* Discrete random variable

Discrete distribution	Continuous distribution
Binomial	normal
poisson	gamma
	rectangular
	exponential
	Beta

* Binomial distribution

→ no of events - independent & finite.

$p \rightarrow$ success

$q \rightarrow$ failure

$$P(X=r) = \begin{cases} {}^{n}C_r \, p^r q^{n-r} & ; r = 0,1,2 \ldots n ; q = 1-p \\ 0 & ; r \neq 0,1,2,3,\ldots n. \end{cases}$$

$n, p, q \rightarrow$ parameters of distribution

binomial expansion of $N \cdot (q+p)^n$.

(Binomial freq distribution)

$$P(X=r) = \begin{cases} {}^{n}C_r \, p^r q^{n-r} & ; r = 0,1,2 \ldots n ; q = 1-p \\ 0 & \text{otherwise} \end{cases}$$

$\mu = np$

$\sigma^2 = npq$

↳ gives spread of data

exercise:- Using binomial; if a perfectly cubical die is thrown a large number of times in sets of 8. The occurrence of 5 or 6 is success. In what proportion of the sets do you expect 3 success?

$p = 1/3$ {prob of getting success}

$q = 2/3$ {prob of fail}

$n = 8$ {process repeated 8 times. hence occurrence of an event = 8}

$P(X = 3) = {}^{8}C_r \, p^r q^{8-r}$ → find for $r = 3$

~~$P(X=5) = 8C_5 (1/3)^5 (2/3)^3$~~

~~$(X=6) = 8C_6 (1/3)^6 (2/3)^2$~~

$$P(X=3) = 8C_3 (1/3)^3 (2/3)^5 = \underline{0.2731}$$

exercise:- 6 dice are thrown 729 times. How many times do you expect, 3 dice to show a 5 or 6?

at least

$p = 1/3$

$q = 2/3$

$n = 729$

$$P(x \text{ atleast } 3) = \sum_{r=3}^{6} nCr\, p^r q^{n-r}$$

$$= 729C_3 (1/3)^3 (2/3)^{726} + 729C_4 (1/3)^4 (2/3)^{725} + 729C_5 (1/3)^5 (2/3)^{724} + 729C_6 (1/3)^6 (2/3)^{723}$$

$\frac{233}{729}$ //

exercise assuming half the population are consumers of chocolate so that chance of individual being a consumer is $1/2$; and assuming 100 investigators are used to see if they consume chocolate how many investigators would you expect to report 3 people or less were consumers

exercise in a sampling a large number of parts manufactured by a machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples how many would be expected to contain

a) at least 3 defective parts

b) none defective

* Poisson distribution

- used when p, q very small; n very large.

$$P(X = r) = \begin{cases} \frac{e^{-m} m^r}{r!} & r = 0, 1, 2 \dots \\ 0 & \text{otherwise.} \end{cases}$$

↳ success

10-3-25

exercise

Calculate:- $e^{-0.5} = 0.61$

deaths	0	1	2	3	4
freq	122	60	15	2	1

find theoretical frequency if $e^{-0.5} = 0.61$

total freq $\Rightarrow$ 122+60+15+2+1 = 200

$$\text{mean} = \frac{\Sigma f x}{N} = \frac{(122 \times 0) + (1 \times 60) + (2 \times 15) + (2 \times 3) + (4 \times 1)}{200}$$

$\Rightarrow 0.5$

$e^{-\text{mean}} = e^{-0.5} = 0.61$ approx

$$N e^{-m} \frac{m^r}{r!} = 200 \times (0.61) \times \frac{(0.5)^r}{r!}$$

when $r = 0 \Rightarrow 200 \times (0.61) \times \frac{(0.5)^0}{0!} \Rightarrow 122$

when $r = 1 \Rightarrow 200 \times (0.60) \times \frac{0.5^1}{1!} = 61$

$r = 2 \Rightarrow 15$

$r = 3 \Rightarrow 2$

$r = 4 \Rightarrow 0$

→ expected frequency

Σ expected must be equal to Σ actual.

