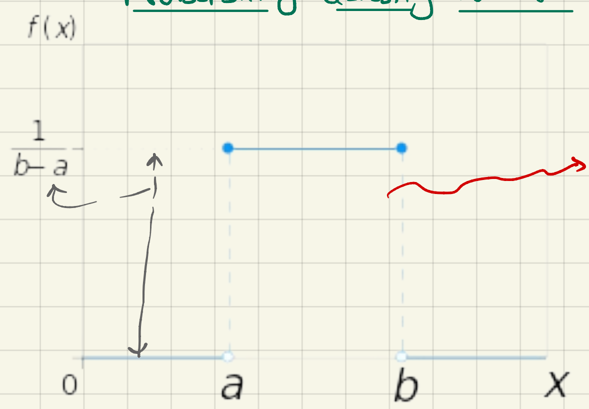
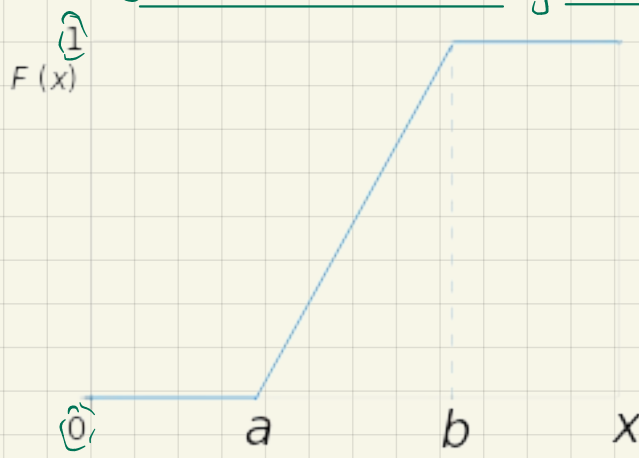
Continous uniform distribution

Probability density function (not mass function)



→ ar. under the curve should be one.

Cummolative distribution function



Notation :- $U(a, b)$

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

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

0 otherwise

Random number Generator (uniform distribution)

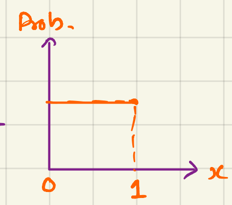
C, Java → stdlib.h → rand()

↳ pseudo random number generators.

→ If not it's explicitly called out, most random no. generator are uniform generator.

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

→ 0, 1



↳ If you generate '1000' values & plot it. It will look like this.

→ These no.s are randomly generated with uniform distribution.

↓ it means

the no. b/w '0 to 1' have equal probability to get selected.

→ suppose → n data points & I want to uniformly sample "m" data points from this data points

↓

The chances or probability for getting selected in 'm' is equal for all data points in 'n'.

Q) If you don't know about 'Pandas' & 'Numpy' and your task is

$D = x_1, x_2, x_3 \ldots\ldots\ldots x_{150}$

There is $\frac{30}{150}$ = 20% chance that this 'x_1' belong to D'.

↓ Sample uniformly

$D' = x_1', x_2', x_3', \ldots\ldots\ldots x_{30}'$

Random Number Generator (Uniform Distribution)

```
[1] import random
    print(random.random())

0.05856684124159628
```

```
#load IRIS dataset with 150 points.
from sklearn import datasets
iris = datasets.load_iris()
d = iris.data
d.shape
```

→ I want to generate random 30 point data set from this data set.

→ (PL, PW, SL, SW)

```
(150, 4)
```

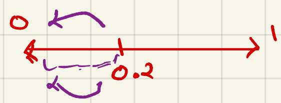
```
# Sample 30 points randomly from the 150 point dataset
n=150
m=30
p = m/n
print(p)
sampled_data =[];

for i in range(0,n):
  a = random.random()
#   print(a)
  if  a <= p:
    sampled_data.append(d[i,:])

print(sampled_data)
print(len(sampled_data))
```

→ 0.2

→ I know this generate a random value from '0 to 1' with equal probability.



→ U(0,1)

→ random no. generator are never perfect. So, whenever you run this it will keep on changing. (It will not always be '30' it will be roughly around '30'.

```
0.2
[array([4.9, 3. , 1.4, 0.2]), array([4.6, 3.4, 1.4, 0.3]), array([5.4, 3.7, 1.5,
32
```

Bernoulli and Binomial Distribution

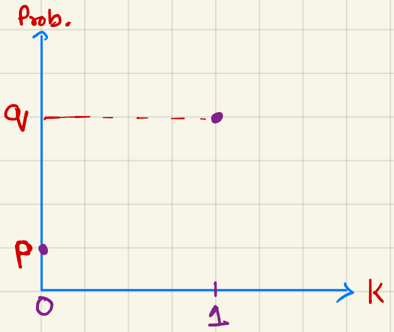
Bernoulli's

→ It has only two outcomes → 1 → with probability 'p'
↳ 0 → " " '1−q'

