Analysis

- Descriptive
- Diagnostic
- Predictive
- Prescriptive

02) Descriptive statistics

Measures of central tendency:

Mean $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Median: Middle value → Firstly it sorts the data and then compute the median

→ n is odd = $\left(\frac{n+1}{2}\right)^{th}$ item → Middle one

→ n is even = $\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$ item.

↓

avg of two middle values.

→ Mode: Most frequently occuring value.

Program:

```
s1 = pd.Series([2, 1, 0, 4, 3, 15])
s1

s1.mean()
=> 4.1666

s2 = pd.Series([2, 1, 3, 5, 0])
s2
s2.median()
=> 2
```

```
s3 = pd.Series([3, 3, 1, 0, 2, 4, 1])
s3.mode()
```

=> 3, 1 → multimodal data => coz these are 2 no.s which are repeating the most in same no.

```
df = pd.read_excel('ABC attrition data.xlsx', sheet_name =
                   'Employee data class training')
df.head()
df.shape()
df.columns()
```

Task1- Get average & median distance from home :

```
df['Distance From Home'].mean()
df['Distance From Home'].median()
```

Task 2 - Which Dept has max no. of employees ? & how many?

```
df['Department'].value_counts()
```

=> Reasearch & Development 333
Sales 153
Human Resources 14

```
df['Department'].mode()
```

=> Research & Development.

Measures of Variability / Dispersion:

- Range: Largest obs - Smallest obs
- Variance: How far each observation is from mean.
  These differences from the mean are called deviations
- The avg squared deviation from mean is called variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

- Standard Deviation: Sq. root of variance.

Task 3: Get avg, median, min, max, range, var & std.dev for the monthly income.

```
print('Mean = ', df.MonthlyIncome.mean())
print('Median = ', df.MonthlyIncome.median())
print('Minimum = ', df.MonthlyIncome.min())
print('Maximum = ', df.MonthlyIncome.max())
print('Range = ', df.MonthlyIncome.max() - df.MonthlyIncome.min())
print('Variance = ', df.MonthlyIncome.var())
print('Std.dev = ', df.MonthlyIncome.std())
```

Ex: Stock A = annual return 15%
Std deviation 30%
Stock B = annual return 12%
Std deviation 8%
Which one is better for risk prevention.

=) Ans Stock B.

Stock A → 15% → −30% to +30% of 15%

15% + 30% = 45% profit.

15% − 30% = −15% → loss.

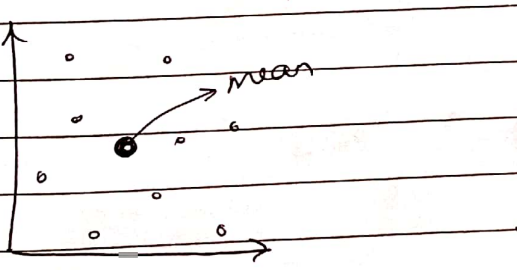
There is 45% of profit but at the same time 15% of loss.

Stock B → 12% → −8% to +8% of 12%.

12% + 8% = 20% profit.

12% − 8% = 4% profit.

At 8% deviation also we have max 20% profit & atleast min of 4% profit at the worst case. So stock B would be preferable.



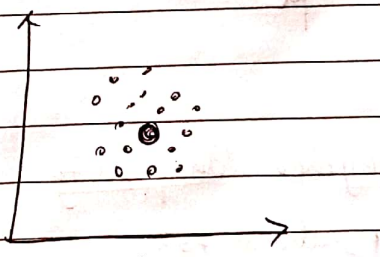
more scatteredness

↓

more deviation

↓

high std deviation



less scatteredness

↓

less deviation

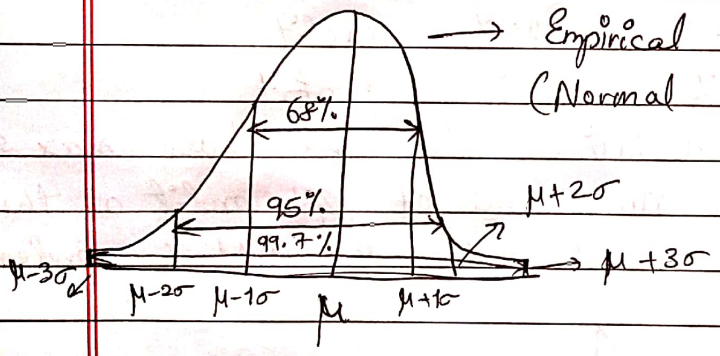
↓

low std deviation

Chebyshev's Inequality

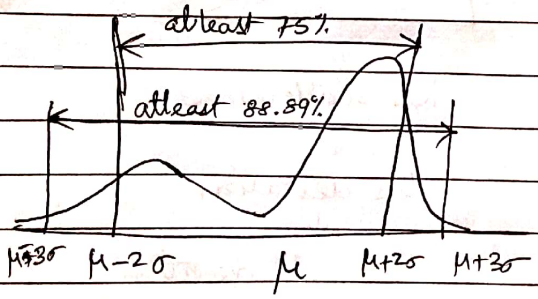
For any data set, it can proved mathematically that
- Atleast 75% of all data will lie in within 2 std dev. of the mean.
- And atleast 89% of within 3 std dev.

```
from IPython.display import Image
Image(filename='ChebyshevInequality.png')
```



