biased towards heads}) = p(\text{obs} \mid H_0)$

(obs: observation)

Here, we will check, how the hypothesis is coin is not biased towards heads.

not biased ⇒ H_0 ⇒ we flip 5 time

↳ $P(\text{head}) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{2^5} = \frac{1}{232} = 0.03 = 3\%$.

But this shows that when a coin is flipped 5 times, there are 3% probability that all five head head occur.

Here 3% < 5%. So, our hypothesis is false, and as by observation we got all five head ∴ coin is biased towards head.

Again we need to choose hypothesis and testing very carefully, because this changes the whole scenario.

Suppose we trice the (the no. of times coin was tossed) tossed to 3.

Now, observation = 1×1×1 = n×n×n = n^3

toward coin it is not biased = $\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} = \frac{1}{8}\times 100 = 12.5\% > 5\%$

12.5% times we can observe n^n. which is Less.

Here, H_0 is false. ∴ Coin is not biased.

Q&A ⇒ Here observation is done only once, because, doing it n times would cost you a lot of money.

P-value is defined for single experiment.

K-S test for Similarity of two distribution.

Here, K-S stands for Kolmogorov - Smirnov Test.

This test is used when we have to check if two distributions are similar or not.

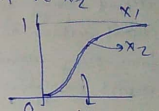
Suppose, we have, two set of random variables here,

$X_1 \Rightarrow [x_{11}, x_{12}, \dots x_{1n}]$ → n variable.

$X_2 \Rightarrow [x_{21}, x_{22}, \dots x_{2m}]$ → m variable

Now we have to check, if X_1 and X_2 come from same distribution or not.

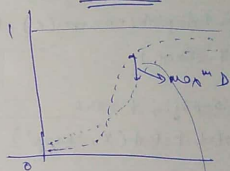
So, for this to happen, for two distribution to be same, cdf of X_1 should lie on each other to cdf for X_2



Here cdf of X_1 overlaps cdf of X_2 hence, comes from same pdf.

if n and m are large, then distance b/w two cdf must be 0.

if we have small datasets, or large there should be a min Distance(D) after which our hypothesis will fail.



Now, we have two R.V.s as X_1 and X_2

H_0 → Null hypothesis → two R.V. X_1 and X_2 comes from same distribution.

Now, max distance = $D_{n,m} = \sup |F_{1,n}(x) - F_{2,m}(x)|$

sup → supremum function. empirical distribution. max gap.

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

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

for rejection of H_0.

α	0.10	0.05	0.025	0.01
c(α)	1.073	1.224	1.358	1.517

5%,

To reject H_0 $D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$

Let $D_{n,m} = D$

p-value < α

$D\sqrt{\frac{nm}{n+m}} > c(\alpha) = \sqrt{-\frac{1}{2}\ln\alpha}$

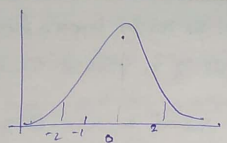
$D\sqrt{\frac{nm}{n+m}} > \sqrt{-\frac{1}{2}\ln(\alpha)}$

$-D^2\left(\frac{nm}{n+m}\right) < \frac{1}{2}\ln(\alpha)$

$\exp\left[-2D^2\left(\frac{nm}{n+m}\right)\right] < \alpha$

p-value.

Code snippet of k-s test



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

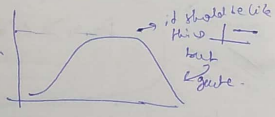
# Generate Gaussian(Normal) r.v of X.
x = stats.norm.rvs(size=1000); → This will generate a gaussian/Normal r.v of size 1000.
sns.set_style('whitegrid')
sns.kdeplot(np.array(x), bw=0.5) → scaling factor for Std. of data.
→ It takes data as 1D array.
plt.show()

⇒ # applying k-s test → data: 1-D array or string
↳ stats.kstest(x, 'norm')

Out:- (statistic = 0.02150, pvalue = 0.758)
↳ Distance D
↳ 75% ⇒ ∴ Hypothesis accepted.