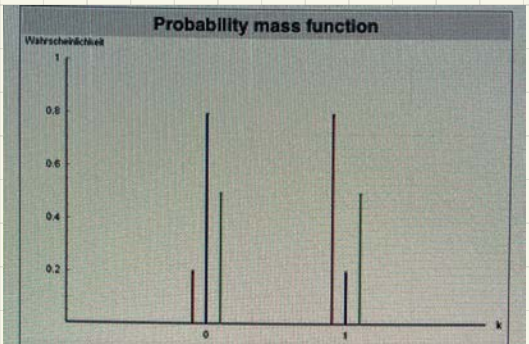
→ X ~ Bernoulli(P=0.5)
↳ it is like coin toss → H(1)
→ T(0)

→ it is a discrete random variable.

→ Pmf



$p+q=1$



Three examples of Bernoulli distribution:
- $P(x=0)=0.2$ and $P(x=1)=0.8$
- $P(x=0)=0.8$ and $P(x=1)=0.2$
- $P(x=0)=0.5$ and $P(x=1)=0.5$

Parameters	$0 \le p \le 1$, $P\in R$ $q = 1-p$
Support	$k \in \{0,1\}$
PMF	$\begin{cases} q=1-p & \text{if } k=0 \\ p & \text{if } k=1 \end{cases}$
CDF	$\begin{cases} 0 & \text{if } k<0 \\ 1-p & \text{if } 0\le k<1 \\ 1 & \text{if } k \ge 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0 & \text{if } p<1/2 \\ [0,1] & \text{if } p=1/2 \\ 1 & \text{if } p>1/2 \end{cases}$
Mode	$\begin{cases} 0 & \text{if } p<1/2 \\ 0,1 & \text{if } p=1/2 \\ 1 & \text{if } p>1/2 \end{cases}$
Variance	$p(1-p)=pq$
MAD	$\frac{1}{2}$
Skewness	$\frac{q-p}{\sqrt{pq}}$
Ex. kurtosis	$\frac{1-6pq}{pq}$
Entropy	$-q\ln q - p\ln p$
MGF	$q+pe^t$

Bernoulli and binomial distributions are closely related, but they differ in how they model events:

- Bernoulli Models a single event or non-event trial. It has two possible outcomes: 0 or 1.
- Binomial Models multiple events occurring at the same time. It's the sum of independent and identically distributed Bernoulli random variables.

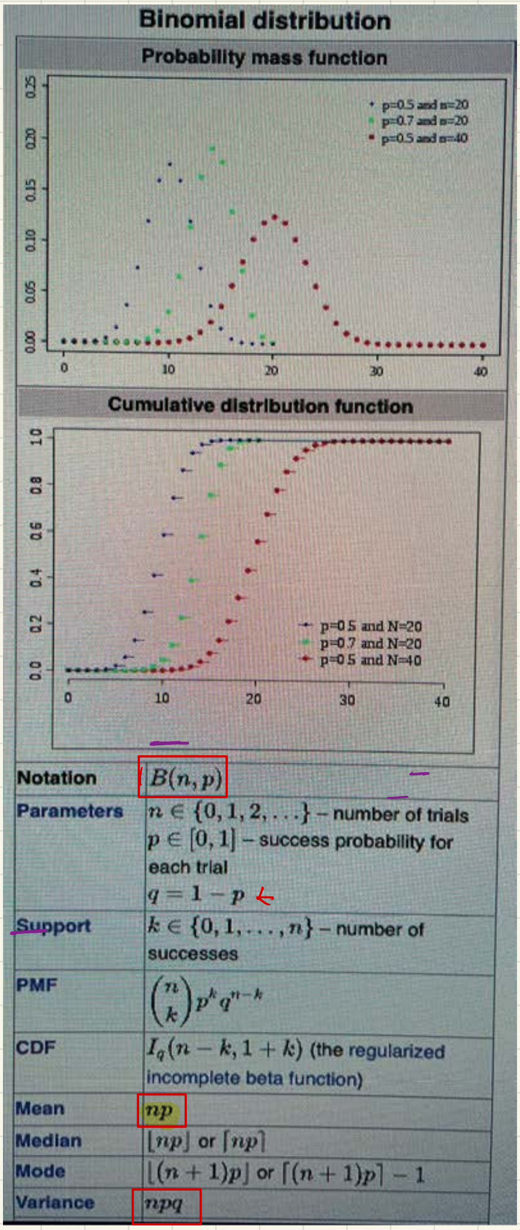
Bernoulli Distribution Examples

1. Flipping a coin: The outcome of a single coin flip can be modeled as a Bernoulli random variable, where "heads" is considered a success (1) and "tails" is a failure (0). The probability of success (getting heads) is p = 0.5.
2. Rolling a die: The outcome of rolling a single die can be modeled as a Bernoulli random variable, where rolling a "6" is considered a success (1) and any other number is a failure (0). The probability of success (rolling a 6) is p = 1/6.
3. Passing a test: The outcome of a student taking a test can be modeled as a Bernoulli random variable, where "passing" the test is a success (1) and "failing" the test is a failure (0). The probability of success (passing the test) is the student's exam score.