→ Empirical Rule (Normal Distribution

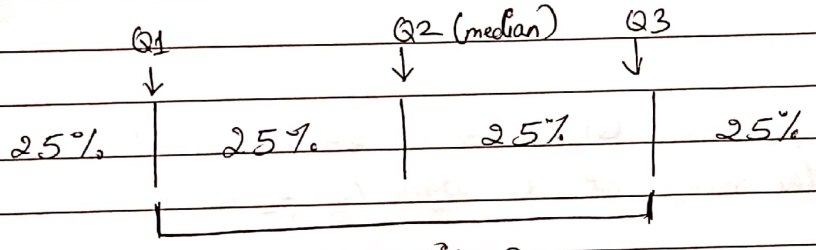
Chebyshev Inequality



Quantiles, Quartiles & Percentiles of data:

- Quantile → certain portion of given data
- Quartile → one portion among four equal divisions of data.
  - 1st quartile => 25% of data (lowest)
  - 2nd quartile => 50% of data (=median)
  - 3rd quartile => 75% of data
- Percentile is one portion among 100 equal divisions of data.
  - P1 to P99.
  - P25 = Q1
  - P50 = Q2
  - P75 = Q3

Q = Quartile



Interquartile Range

IQR

Q3 − Q1

Task 4: What is 25th, 50th & 75th percentile of emp age?

print(df['Age'].quantile(0.25)) → This is Q1
print(df['Age'].quantile(0.50)) → This is Q2
print(df['Age'].quantile(0.75)) → This is Q3

=> 30.0 → 25% of 500 employees = 125 emps are below 30 age
- 36.0 → 50% of 500 emps = 250 emps are within below 36 age
43.0 → 75% of 500 emps = 375 emps are within 43 age

* This clearly indicates that organization has a young pools of employees.

IQR

Task 5: IQR of monthly income.

print ('IQR of income = ', df.Monthly Income.quantile(0.75)
- df.Monthly Income.quantile(0.25))

=> 5841.5

⇓

Q1 of income = 2900.25

Q3 of income = 8741.75

Inference:

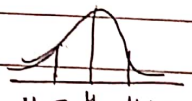
50% of emps (blw Q1 & Q3) = 250 emps have salary in the range of Rs. 2900/- to Rs. 8742/-

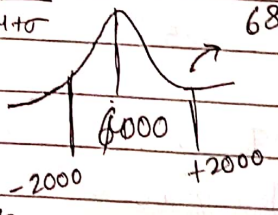
∴ 50% of emps have of salary of variation 5841.5/-

$\mu - \sigma$

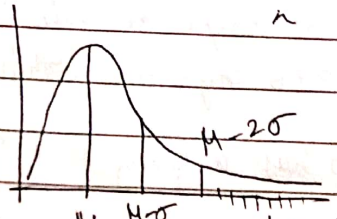
↓ something 6000/- (MI)

↓ something 10000/- (MI)





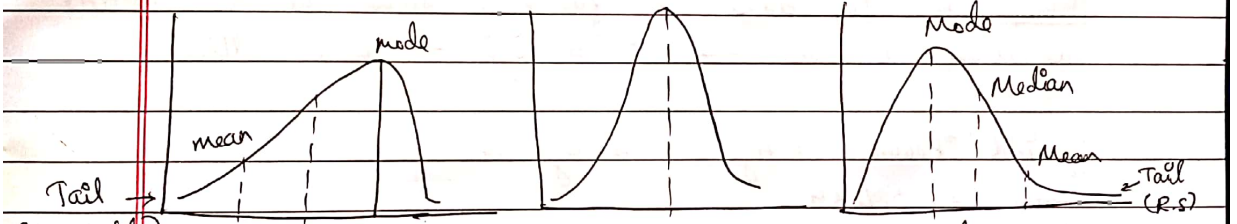
The more the deviation, in salaries the more is difference blw salaries of the employees.



↳ emps with more salaries / less salaries are very few

Skewness :

- Skewness is a measure of symmetry, or more precisely the lack of symmetry.
- A distribution, or data set, is symmetric if it looks like a bell-shaped curve (normal / Gaussian distribution).
- If the skewness is b/w
  - -0.5 to 0.5, the data is nearly symmetrical
  - -1 to -0.5, then negatively skewed.
  - 0.5 to 1, the distribution is positively skewed.
- If the skewness is lower than -1 to greater than 1, the data are ~~exp~~ extremely skewed.



mean < median < mode — negatively skewed.

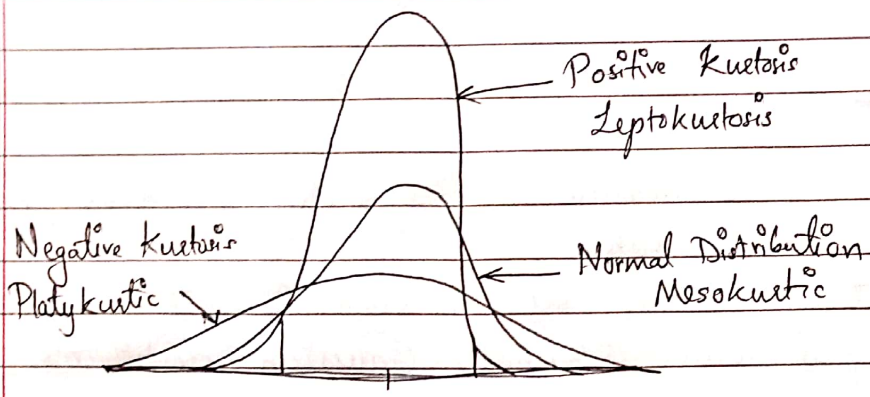
Mean = Median = Mode

mode < median < mean — positively skewed

24/02

Kurtosis :