mean

find theoretical freq

x	0	1	2	3	4
f	192	100	24	3	1

tot freq = 320

mean = 0.5

$e^{-0.5} \approx 0.61$

poisson → $N e^{-0.5} \cdot \frac{(0.5)^x}{x!}$

	expected
0 →	195
1 →	98 → 97
2 →	24
3 →	4
4 →	0.5 → 1
	322

Continuous random distribution

- Uniform random distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & ; a \le x \le b \\ 0 & ; \text{otherwise} \end{cases}$$

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

- normal / gaussian distribution / mandate distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-1/2\left(\frac{x-\mu_x}{\sigma_x}\right)^2}$$

$-\infty < x < \infty$

- exponential random variable

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

* Hypothesis testing
  - null hypothesis (H_0) → basic → tested for possible rejection under the assumption it is true
  - alternate hypothesis / composite hypothesis / complementary hyp..
    - when H_0 is rejected, H_1 is accepted.

before starting the exp. we are assuming default mean value μ_0.

Null hypothesis is given as $H_0 : \mu = \mu_0$

Experiment starts :

Case 1 → in some experiments mean can be $> \mu_0$ or $< \mu_0$ ⇒ 2 tailed alternative

Case 2 → After exp. we found $\mu > \mu_0$ ⇒ right tailed alternative.

Case 3 → " $\mu < \mu_0$ ⇒ left tailed alternative

error types:

- reject null hypothesis : even if true; → first kind / type 1 error → α
- reject null hypothesis, though false (or to be rejected) → 2nd kind / type 2 error → β

β = P(accept H_0 when false) = P(accept H_0 / H_1)

α = P(reject H_0 when it is true) = P(" H_0 / H_0)

There are many ways to test the hypothesis

1) T-test → student T test
2) P value
3) Z value.

* T-test

tests if there is significant difference btwn means of two groups

The t tests is performed on different samples in population. A statistical variation (in terms of mean, median, mode) is found out btwn these samples

1 sample t

↳ eg:- a manufacturer of choc bars claims bars weigh 50g avg. To verify, sample of 30 is taken but mean is 48g

independent sample t

↳ to test effectives of drug, 60 subjects are divided randomly
1st group → drug A
2nd group → drug B
We can independently test if significant changes present in groups for drugs.

paired sample t

↳ how effective diet plan is. weigh 30 people before, weigh same 30 people after

Hypothesis for 1 sample ; $t = \frac{\bar{x} - \mu}{\sqrt{\frac{x_i - \bar{x}}{n}} / \sqrt{n}}$ ← reference value; ← std deviation; ← no of cases

H_0 :- sample mean ($\bar{x}$) = μ

H_1 :- $\bar{x}$ is different to μ

Hypothesis for independent samples

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

std deviation; no of cases

H_0 :- means in 2 groups are equal. (no difference btwn groups)

H_1 :- means are not equal

Hypothesis for paired t-test

$$t = \frac{\bar{x_d} - 0}{S/\sqrt{n}}$$

H_0 :- mean of differences between the pairs is zero

H_1 :- mean of " btwn pair is non zero.

- Procedure for finding t

given :- reference t value for certain degree of freedom & significance.

calculate t acc to sample.

if calculated t > reference ⇒ reject null hypothesis

ex:- 10 individuals are chosen @ random and heights in inches are 63, 63, 64, 65, 66, 69, 69, 70, 70, 71. Discuss proposal that mean height is 65 inches given for 9 degrees of freedom, value of t at 5% level of significance is 2.262.