Binomial Distribution Examples

1. Flipping a coin 10 times: The number of heads obtained in 10 coin flips can be modeled as a binomial random variable, with n = 10 trials and p = 0.5 for the probability of getting heads on each flip.
2. Rolling a die 20 times: The number of times a "6" is rolled in 20 die rolls can be modeled as a binomial random variable, with n = 20 trials and p = 1/6 for the probability of rolling a 6 on each roll.
3. Students passing a test: If 100 students take a test, and each student has a 70% chance of passing, the number of students who pass the test can be modeled as a binomial random variable, with n = 100 trials and p = 0.7 for the probability of passing.

Binomial



Binomial distribution

Notation	$\boxed{B(n,p)}$
Parameters	$n \in \{0, 1, 2, \ldots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial $q = 1 - p$ ←
Support	$k \in \{0, 1, \ldots, n\}$ – number of successes
PMF	$\binom{n}{k} p^k q^{n-k}$
CDF	$I_q(n-k, 1+k)$ (the regularized incomplete beta function)
Mean	$\boxed{np}$
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$\boxed{npq}$

→ Two parameter → n → no. of trials (eg. no. of coin tosses)
↳ p → prob. value (eg. prob. of getting head)

→ $Y \sim Bin(n,p)$

The binomial distribution is a discrete probability distribution that describes the number of successes in a sequence of independent experiments, each with a yes–no question and a Boolean-valued outcome. The Gaussian distribution is a special case of the binomial distribution when the number of tries is large enough. It applies to a large number of variables and has a symmetric and unimodal shape with tails that extend to positive and negative infinity.

A binomial distribution has a finite amount of events, while a normal distribution has an infinite number. In a binomial distribution, there are no data points between any two data points, while a normal distribution has continuous data points.

Log-normal distribution

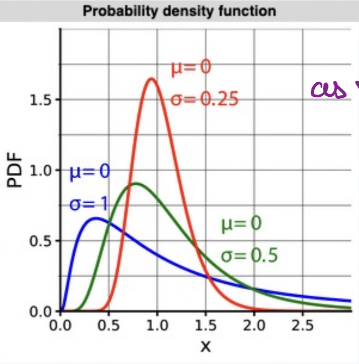
- $X \sim$ log-normal (μ, σ) if $\ln(x)$ has a normal distribution

→ you can convert 'log-normal distribution' to 'Gaussian / Normal distribution'.

we want to convert log → Gaussian coz → we have many algo. or concepts for Gaussian.

Log-normal

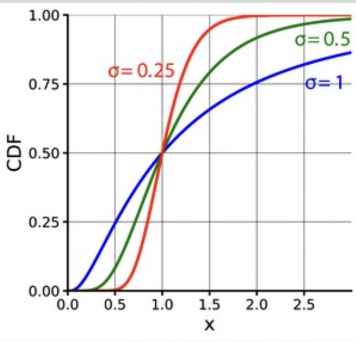
Probability density function



Identical parameter μ but differing parameters σ

as σ (↑), spread also (↑)

Cumulative distribution function



$\mu = 0$

Notation	$\mathrm{Lognormal}(\mu, \sigma^2)$
Parameters	$\mu \in (-\infty, +\infty)$, $\sigma > 0$
Support	$x \in (0, +\infty)$
PDF	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$
CDF	$\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\exp(\mu + \sqrt{2\sigma^2}\,\mathrm{erf}^{-1}(2p - 1))$
Mean	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$
Median	$\exp(\mu)$

Q) if we know $x \sim$ log-normal (μ, σ)

$x_1, x_2, x_3, \ldots\ldots x_n$

Q) is x is ~log normal?