- Kurtosis is a measure of whether the data is heavy-tailed or light-tailed relative to a normal distribution.
- That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.
- Kurtosis for std normal distribution is 3.
- Leptokurtic → kurtosis > 3
- Platykurtic → kurtosis < 3



Leptokurtosis: Most of the data lies close to the mean, so the deviation is not that much. σ is less.
Mesokurtic: Normal Distribution - The deviation is in control although the sp data is little spread.
Platykurtic: The data lies far away from the mean, means deviation is high → σ is high. → more outliers.

print('Skewness of Age = ', df.Age.skew())
print(' => 0.439.
print('Skewness of MI = ', df.MonthlyIncome.skew())
=> 1.3265
print('kurtosis of Age = ', df.Age.kurtosis())
=> -0.38227
print('kurtosis of MI = ', df.MonthlyIncome.kurtosis())
=> 0.81824

Covariance and Correlation

- Covariance and correlation both primarily assess the relationship between variables.
- Using covariance, we can only gauge the direction of the relationship (whether the variables tend to move in

tandem or show an inverse relationship). However, it does not indicate the strength of the relationship, nor the dependancy between the variables.

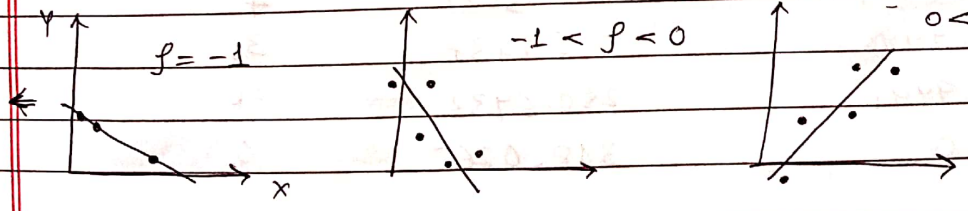
- On the other hand, correlation measures the strength of the relationship between variables. Corelation is the scaled measure of covariance. It is dimensionless. In other words, the correlation coefficient is always a pure value and not measured in any units.

$$Cov(X,Y) = \frac{\Sigma[(x_i - \bar{x})(y_i - \bar{y})]}{n}$$

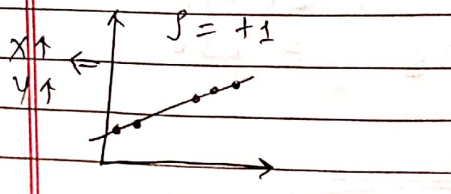
$$\rho = \text{Correlation coefficient} = r = \frac{\Sigma[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

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

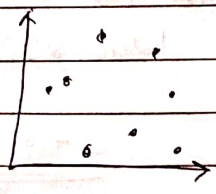
The value of correlation coefficient ranges from -1 to $+1$.



X↑ Y↓



X↑ Y↑



ρ = 0 → No correlation
i.e, X & Y have no relation

Example:

$X = [43, 21, 25, 42, 57, 59]$ = Age of people.
$Y = [99, 65, 79, 75, 87, 81]$ = Glucose level.

X	Y	$X - \bar{x}$	$Y - \bar{y}$
43	99	$43 - 41.1667 = 1.8333$	18
21	65	-20.1667	-16
25	79	-16.1667	-2
42	75	0.8333	-6
57	87	15.8333	6
59	81	17.8333	0
Avg = 41.1667	81		

$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{x})^2$	$(Y - \bar{y})^2$
32.9994	3.3609	324
322.6672	406.6957	256
32.3334	261.3621	4
-4.9998	0.69438	36
94.9998	250.6933	36
0	318.0265	0

$\Sigma (X - \bar{x})(Y - \bar{Y}) = 478$

$\Sigma (X - \bar{x})^2 = 1240.8333$

$\Sigma (Y - \bar{y})^2 = 656$

Correlation coefficient $= \dfrac{478}{\sqrt{1240.8333 \times 656}}$

$= 0.5298$

So, since the correlation coefficient is positive, ∴ we can say that as Age increases, glucose level also increases.

0.5298 lies b/w 0 to +1, some deviation makes sense.

Program:

```
import numpy as np
Age = np.array([43, 21, 25, 42, 57, 59])
GlucoseLevel = np.array([99, 65, 79, 75, 87, 81])

R = np.corrcoef(Age, GlucoseLevel)
print(R)
```

=> `[[1.  0.5298]`
`[0.5298  1.  ]]`

	Age	Glucose
Age	1	0.5298
Glucose	0.5298	1

Correlation coefficient = R[0][1] or R[1][0] = 0.5298

Age - Age = 1
glucose - glucose = 1
Age - glucose = 0.5298

~~25/02~~

Inferential Statistics:

Probability Basics

Tossing two coins = HH, HT, TH, TT

Probability distribution. → Considering getting a tail.

0 Tail → 1/4 → HH
1 Tail → 2/4 → TH, HT
2 Tails → 1/4 → TT

x_i	0	1	2
$P_i(X = x_i)$	1/4	2/4	1/4

$\sum P_i = 1$

PMF - Probability Mass Function → Discrete → 0, 1, 2
PDF - Probability Density Function → Continuous → 0.5, 0.8, 0.333, 1.53 etc

Few rules of Probability:

1. Complementary event of A' of A
   $P(A') = 1 - P(A)$ → Toxing a coin

2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

3. When two events are mutually exclusive / disjoint then $P(A \cap B) = 0$

4. Multiplication theorem: If A & B are independent events then $P(A \cap B) = P(A)\,P(B)$

Mutually exclusive / disjoint
Rolling a dice
A: Getting an even no. → {2, 4, 6}
B: Getting an odd no. → {1, 3, 5}
When event A occurs, B does not. No commonality.
→ $P(A \cap B) = 0$ → no intersection

Mutually inclusive

A: Getting an even no. → {2, 4, 6}
B: Getting a multiple of 3 → {3, 6}
6 is common, when 6 occurs, both the events are said to satisfied.