Y – Continuous Uniform Distribution (0,1)

y = np.random.uniform(0,1,1000); → generate 1000 uniform random variable
sns.kdeplot(np.array(y), bw=0.1)
plt.show()



⇒ stats.kstest(y, 'norm') → it can be any distribution as we want.

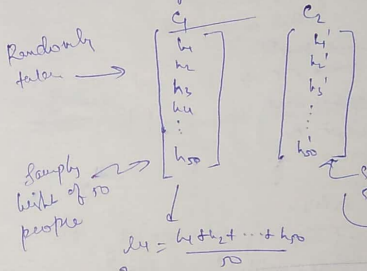
Out:- (statistic = 0.50159, pvalue = 0.0)
↳ D → max bar
↳ 0% ⇒ Hypothesis fails here that Y is Gaussian distributed.

Hypothesis Testing: Another Example

Task:- Determine if population mean of height of people in these two cities are equal or not.

Suppose City 1 has population of 1 million. (C_1)
City 2 has population of 2 million (C_2)

Since asking height to 1 million people is a tough task. So, we take 50-50 samples from two cities.



$\bar{h}_1 = \frac{h_1 + h_2 + \cdots + h_{50}}{50}$ → mean

$\bar{h}_2 = \frac{h_1' + h_2' + \cdots + h_{50}'}{50}$

↳ $\bar{h}_2 = 167$ cm

Suppose $\bar{h}_1 = 162$ cm

↳ Mean heights are the observed values.

Now,

Test statistics :- $\bar{h}_2 - \bar{h}_1 = x = -162 + 167$ cm $= 5$ cm
(C_2 C_1)
↳ difference in population means of two cities

Null-Hypothesis(H_0) :- There is no difference in population mean.

Compute for: $P(x = 5\text{cm} / H_0)$

probability of observing a diff. of 5 cm in sample mean heights of sample size 50 b/w C_1 and C_2, if there is no population difference in mean heights.

Case 1: $P(x=5/H_0) = 0.2 = 20\%$ (Suppose)

There is 20% chance of observing a difference of 5cm in sample mean heights of C_1 & C_2 with sample size of 50, if there is no population mean difference.

Here 20% is significant.
∴ assumption must be true.
∴ accept H_0.

Case 2: $P(x=5/H_0) = 0.03 = 3\%$

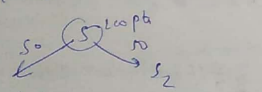
$P(\text{obs}/\text{assumption}) = 3\% \rightarrow$ small $\rightarrow < 5\%$

⇒ assumption must be incorrect
⇒ reject H_0 ⇒ accept H_1

Resampling and Permutation tests -

Step 1: $S = \{h_1, h_2, h_3, \ldots h_{50}, h_1', h_2', \ldots h_{50}'\}$

$C_1 \cup C_2 \rightarrow$ Concatinating the two set.



random selection

$\{h_1, h_2', h_3, h_4, \ldots h_{50}'\}$ → H_1

$\{h_4, h_7', \ldots\}$ → H_2 → resampling

→ Simulate H_0

Assuming first resampling:

① $h_2 - h_1 \rightarrow 3\text{cm} \rightarrow \delta_1 \leftarrow$ diff 1

repeat ② $h_2 - h_4 \rightarrow -2\text{cm} \rightarrow \delta_2$

③ $h_2 - h_4 \rightarrow 1\text{cm} \rightarrow \delta_3$

⋮

Let's take = 1000

④ $H_2 - H_1 \rightarrow 6\text{cm} \rightarrow \delta_{1000}$

Step 3: sort δ_i

$\delta_1' \leq \delta_2' \leq \delta_3' \cdots \leq \delta_{1000}'$

Simulated diffs

Case 2: Since observed diff = 5cm.

Suppose $\delta_1' \leq \delta_2' \leq \delta_3' \cdots \delta_{800}' \leq 5\text{cm} \leq \delta_{801}' \cdots \leq \delta_{1000}'$

80% of sim diff. ≤ 5cm

20% of simulated diff > observed diff.