→ $x_1, x_2, x_3, \ldots\ldots x_n$

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

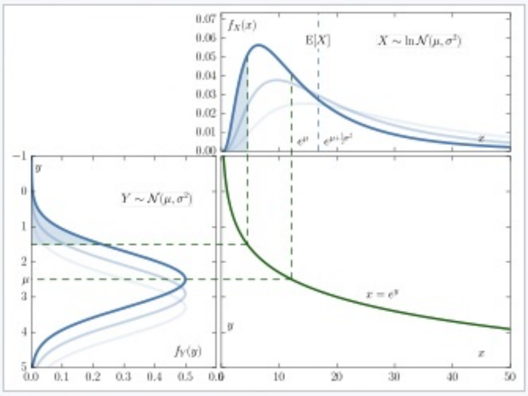
$y_1, y_2, \ldots\ldots y_n$

→ (Q-Q plot)

y_i's are Gaussian

If it's true then

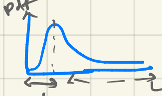
$X \sim$ log-normal



Relation between normal and log-normal distribution. If $Y = \mu + \sigma Z$ is normally distributed, then $X \sim e^Y$ is log-normally distributed.

→ log-normal distⁿ occurs a lot in nature.

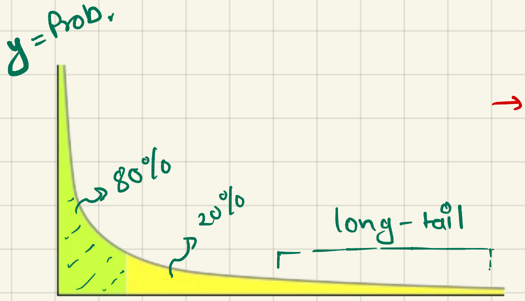
→ on internet (like comments)



→ long comment / more salaries / etc

→ short comment / less salaries / etc

Power law and Pareto dist^n



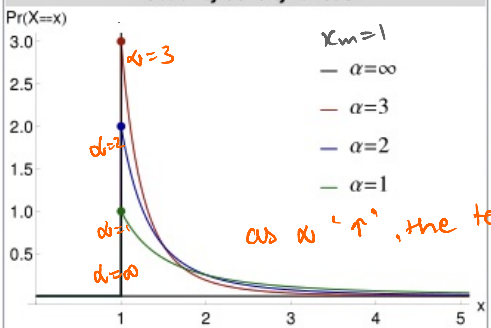
→ it follows 80 - 20 rule → 80% of time you will find the result in 20% of distribution. (approx).

→ They occur lot in 'internet' (on comments)

→ Whenever log-low follows a dist^n it's called 'Pareto dist^n'.

Pareto Type I

Probability density function

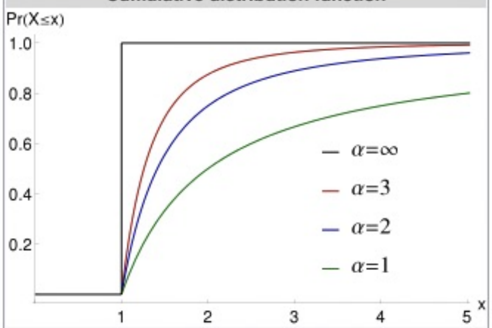


Pareto Type I probability density functions for various α with $x_\mathrm{m} = 1$. As $\alpha \to \infty$, the distribution approaches $\delta(x - x_\mathrm{m})$, where δ is the Dirac delta function.

→ have two parameter.

as α '↑', the tail-flatness also '↑'.

Cumulative distribution function



Pareto Type I cumulative distribution functions for various α with $x_\mathrm{m} = 1$.

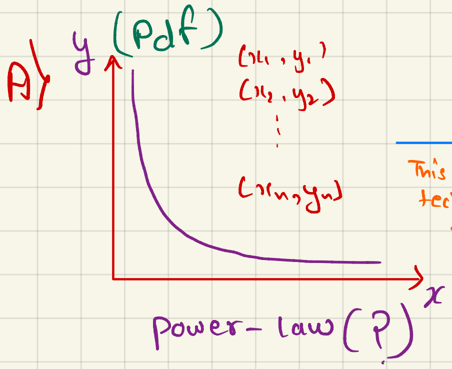
Parameters	$x_\mathrm{m} > 0$ scale (real) ← $\alpha > 0$ shape (real) ←
Support	$x \in [x_\mathrm{m}, \infty)$
PDF	$\frac{\alpha x_\mathrm{m}^\alpha}{x^{\alpha+1}}$
CDF	$1 - \left(\frac{x_\mathrm{m}}{x}\right)^\alpha$
Quantile	$x_\mathrm{m}(1-p)^{-\frac{1}{\alpha}}$
Mean	$\begin{cases} \infty & \text{for } \alpha \le 1 \\ \frac{\alpha x_\mathrm{m}}{\alpha - 1} & \text{for } \alpha > 1 \end{cases}$
Median	$x_\mathrm{m}\sqrt[\alpha]{2}$
Mode	x_m
Variance	$\begin{cases} \infty & \text{for } \alpha \le 2 \\ \frac{x_\mathrm{m}^2 \alpha}{(\alpha-1)^2(\alpha-2)} & \text{for } \alpha > 2 \end{cases}$