~~$\bar{x} \rightarrow 67$~~ ~~$\mu = 65$~~

~~std dev $= \sqrt{\frac{(x_i - \bar{x})^2}{n}}$~~

$\bar{x} = 67$

$\mu = 65$

std dev $= \sqrt{\frac{(x_i - \bar{x})^2}{n}}$ ~~std dev $\frac{\sqrt{(x_i - \bar{x})^2}}{n} = \frac{8.8}{10} = 0.88$~~

$t = \frac{x_i - \bar{x}}{\text{std}/\sqrt{n}} \Rightarrow \frac{67 - 65}{2.96/\sqrt{10}} = 2.136$

H_0 : → sample mean = population mean

~~calcu t < reference t~~

~~null hypothesis rejected~~

calcu t is not > reference t

∴ null hypothesis accepted.

calcu t < reference t
calcu t = reference t } H_0 accepted
calcu t > reference t } H_0 rejected

• 2 horses A & B tested

A	28	30	32	33	33	29	34
B	29	30	30	24	27	29	

Can u descriminate btwn horses? if 5% value of t for 11 df {degree of freedom} is 2.20

null hypo :- 2 horses same. No discrimination

$\bar{x}_1 \Rightarrow 31.28$ $\quad S_1 \Rightarrow 2.1189$

$\bar{x}_2 \Rightarrow 28.1667$ $\quad S_2 \rightarrow 2.1147$

$$t \Rightarrow \frac{31.28 - 28.17}{\sqrt{\frac{2.1189^2}{7} + \frac{2.1147^2}{6}}} = 2.64$$

calcu t > refer t

∴ null hypo rejected

Performance of 2 horses are different

g_1 :- 18 20 36 50 49 36 34 49 41

g_2 :- 29 28 26 35 30 44 46

examine differences in marks

t for 14 df @ 5% level of significance is 2.14

$\bar{g}_1 \Rightarrow 37$

$\bar{g}_2 \Rightarrow 34$

$S_1 \Rightarrow 13.53$

$S_2 \Rightarrow 7.42$

$t = 0.5$

Accept ho

Ten soldiers tour a rifle range once in a week for two successive weeks. Their scores in the first week were 67, 24, 57, 55, 63, 54, 56, 68, 33, 43. Their scores in the second week (in the same order) were 70, 38, 58, 58, 56, 67, 68, 75, 42, 38. Is there any significant improvement. Given $t_{9\,df\,@\,0.05} = 2.262$.

$W_1 \rightarrow$ 67; 24; 57; 55; 63; 54; 56; 68; 33; 43

$W_2 \rightarrow$ 70; 38; 58; 56; 67; 68; 75; 42; 38.

for uniformity

$W_2 - W_1 \rightarrow$ 3; 14; 1; 1; 4; 14; 19; -26;

W_2	W_1	$W_2 - W_1$
70	67	3
38	24	14
58	57	1
58	55	3
56	63	-7
67	54	13
68	56	12
75	68	7
42	33	9
38	43	-5

$\bar{x} \Rightarrow 5$

$s \Rightarrow 6.94$

$$t = \frac{5}{6.94/\sqrt{10}} = 2.278$$

Null hypo rejected. There is significant change in performance

* Association rule mining

- defines dependency btwn 2 sets of objects
- written as 'antecedent → consequent'
- main idea behind data mining

examples:-

transaction id	items
t1	milk, bread, coffee, tea
t2	milk, bread
t3	milk, coffee
t4	bread, ketchup
t5	milk, tea, sugar

itemset → collection of items in a single transaction

support count → frequency of an item in a transaction

eg:- support count {milk} ⟹ 4

{milk, bread} ⟹ 2.

{milk, egg} ⟹ 0

min (support count) ⟹ frequency of itemset at which it becomes relevant enough to be included in association mining process

frequent item ⟹

support ⟹ probability that the itemset is present in a transaction.