25/02

Joint Probability:

- It is a statistical measure that calculates the likelihood of two events occuring together and at the same point of time.
- Ex: From 52 cards, the joint probability of picking up a card that is both red & 6 is $P(6 \cap red) = \frac{2}{52} = \frac{1}{26}$
Since it contains 2 red sixes
heart ^ diamond

$$P(6 \cap red) = P(6) \times P(red) = \frac{4}{52} \times \frac{26}{52} = \frac{1}{26}$$

Marginal Probability:

- It is the probability of an event irrespective of the outcome of another variable.

Conditional Probability:

- The probability of event A given that event B has already occured is $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Ex: U have studied everyday for 5 hrs., what can be the probability of getting 90%.

Example: Consider a random experiment of rolling a dice, what is the probability of getting an even number or a no. divisible by 3?

=> A: {2, 4, 6} S = {1, 2, 3, 4, 5, 6}
B: {3, 6}

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= \frac{3}{6} + \frac{2}{6} - \frac{1}{6}$$
$$= \frac{4}{6} = \frac{2}{3}$$ //

(2) A : odd no. $= \{1, 3, 5\}$
B : $<= 3$ $= \{1, 2, 3\}$
Then what is the probability A given B, $P(A|B)$?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/6}{3/6} = \frac{2}{3}$$

(3) What is the probability a randomly selected person is male, given that they own a pet?

	Have Pets	Dont have pet	Total
Male	0.41	0.08	0.49
Female	0.45	0.06	0.51
Total	0.86	0.14	1

A : Male
B : own a pet

$\sim A$ = Female
$\sim B$ = Do not own a pet } negation

$P(A \cap B) = 0.41$
$P(\sim A \cap B) = 0.45$
$P(A \cap \sim B) = 0.08$
$P(A \cap \sim B) = 0.14$

$P(A) = 0.49$
$P(B) = 0.86$
$P(\sim A) = 0.51$
$P(\sim B) = 0.14$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.41}{0.86} = 0.4777$$

UsedCarsPrice

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

cars_data = pd.read_csv('UsedCarsPrice.csv')
cars_data.head()

cars_data.shape
=> (1436, 11)

cars_data.info()

cars_data.isna().sum()
```

When data has some missing values but that can be only read & interpreted by humans, like ?? or ????, python only fills blanks with NaN not these type of missing values (??, ????). So we need to tell python that these are also the missing values. We have ?? in KM column so that it is detected as object type but not an integer, and no. of missing values in KM column came out to be 0, as ?? was treated as a string/object type.

∴

```
cars_data = pd.read_csv('UsedCarsPrice.csv', index_col=0,
                        na_values = ["??", "????"])
```

Now the KM and HP became float. And missing values count for KM = 15 & HP = 6 was made.

Frequency Table using crosstab()

- To compute a simple cross tabulation of one or more factors.
- By default, computes a frequency table of factors.

```
pd.crosstab(index = cars_data['FuelType'], columns = 'count',
                                    dropna = True)
```

↓

no. of cars having various types.

FuelType	col_0 count
CNG	15
Diesel	144
Petrol	1177

Two way tables

```
pd.crosstab(index = cars_data['Automatic'],
            columns = cars_data['FuelType'], dropna=True)
```

values of Automatic	FuelType	CNG	Diesel	Petrol
No	← 0	15	144	1104
Yes	← 1	0	0	73

CNG → with manual gears = 15
Diesel → " " " = 144
Petrol → " " " = 1104
Petrol → with automatic gears = 73

Two way table with joint probability

```
pd.crosstab(index = cars_data['Automatic'],
            columns = cars_data['FuelType'],
            normalize = True,
            dropna = True)
```

FuelType	CNG	Diesel	Petrol
Automatic			
0	0.011228	0.107784	0.826347
1	0.000000	0.000000	0.054641

Interpretation:

- Probability of a car with manual gearbox & fueltype CNG is 0.011228.
- Probability of a car with automatic gearbox & fuel type petrol is 0.054641

⇓

What % of probability of petrol cars having automatic gear?

Two way table with Marginal probability

```
pd.crosstab(index = cars_data['Automatic'],
            columns = cars_data['FuelType'],
            normalize = True,
            margins = True,
            dropna = True)
```

FuelType	CNG	Diesel	Petrol	All
Automatic				
0	0.011228	0.107784	0.862347	0.945359
1	0.000000	0.000000	0.056641	0.056641
All	0.011228	0.107784	0.880988	1.000000

Interpretation:

94.5% → Manual gears
5.5% → Automatic gears.

1.1% → CNG cars
10.7% → Diesel cars
88.1% → Petrol cars

Two way table with conditional probability:

- Given the type of gear box, compute the probability of different fuel type → prob. of petrol car → automatic gear.

```
pd.crosstab(index = cars_data['Automatic'],
            columns = cars_data['FuelType'],
            normalize = 'index',
            margins = True,
            dropna = True)
```

FuelType	CNG	Diesel	Petrol
Automatic			
0	0.011876	0.114014	0.874109
1	0.000000	0.000000	1.000000
All	0.011228	0.107784	0.880988

Interpretation:

- Gearbox = manual
  * probability of car being CNG FT = 0.011876
  * " " " " Petrol " = 0.874109
- Gearbox = Automatic
  * Probability of car being petrol FuelType = 1.

Q. Given the fuelType, compute the probability of diff. gear box.