Application:-

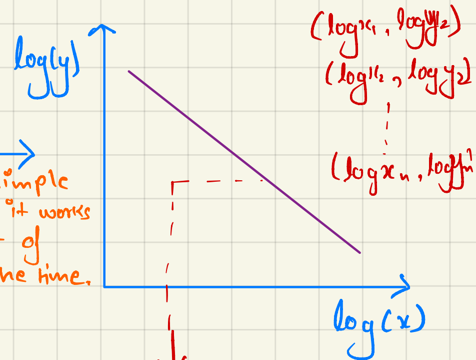
- size of human settlements (cities, rural etc)
- oil reserve (few large oil reser. and large few reser.)

Q) How to determine the dist^n we have is a pareto-dist^n or not?

A)



This is simple techn. it works most of the time.



if you get st. line we can say it follow power-law.

OR

Q-Q plot → (X, Y)

X ↓ obs.

Y ↳ pareto

Power-transform (Box-cox transform)

$(x_1, x_2, \dots x_n)$ $X \xrightarrow{\ln} Y$ $(\ln x_1, \ln x_2, \dots)$ $(y_1, y_2, \dots)$

X ↓ log-normal

Y ↓ Gaussian

ML :- assume that a feature/variable Gaussian disbⁿ.

Q) Can we convert our "power-law / pareto" to Gaussian disbⁿ?

↳ we can do that by Power-transform (Box-cox transform)

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

Gaussian ~ $Y : [y_1, y_2, \dots y_n]$

conversion

① box-cox (x) = (λ) lambda

$x_1, x_2 \dots x_n$

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

gaussian disbⁿ.

if $\lambda = 0$

$\Rightarrow x_i \rightsquigarrow$ log-normal

else

$X \longrightarrow$ box-cox$(x) \longrightarrow \lambda \longrightarrow Y$ (gaussian)

(pareto)

library, module, function

if you know, give it. Otherwise, it will compute by itself.

```
scipy.stats.boxcox(x, lmbda=None, alpha=None, optimizer=None)
```

Return a dataset transformed by a Box-Cox power transformation.

↳ it will return 1-D array

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

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

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

→ creating random variable from disbⁿ called "loggamma".

→ plotting 'Q–Q' plot with normal disbⁿ.

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

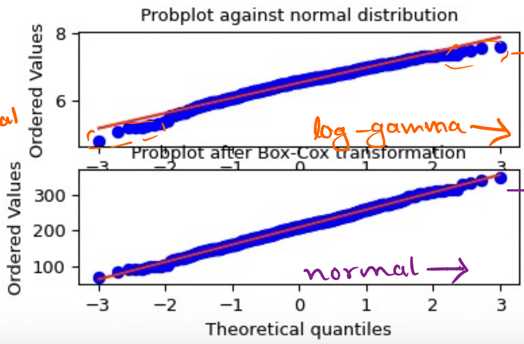
```
>>> ax2 = fig.add_subplot(212)
>>> xt, _ = stats.boxcox(x)
>>> prob = stats.probplot(xt, dist=stats.norm, plot=ax2)
>>> ax2.set_title('Probplot after Box-Cox transformation')
```

x → loggamma

xt → Gaussian

→ using 'Q–Q' plot we have verifyed that 'xt' is gaussian.

```
>>> plt.show()
```



normal ↑

log-gamma →

xt ↑

normal →

→ lots of deviation here, so, r.v x & y not coming from similar disbⁿ.

loggamma ← x

normal ← y

→ most of pt. lie close to line, so, 'xt' and 'normal' follow the same disbⁿ.

Application of non-gaussian disbⁿ

- uniform → {random-no. generating}
- Bernoulli's, Binomial
- log-normal, Pareto

There are 100's of disbⁿ.

Q) why do we have this much disbⁿ?

random variable follows → (well-studies disbⁿ) → Theoritical model

"weibull" disbⁿ

→ dams → how high dam?
- water-storage
- how much rainfall could occur → (to mitigate accidents)

1950's Civil engg.

(maximum one-day rainfall)

$P(x \geq 20\text{cm})$

if $P(x > 20\text{cm})$ is v.v.small ($\leq 0.1\%$)

rainfall data → 30 yrs; 50 yrs

10cm
5cm
1cm
16cm

→ Can we fit a known & well-suited disbⁿ / theoritical model

$P(x \geq 20\text{cm}) = ?$

percentiles → $\frac{1}{200} = 0.5\%$ → not trustworthy

(coz only 'one' data-point in '200' data-point has > 20cm)

(so, if there is no observation what should we do?