↳ support count / no. of transactions

confidence ⟹ if we have association rule $A \rightarrow C$ where A & C are 2 itemsets;

$$\text{confidence}_{A \rightarrow C} \Rightarrow \frac{\text{support of } (A \cup C)}{\text{support } (A)}$$

(if A bought, how confident are we that user would buy C.)

confidence (milk, bread) ⟶ {coffee}

$$\Rightarrow \frac{\text{support}\{\text{Milk, bread, coffee}\}}{\text{support}\{\text{milk, bread}\}}$$

$$\Rightarrow \frac{0.2}{0.4} = 0.5$$

1/5 , 2/5

Lift

(a strength of association rule

$$\text{Lift}(A \rightarrow C) = \text{support}\left[\frac{(A \cup C)}{\text{support}(A) * \text{support}(C)}\right]$$

$$= \text{support}\left[\frac{(A \cup C)}{\text{support}(A) \times \text{support}(C)}\right]$$

if lift < 1 ; antecedent & consequents are substitutes for each other

lift > 1 ; antecedent & consequents are dependent on each other

lift = 1 ; antecedent & consequent are independent

Types of data mining

- apriori algorithm
- FP growth algorithm
- CARMA { classification Association Rules based on Multiple associations }
- Eclat Algorithm : { equivalence class clustering and Bottom up lattice traversal }

* **Apriori association rule.**
  - Used in market basket analysis
  - generate frequent itemsets in a transaction dataset
  - iterative process

Procedure
- generate frequent itemsets
- use frequent itemsets to generate association rules
- Finalize rules.

Exercise:- Perform a apriori association algorithm to find the support & confidence for the most frequently bought items given in the transaction.

Transaction id	items
1	$I_1; I_2; I_4$
2	$I_2; I_3; I_5; I_6$
3	$I_1; I_2; I_3; I_5$
4	$I_2; I_5$
5	$I_1; I_3; I_5$

① Find transaction table:

	I_1	I_2	I_3	I_4	I_5	I_6
T_1	1	1	0	1	0	0
T_2	0	1	1	0	1	1
T_3	1	1	1	0	1	0
T_4	0	1	0	0	1	0
T_5	1	0	1	0	1	0

② Frequent itemset with one item

item	support count (bought how many times)
I_1	3
I_2	4
I_3	3
I_4	1
I_5	4
I_6	1

③ Ignore items with less than 2 support count

i_1 — 3

i_2 — 4

i_3 — 3

i_5 — 4

④ frequent itemsets for 2 items

$\{i_1, i_2\}$

$\{i_1, i_3\}$

$\{i_1, i_5\}$

$\{i_2, i_3\}$

$\{i_2, i_5\}$

$\{i_3, i_5\}$

⑤ from question; find support count

$i_1, i_2 \longrightarrow 2$

$i_1, i_3 \longrightarrow 2$

$i_1, i_5 \longrightarrow 2$

$i_2, i_3 \longrightarrow 2$

$i_2, i_5 \longrightarrow 3$

$i_3, i_5 \longrightarrow 3$

Since support count < 2 is not available; we consider all the 2 itemsets

⑥ 3 itemsets

$\{i_1, i_2, i_3\} \longrightarrow 1$

$\{i_1, i_2, i_5\} \longrightarrow 1$

$\{i_1, i_3, i_5\} \longrightarrow 2$

$\{i_2, i_3, i_5\} \longrightarrow 2.$

from the 4; 3 itemset we will calculate subitem within list. This is called pruning

itemset	subset	all subsets freq itemsets?
$\{i_1 i_2 i_3\}$	$i_1 i_2, i_1 i_3, i_2 i_3$	ye
$\{i_1 i_2 i_5\}$	$i_1 i_2, i_1 i_5, i_2 i_5$	ye
$\{i_1 i_3 i_5\}$	$i_1 i_3, i_1 i_5, i_3 i_5$	ye
$\{i_2 i_3 i_5\}$	$i_2 i_3, i_2 i_5, i_3 i_5$	ye

* feature engineering

- feature selection is the process of selecting sub features from a dataset based on certain criteria - reduces no of features

- filter methods
- wrapper methods
- embedded methods

i) filter method - evaluate each feature independently with target variable
   - fast, remove redundancy
   - chi-square method
   - info gain
   - correlation coefficient

ii) wrapper method → greedy alg that train algorithm
   - precisely selected

iii) embedded methods → combine both

---

mid term