```
pd.crosstab(index = cars_data['Automatic'],
            columns = cars_data['FuelType'],
            normalize = 'columns',
            margins = True,
            dropna = True)
```

FuelType	CNG	Diesel	Petrol	All
Automatic				
0	1.0	1.0	0.937978	0.945359
1	0.0	0.0	0.062022	0.054641

Interpretation:

- CNG
  - probability of car being automatic is 0
  - " " " " manual is 1
- Diesel → same
- Petrol
  - auto → 0.054641 = 5.46%
  - manual → 0.945359 = 94.53%

Probability Distribution :

* Discrete distribution
  - Bernoulli
  - Binomial
  - Poisson
* Continuous distribution
  - Normal
  - Unifor
  - Exponential

Binomial distribution :

- It is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N.
- p = probability of success n = no. of trials
  $q = 1-p$ or failure
  x = successes
- Each success/failure experiment → Bernoulli trial
  $n = 1$.
- PMF is given by: of $B(n, p)$
  $$P_x = {}^nC_x p^x q^{n-x}$$
- Mean of Binomial distribution is np, variance is npq
  $SD = \sqrt{npq}$

from scipy.stats import binom

Q. What is the probability of getting exactly 0 heads?

=> Here $n=6$, $p = q = \frac{1}{2}$ and $x=0$

We need to find $P(x=0)$ and $P(x \leq 2)$

binom.pmf(0, 6, 0.5)

=> 0.015625 0

binom.pmf(x, n, p)

x = no. of success

n = no. of trial

p = probability of success

Q. Probability of getting <= 2 heads.

=> Here $n=6$, $p=1/2$

$P(x <= 2) = P(x=0) + P(x=1) + P(x=2)$

binom.pmf(0, 6, 0.5) + binom.pmf(1, 6, 0.5) + binom.pmf(2, 6, 0.5)

=> 0.343750

Alternative way:

Cumulative probability:

$P(x \leq 2) = 1 - P(x > 2)$ → Cummulative dist^n function

binom.cdf(2, 6, 0.5)

↓

<= 2

Tasks :

- There is road junction at which the drivers stops and seek for directions. The probability for seeking direction is 0.45. On a given day, if 200 vehicles pass, then

i) prob. for exactly 100 drivers seeking help
ii) atleast 50 drivers
iii) at most 20 drivers

=> i) $P(x=100) = {}^{200}C_{100} (0.45)^{100} \times (0.55)^{200-100}$ (p, q)

binom.pmf (100, 200, 0.45)
=> 0.020625

ii) $P(X \geq 50) = 1 - P(X < 50)$
$= 1 - \text{binom.cdf}(49, 200, 0.45)$
$\Rightarrow 0.999999 \simeq 1$

iii) $P(X \leq 20) = \text{binom.cdf}(20, 200, 0.45)$
$\Rightarrow 3.978917e-27$
$\Rightarrow 3.97 \times 10^{-27}$

```
import pandas as pd.

camp = pd.read_excel('ProbDist_Data.xlsx', sheet_name = 'binom')

camp.shape
=> (297, 5)
```

camp['Campaign_Response'].value_counts()

=> False 248 ⟶ 248/297 } probability of response
True 49 ⟶ 49/297 }

To know the probability by coding.

↓

camp['Campaign_Response'].value_counts(normalize=True)

=> False 0.835017 => 83.5%
True 0.164983 => 16.49% ≃ 16.5%

Q. What is the probability that upto 15 customers will respond to campaign out of 150 randomly selected ~~co~~ customers?

=> P(x <= 15)

n = 150, p = 0.165, P(x <= 15) = ?

binom.cdf(15, 150, 0.165)

=> 0.016547

Q. What is probability that between 15 to 20 customers will respond to campaign out of 150 randomly selected customers?

=> P(15 ≤ x ≤ 20)

=> P(0 ≤ x ≤ 20) − P(0 ≤ x ≤ 15) ↗ P(0 ≤ x <= 14)

=> binom.cdf(20, 150, 0.165) − binom.cdf(14, 150, 0.165)

=> 0.1669922 ≃ 16.7%

Poisson Distribution:

$$P(x=x) = \frac{e^{-\lambda} \lambda^{x}}{x!} \quad \text{for all } x = 0, 1, 2, 3, 4, \ldots$$

Mean = variance = λ

- It is a discrete probability distribution that expresses the prob. of a given no. of events occuring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.
- The no. of successes(p) are very small compared to the no. of trials(n), then we go for Poisson distribution.

Q. Find the probability of exactly 4 accidents in a month, given that average is 5 accidents/month.

-> Here $\lambda = 5$, we need to compute $P(x=4)$