(we can't say the $p(x > 20\text{cm})$ is zero)

disbⁿ → theoritical model (how data is distributed)

100's of disbⁿ
- Gaussian — X
- log-normal — X
- weibull disbⁿ fit ✓

To understand particle size

1. → X = max-daily rainfall
2. → it follows weibull's disbⁿ
3. → CDF or PDF $(P > 20\text{cm}) = 1 - P(x < 20\text{cm})$ → more trustworthy; cdf of x

→ when I tried to use non-parametric method (mean, median, percentile) there we got "0.5%". But, I can't trust this because I have only one observation / or I am relying only on one-data point. Where if I use all observation data and fit in a model that result is more trustworthy than non-parametric method because I am relying on many data points.

→ weibull → particle size
→ Civil Engg.

box-cox transform / Power-transform (sometimes)

$X \xrightarrow[\text{(assume)}]{\text{box-cox}} X' \longrightarrow$ all of my analyses

X ↓ non-gaussian dist^n

X' ↳ gaussian dist^n

x_1
x_2
⋮

$p(x \geq 20cm) \longrightarrow p(x' \geq 8) \longrightarrow$ cdf of $N(\mu, \sigma)$

x_1'
x_2'
⋮

Co-variance

↳ how to measure relationship b/w random variable.

→ X : heights
Y : weights

	X=h	Y=w
s_1	160	62
s_2	150	54
s_n	140	48

Q) is there any relationship b/w "X" & "Y"
for eg! — X↑, Y↑
or
X↑, Y↓

→ There are three ways to do that.
- → co-variance
- → Pearson correlation coeff.
- → Spearman rank corr. coeff.

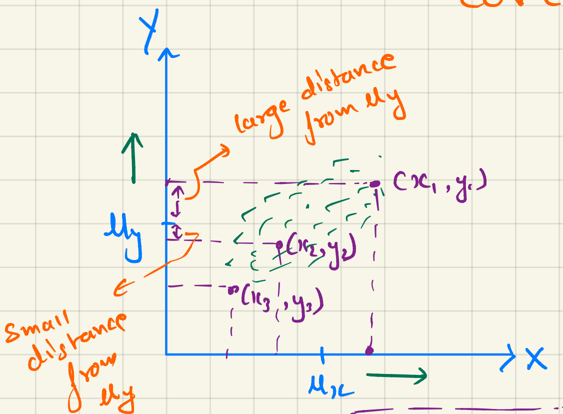
Co-variance

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

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

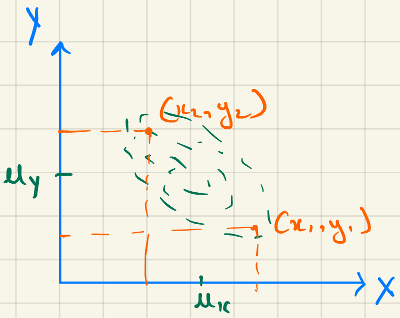
$$\mathrm{cov}(x,x) = \mathrm{Var}(x)$$

if $\mathrm{cov}(x,y)$ = +ve as X↑, Y↑
$\mathrm{cov}(x,y)$ = -ve as X↑, Y↓



co (x_1, y_1) =	+	+	+ → large value	(coz distance from 'μ' is large).
Co (x_2, y_2) =	−	−	+ → small value	
Co (x_3, y_3) =	−	−	+ → v. large	

here as X↑ → Y↑



(x_1, y_1) →	+	−	− → large '-ve'
(x_2, y_2) →	−	+	−

here as X↑ → Y↓

Problem with co-variance

1. $cov(x, y)$ — $x \rightarrow$ cm (height), $y \rightarrow$ kg (wt.)

   ↓ same data set just changed the unit/physical → Co-variance does not match.

   $cov(x', y')$ — $y' \rightarrow$ lbs (wt.), $x' \rightarrow$ ft (height)

→ we can fix that problem by using → pearson correlation coefficient.

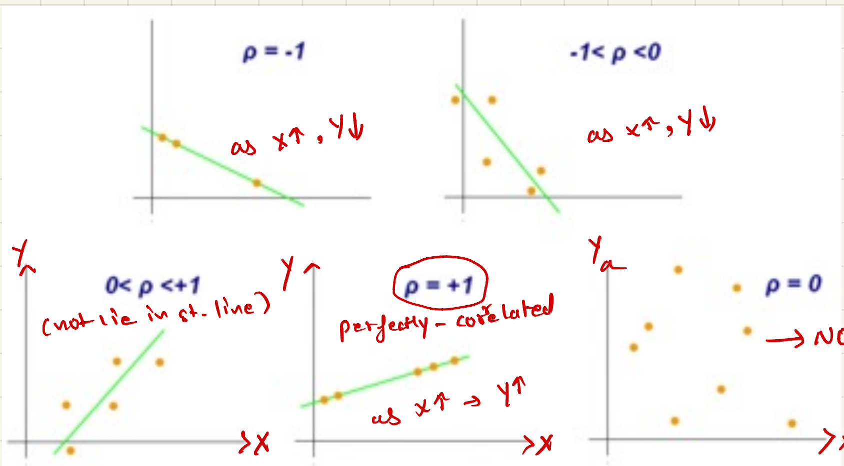
Pearson correlation coefficient (pcc)