∴ Probability $P(\text{obs-diff}/\text{assumption}) = 20\% > 5\%$.

∴ assumption = true
accept H_0.

if $P(\text{obs-diff}/\text{assumption}) = 3\% < 5\%$.

assumption = false
reject H_0.

Hypothesis testing. (How to use)

↳ we have seen (k-s test) that tells us about if the random variable follows same distribution or not

⇒ Suppose we have two drugs here D_1 and D_2 in which D_1 claims that it can reduce fever in an hour and D_2 claims that it can reduce fever before D_1.

(D_1 in the market) $D_2 \rightarrow$ new drug.

So, here court of India has to decide whether D_2 is claiming right or not. Should it be giving ~~licence~~ licence to D_2 company to sell it or not. So, it can be tested using hypothesis testing.

Here,

Task :- expt :- to determine if claim is true or not.

experiment

D_1 → 100 random Pick → D_2

S_1 = | 41 | 2 | | . . . | 38 |
1 2 . . 50

| 11 | 05 | 15 | . . | 24 |
1 2 3 . 50

So, they decided to experiment this with 50, 50 samples of people, and the timing is also noted.

Suppose mean time ← $\mu_1 = 4$ hrs.

mean time for D_2 $\mu_2 = 2$ hrs.

↑ This is the observed value.

Since we can't risk the life of any human.

∴ so it will use α or p-value $< 1\%$.

Hypothesis testing :- ① H_0 : D_2 and D_2 takes the same time to reduce fevers.

② Test statistic :- $\bar{X} = \mu_2 - \mu_1 = 4 - 2 = 2$ ⇒ observation.

Now, $P(\bar{X} \geq 2 / H_0) = V.small$ (Suppose 1%)

obs

p-value

If there is no difference in D_1 and D_2, the prob. of obs. that $\bar{X} \geq 2$ is very small (1%)

∴ H_0 and obs do not agree with each other.

∴ H_0 is rejected (incorrect)

for e-commerce company like amazon, we can use $\alpha \leq 5\%$. because for a little loss of money can be tolerable.

Proportional Sampling. (probability sampling)

Suppose we have taken 5 random elements

d_1 d_2 d_3 d_4 d_5 . . d_i

d = | 2.0 | 6 | 12 | 5.8 | 20.0 |
1 2 3 4 5 = n

Task :- Pick an element among the n elements such that probability of picking an element is proportional to d_i's.

Sum of all the elements = 35.

Step 1 :- ⓐ Compute the sum ⇒ $s = \sum_{i=1}^{5} d_i = 35$.

ⓑ Normalizing using the sum.

$d_i' = d_i / s$

$\sum d_i' = \sum \frac{d_i}{s}$

$d_1' = 0.0571$
$d_2' = 0.171428$
$d_3' = 0.0343$
$d_4' = 0.1657$
$d_5' = 0.5714$

0 to 1
sum to 1

$\tilde{d}_1 = d_1' = 0.0571$
$\tilde{d}_2 = \tilde{d}_1 + d_2' = 0.228528$
$\tilde{d}_3 = \tilde{d}_2 + d_3' = 0.262828$
$\tilde{d}_4 = 0.428528$
$\tilde{d}_5 = 1.00$

ⓒ Cumulative Normalized Sum.

$\tilde{d}_i$ = cum-norm-value

Step 2 Sample one random uniform value (x) ⇒ unif(0, 1)

x = numpy.random.uniform(0, 1, 1) ← max
↑ range

Let $r = 0.6$.

Step (3) Prop sampling.

if $r \le \tilde{d}_1$
return 1
else if $r \le \tilde{d}_2$
return 2
else if $r \le \tilde{d}_3$
return 3
:

Here $r = 0.6 < \tilde{d}_5$
∴ pick d_5

Ans. prob. of picking 4.
= prob of r lying b/w $\tilde{d}_3$ & $\tilde{d}_4$
$= d_4'$
$\propto d_4$. $\Rightarrow d_4' = \frac{d_4}{5}$

~~Q~~ Q/A → 3-college student examples.