```
from scipy.stats import poisson
poisson.pmf(4,5)
```

=> 0.17547 = 17.55%

```
poisson.pmf(x, λ)
```

Q. What is the probability of 3 or lesser accidents in a moth

=> $P(x <= 3)$

~~binom~~ poisson.cdf(3, 5)

= 0.265 -> 26.5%

Q. 5 or more accidents

$P(x >= 5)$

=> $1 - P(X < 5)$

=> 1 - poisson.cdf (4, 5)

=> 0.5595

cart = pd.read_excel('ProbDist_Data.xlsx', sheet_name='poisson', usecols = [0])

Q. What is probability of seeing atleast 2 items in cart?

=> $P(X >= 2)$

$= 1 - P(X < 2)$

$= 1 - \{P(X=0) + P(X=1)\}$

$= 1 - \left\{ \frac{e^{-1.4489} (1.4498)^0}{0!} + \frac{e^{-1.4489} (1.4489)^1}{1!} \right\}$

BTW 1.4489 = Mean of carts addition.

or in coding:

=> 1 - poisson.cdf (1, 1.4489)

=> 0.4249

Q. min 6 items, max 9 items.

=> poisson.cdf (9, 1.45) - poisson.cdf (5, 1.45)

=> 0.00379

Normal Distribution

- It is a continuous distribution.
- The probability density function of Normal distribution with parameters μ and σ is given by -

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Q. Let the average sales of a particular product is 10000 and std dev. is 2400. Then what is the probability of getting more than 12000 sales?

=> $\mu = 10000$

$\sigma = 2400$

$P(x \geq 12000) = ?$

$X \sim N(10000, 2400)$

$Z = \frac{X - \mu}{\sigma} \sim N(0, 1) \rightarrow$ standard N.D.

$$Z = \frac{12000 - 10000}{2400} = \frac{2000}{2400} = \frac{10}{12} = \frac{5}{6} = 0.8333$$

$$P(X \geq 12000) \approx P(Z \geq 0.8333)$$
$$\approx P(Z \leq 0.8333)$$

Google for **Std. normal table:**

Under the z in the at 0.8 and at column 0.03 we have 0.7967

$\therefore 1 - 0.7967 = 0.2033$

Therefore, chances of getting ~~sales~~ more than 12000 products sales is just 20.3%.

By coding:

```
from scipy.stats import norm
1 - norm.cdf(12000, 10000, 2400)
```

=> 0.2023

Business inference:

When we have avg sales of 10000 with std dev of 2400, then the chance of having more than 12000 sales is around 20.23%.

Q. What is the probability of getting fewer than 9200 sales?

=> $P(x <= 9200)$

$\mu = 10000$

$\sigma = 2400$

$z = \frac{9200 - 10000}{2400} = -0.333$

According to the table, from 0.3 to 0.03 we have intersection is 0.3745.

By coding:

```
norm.cdf(9200, 10000, 2400)
```

=> 0.37

Risk management using normal distribution:

Q. Consider a stock with returns mean = 10, std = 5. What are the chances that the returns will be < 0?

=> norm.cdf (0, 10, 5)

=> 0.02275 → 2.3%

Q. What are the chances of losing money when std = 10%?

=> norm.cdf (0, 10, 10)

=> 0.15865 → 15.8%

So as std. dev increases, risk of loosing money also increases

Q. Mean A monthly balance in the bank account of credit card holders is assumed to be normally distributed with mean 500 USD and variance 100 USD.

→ Prob. of balance more than 513.5 USD?

→ If there are 1000 customers,

- how many customers have balance > 513.5 USD
- " " " " " < 520 USD

=> mean = 500

variance
~~SD~~ = 100

SD = 10

1 − norm.cdf (513.5, 500, 10)

=> 8.85%

N = 1000

round (N * (1 - norm.cdf (513.5, 500, 10)), 0)

=> 89.0 → 89 people have acct. bal. >= 513.5 USD

It is rounded because, we are find no. of people.

round (N * (norm.cdf (520, 500, 10)), 0)

=> 977.0 → 977 people have acct. bal. <= 520 USD

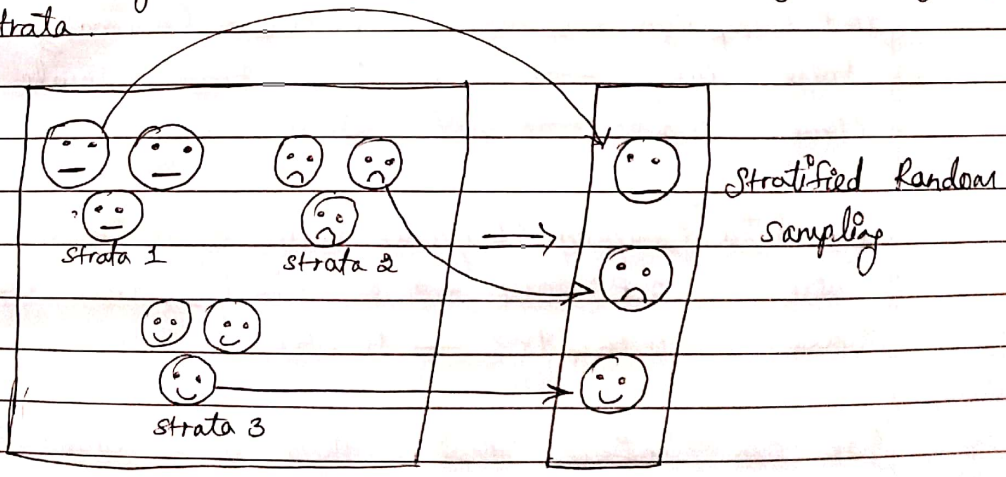
Sampling

i) **Random Sampling:**

Randomly selected items with no repetition, which is called random sample.

ii) **Stratified Sampling:**

- Selecting the samples with some strategy.
- Involves dividing the entire population into homogenous groups called strata.



```
import pandas as pd

data = pd.read_excel ('Credit Card Data.xlsx')

data.shape
(297, 5)
```

Random:

```
df1 = data.sample (n=5, random_state = 44)
df1
```

↓ size of 5

If we remove random_state parameter, then everytime we execute we get randomly selected samples from the dataset. To keep the first random samples, we pass some value to the random_state, so that for analysis purpose it makes easier.