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

$$\boxed{-1 \leq \rho \leq 1}$$

→ as $x\uparrow \rightarrow y\uparrow \rightarrow co(x,y) \rightarrow (+) \rightarrow$ but don't know how '+ve'

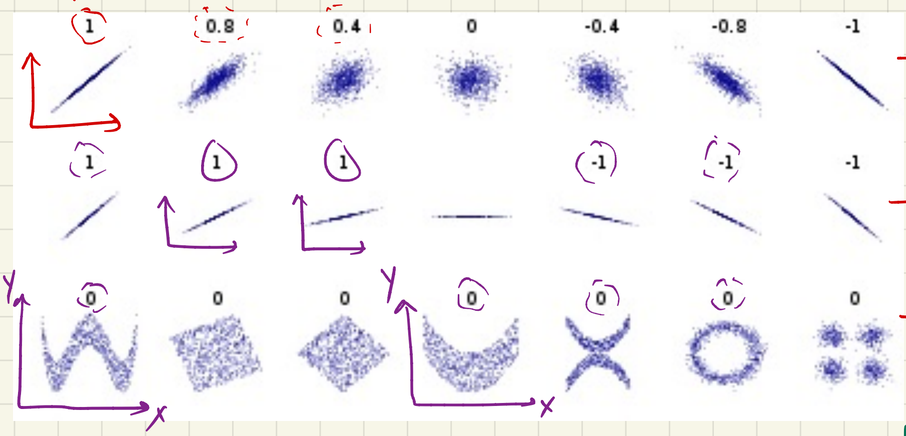
$x\uparrow \rightarrow y\downarrow \rightarrow co(x,y) \rightarrow (-) \rightarrow$ " " " " much '-ve'



→ NO relationship (Not linear relationship) b/w x, y.

→ PCC & Co-var. are biased towards linear relationship.

PCC



→ showing biasness towards linearity.

→ does not depent on slope.

→ does not capture non-linear relationship.

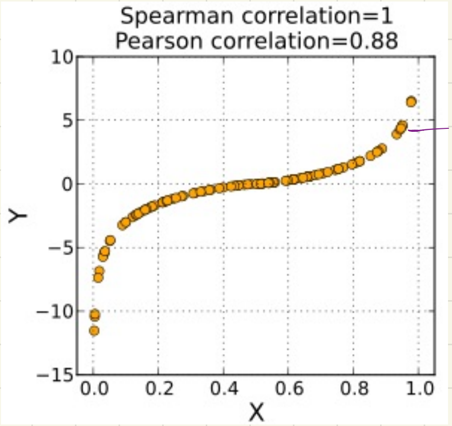
$Y = \sin(x) + err\_small$ $\quad$ $Y = x^2 + err\_small$

→ for finding full-fledge model we learn regression model in ML.

Spearman's rank correlation (γ) → Two ways of calculating that → $\gamma = \rho_{R(x),R(y)}$

→ monotonically non-decreasing → if $x_2 > x_1$ and $y_2 > y_1$ (always)

→ " decreasing → if $x_2 < x_1$ and $y_2 < y_1$ (always)



→ This data have monotonically non-decreasing fun^n but still pcc → 0.88, so, we want to correct it.

so, we want that relationship which can caught a little bit of non-linearity too.

→ sort in asscending order

⇒

	X	Y	γ_x	γ_y
S_1	160	52	4	3
S_2	150	66	2	4
S_3	170	68	5	5
S_4	140	46	1	1
S_5	158	51	3	2

$\rho_{x,y}$ → linear relationship.

$$\boxed{\gamma = \rho_{\gamma_x, \gamma_y}}$$

English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

Where d = difference between ranks and d^2 = difference squared.