```
df1.Campaign_Response.value_counts (normalize = True)
```

=> False 0.6 → 60% } → From sample
True 0.4 → 40% }

```
df1 data.Campaign_Response.value counts (normalize = True)
False 0.835017 → 83.5% } → From Dataset
True 0.164983 → 16.5% }
```

We can infere that there is a huge difference b/w the acceptance of campaign (Response) from both the sample & original dataset.

Therefore the randomly selected sample is not true representative of the original dataset.

```
df1 = data.sample(n = 100, random_state = 44)
```
↓ size of 100

```
df1.Campaign_Response.value_counts(normalize = True)
```

=> False 0.86 → 86% } almost near to the original dataset.
True 0.14 → 14%

Therefore the bigger sample size, the bigger is similarity, in case of random sampling.

Task: Select 10% of records randomly, & create a dataframe df2

```
df2 = data.sample(frac = 0.1, random_state = 44)
df2.shape
```
=> (30, 5)

```
df2.Campaign_Response.value_counts(normalize = True)
```
=> False 0.8 → 80%
True 0.2 → 20%

Stratified:

```
from sklearn.model_selection import train_test_split
```

For 30% of records

df1, df2 = train_test_split(data, test_size = 0.3, stratify = data['Gender'], random_state = 44)

Train

df1.shape[0] / data.shape[0]

=> 0.6969

Test

df2.shape[0] / data.shape[0]

=> 0.3030

df2['Gender'].value_counts(normalize = True)

=> M 0.5777

F 0.4222

df1.Gender.value_counts(normalize = True)

=> M 0.570048

F 0.429952

Hypothesis Testing

- Claim made by a person / organisation
- Claim is usually about the population parameters such as mean or proportion & we seek evidence from a sample for the support of the claim.

i) Null hypothesis → H_0

ii) Alternate hypothesis → H_a / H_1

In Hypothesis testing we either accept one hypothesis
Ex: Null / Alternate ~~and~~ or we reject one hypothesis
ex: Alternate / Null.

Type I error

		Actual: True	Actual: False
Observed / tested / predicted (Null Hyp.)	True	True positive	False +ve
	False	Falsely Negative	True -ve

True positive & True negative are correct.
And False negative & False positive are wrong → Errors

- Type I error is conditional probability of rejecting a null hypothesis when it is true, → False positive, negative.
- α, the significance level is value of type I error.
- $P(\text{reject null hypothesis} \mid H_0 \text{ is true}) = \alpha$.

Type II error

- Retaining a null hypothesis when it is true, is called Type II error or False positive.
- $P(\text{Retain null hyp} \mid H_0 \text{ is false}) = \beta$

Power of test

- $(1-\beta)$ is known as power of test
- It is $P(\text{Reject Null hyp.} \mid H_0 \text{ is false}) = 1-\beta$

		Actual H_0 True	Actual H_0 False
Test	True	TP	FP
	False	FN	TN

If FP ↑, FN ↓

If FP ↓, FN ↑

Examples for False +ve & False -ve:

i) Biometrics: If a person is actually an employee but due to some issue if biometrics fails to identify him as an employee, instead given unknown as a result (False Negative FN), then it is OK. But if the so person if not an emp but the test shows him as an authorized person, then it is very problematic situation (FP).

ii) Covid test: If a person is actually not a covid attacked person but the test shows positive, then it is fine to test him again (FP). But if the person is actually affected and is tested negative (FN) then it is dangerous.

Steps involved in solving H.T:

Step 1: Define a Null hypothesis & alternate hypothesis:
- N.H. → No relationship / False claim
- A.H. → There is a relationship / True claim

Step 2: Decide the significance level:

- Controlling the Type I error by determining the risk level, α, the level of significance that you are willing to reject the null hypothesis when it is true. You select a level of 0.01, 0.05 or 0.10.
  - ↓ risk → 1% confidence, 99% confidence
  - → 5% risk, 95% confidence.

Step 3: Identify the test statistic:
- The test statistic will depend on the probability distribution of the sampling distribution.

Step 4: Calculate the p-value or critical values:
- P-value is the conditional probability of observing the test statistic value or extreme than the sample result when the null hypothesis is true.
- Critical value approach:
  Critical values for the appropriate test statistic are selected so that the rejection region contains a total area of α when H_0 is true and the non-rejection region contains a total area of $1-\alpha$ when H_0 is true.

Step 5: Decide to reject or accept null hypothesis:
- Reject N.H. when test statistic lies in the rejection region; retain N.H. otherwise.
- Reject N.H. when p-value $< \alpha$; retain N.H. otherwise.

One sample Z Test:

- Population mean & std dev is known, population is normal (or approx), $N >= 30$ (population size).

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{(n)}}}$$