We then calculate the following:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

We then substitute this into the main equation with the other information as follows:

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

$$\rho = 1 - \frac{6 \times 54}{10(10^2-1)}$$

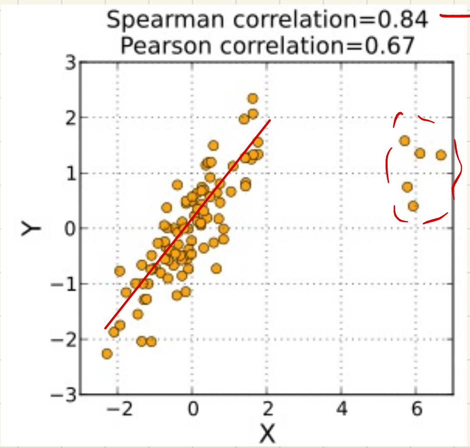
$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

→ if X ↑ and Y ↑ → linear or not → $\gamma = 1$

→ if X ↑ and Y ↓ → linear or not → $\gamma = -1$



→ much more robust towards outliers.

$$\gamma = \rho_{\gamma_x \gamma_y} = \frac{\operatorname{cov}(\gamma_x, \gamma_y)}{\sigma_{\gamma_x}\sigma_{\gamma_y}}$$

or

$$\gamma = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

→ Two ways of calculating Spearman correlation.

Correlation does not imply causation

Correlation Vs Causation

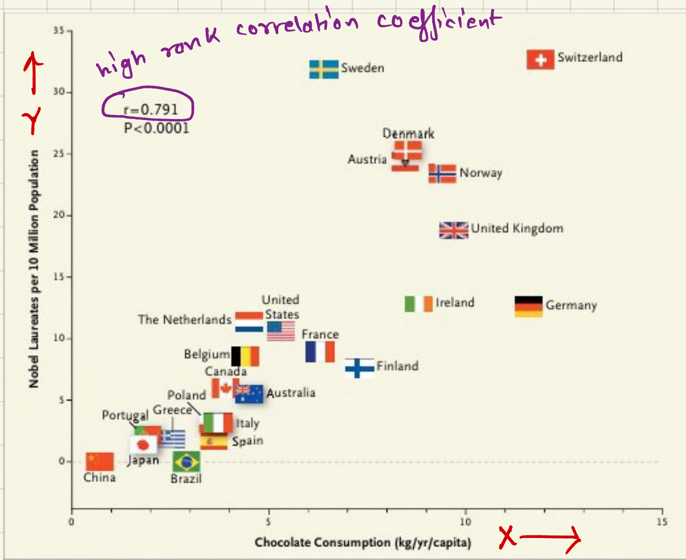


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

→ here as X↑ our Y↑ hence it is correlated

→ Can I say (X causes Y) means as chocolate consumption $\left(\frac{Kg}{person \times year}\right)$ incⁿ our no. of Nobel laureates also incⁿ?

↳ that's absurd/incorrect.

→ Just because two variable are correlated it does not mean one causes other.

→ There is a whole chapter for "Causal model" in adv. statistics & Maths.

Application of Correlation → not causation

Q) Is salary correlated with sq-footage of your home?

↳ I'm not saying higher salary causes you to look for big home.

→ we can say higher salary ppl. <u>tend</u> to buy big home.

as X↑ our Y↑

Q) Is #year-of-education correlated with income?

↳ we are not saying #year-of-eduction caused higher income.

Q) ecommerce : amazon.com

Is time-spent in 24 hrs is correlated money-spent in the next 24 hrs.

after this analysis amzon will try to make website more attractive, so ppl spend more time.

24 hrs	24 hrs
30 min.	$ 20
60 min	$ 100
⋮	⋮

Q) ecommerce :-

# Unique visitors in a day	Vs	# sales in a day
100 K	→	$1M
120K	→	$1.6M
⋮		⋮

after that company will try to inc^n this variable, coz our sale will also inc^n.

Q) Medicine

(dosage of a drug/medicine)	(reduction in blood-sugar)
1 mg	x
2 mg	y
3 mg	z
⋮	⋮

→ as X ↑ our Y ↓ (I know that from data)

Confidence Interval

disb^n $\longleftarrow X$: heights

$\{x_1, x_2, \text{------} x_{10}\}$ — random sample from X of size 10

Q) estimate the population mean of $X = \mu$

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

pop-mean

sample-mean

$\rightarrow$ as $n\uparrow$, our $(\bar{x} \longrightarrow \mu)$

$\mu = \bar{x}$

$\rightarrow$ point estimation

$\Longrightarrow$ $\{x_1, x_2, \text{----} x_{10}\}$

$\{180, 162, 158, 172, 168, 150, 171, 183, 165, 176\}$ $\rightarrow$ height of ppl in 'cm'.

Point estimate of $\mu = \frac{1}{10}\sum_{i=1}^{10} x_i = 168.5\text{cm}$

If I can say (assume)

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

pop-mean

$\rightarrow$ interval

$\rightarrow$ confidence

$\rightarrow$ This is a Confidence-Interval statement.

$\rightarrow$ much more meaningful interms of point estimation.