```
import pandas as pd
import numpy as np

data = pd.excel ('Hypothesis Testing - xlsx', sheet_name = 'One Sample z')
                                                    .iloc [:, 0]
```

Q. A random sample of 35 young adult men was sampled. Each person was asked how many minutes of sports he watched on TV daily. The responses are listed. If is is known that $\sigma = 10$. At 5% significance level, can we conclude that mean amount of television watched daily by all young adult men is greater than 50 minutes?

z → >= 30
t → < N=30

=> • A random sample of 35 young adult men was sampled
• Each person was asked how many mins of sports he watched daily. It is known that $\sigma = 10$.

$N = 35$, pop std = 10, pop mean = 50, $\alpha = 0.05$

Steps:

H_0 : Average amount of daily television minutes watched by young adult men is 50. ($\bar{x} = \mu$)

H_1 : Average amount of daily TV minutes by young men > 50. ($\bar{x} > 50$)

=> $H_0 : \mu = 50$

$H_1 : \mu > 50$

from statsmodels.stats import weightstats as test

test.ztest(data, value = 50, alternate = 'larger') → Right tailed test

To know more ztest function => ?test.ztest

=> (0.841404 , 0.2000609)

↓ z-statistic

$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

↓ p-value

p-value > 0.05

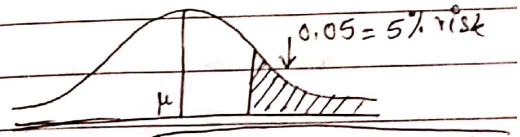
↓

Accept the null hypothesis.

i.e, we don't have enough to reject the N.H.

↓

Avg minutes of watching sports daily is 50.



larger = right tailed
lesser = left tailed
not equal = Two-tailed.

One Sample t test

data = pd.read_excel('Hypothesis Testing.xlsx', sheet_name='One sample T').iloc[:,0]

N = 12 => < 30

$\mu = 6$ $\alpha = 0.05$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Defining Null & A.H.

H_0 : Avg delivery time is 6 hours

H_1 : Avg delivery time is less than 6 hours

lesser → left-tailed

from scipy.stats import ttest_1samp

ttest_1samp(data, 6) # pass data & pop.mean as parameters

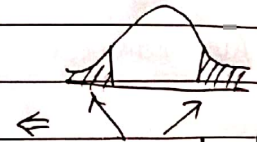
=> ~~(Statistic = 35.8053, pvalue = 1.3997608e-28)~~

(statistic = -0.68498, pvalue = 0.507529)

↓

By-default in t-test, we have risk at both sides

We want only left side, so divide it by 2.



both together makes p = 0.5075

p-value = 0.5075/2
= 0.25375

p-value > 0.05
→ So we accept the N.H.
→ i.e, Average delivery time ≥ 6 hours.

NOTE: Null & Alternate hypothesis always deal with population not samples.

Two Sample Tests

Population 1	Population 2
Parameters : μ_1 & σ_1^2	Parameters : μ_2 & σ_2^2
Sample size : n_1	Sample size : n_2
Statistics : $\bar{x}_1$ and s_1^2	Statistics : $\bar{x}_2$ and s_2^2

```
two = pd.read_excel('Hypothesis Testing.xlsx', sheet_name='Two sample
                    z').iloc[:,0:2]
```

Q. Recent studies seem to indicate that using a cell phone while driving is dangerous.
One reason for this is that a drivers reaction times may slow, while he is talking on the phone. Researchers at Ohio university measured the reaction times of a sample of drivers who owned a car, phone. Half the sample was tested while on phone and other half not on the phone. Can we conclude the reaction times are slower for drivers using cell phone?

from statsmodel.stats import weightstats as test

Define the hypothesis:

- N.H.: Reaction time is same for drivers who use cell phone compared to who don't.
- A.H.: Reaction is more for drivers who use cellphones compared to who don't.

test.ztest(x1 = two['Phone'], x2 = two['Not'], alternative = 'larger')

=> (7.066852, 7.923633e-13)

p-value = 7.9236 e-13

→ p-value < 0.05 (α)

→ Reject Null Hypothesis.

Two Sample t test:

Q. A no. of restaurants feature a device that allows credit card users to swipe thier cards at the table. It allows the users to specify a % or a $ amount to leave as a tip. In an experiment to see how it works, a random sample of credit card users was drawn. Some paid the usual way, and some used the new device. The % left as a tip was recorded and listed. Can we infer that users of the device leave larger tips?

```
df = pd.read_excel('Hypothesis Testing.xlsx', sheet_name = 'Two sample t')
                                                .iloc[0:, 0:2]
```

Define Hypothesis:

- N.H.: Percentage of tip paid are same.
- A.H.: Percentage of tip paid by device are higher.

H0:

$n_1 = 10$ → usual

$n_2 = 11$ → device.

```
from scipy.stats import ttest_ind
ttest_ind(df['Usual'].dropna(), df['Device'], equal_var = True)
```

(equal_var = True: assuming or samples having equal variance)

=> statistics = 2.173009, pvalue = 0.042634

⇓

two-tailed

⇓

we want one-tailed (right-tailed)

∴ pvalue = 0.042634/2 = 0.0213

→ 0.0213 < 0.05 => pvalue < α

→ N.H. is rejected.

→ The Device is getting more tip.

↳ Tips paid by people with cash / usual payment method & through device are not same.

Non parametric : Chi square test of association / independence:

Understanding the relationship between Gender & Campaign Response

data.Campaign_Response.value_counts(normalize = True)
=> False 0.835 → 83.5%
True 0.165 → 16.5%

Null: There is no relationship between gender and campaign response (independent)
Alternate: There is relationship between gender & campaign response (dependent)

obs = pd.crosstab(data['Gender'], data['Campaign_Response'])

Campaign_Response	False	True
Gender		
F	102	25
M	146	24

```
from scipy.stats import chi2_contingency, chisquare
chi2, chi_sq_stat, p_value, deg_freedom,
exp_freq = chi2_contingency(obs)

print('Chi-sq Statistic %3.5f p value %1.6f Degrees
      of freedom %d'
      % (chi_sq_stat, p_value, deg_freedom))
```

=> Chi-sq Statistic 1.25639 p value 0.262335
Degrees of freedom 1

Degrees of freedom:

(no. of rows − 1) * (no. of columns − 1)

⇓

(2 − 1) (2 − 1) = 1

p-value = 0.2623

p-value > 0.05 (α)

N.H. is accepted.

=> There is no relationship between gender & campaign response.