[1] Chi-square test is a statistical procedure for determining the difference between observed and expected data, in other words, it compares obtained results with those expected theoretically on some hypothesis.

thus, it is a measure of the actual divergence of observed and expected frequencies.

Example :-

Handedness and nationality	Right-handed	left handed
American	236	19
Canadian	157	16

A Chi-square test (test of independence) can test wether these observed frequencies are significantly different from the frequencies expected if handedness is unrelated to nationality.

November
Thursday
23

نوفمبر ٢٣
الخميس

٩ جمـاد أول ١٤٤٥ هـ
١٢ هاتـور ١٧٤٠ ق

Example 2: Bird species at a bird feeder

Frequency of visits by bird species at a bird feeder during 24-hour period

Bird species	Frequency
House sparrow	15
house Finch	12
Black-Capped chickadee	9
Common grackle	8
European Starling	8
Mourning dove	6

A Chi-square (Goodness of fit) test can test whether these observed frequencies are significantly different from what was expected, such as equal frequences.

٨ جمـــاد أول ١٤٤٥ هـ
١٢ هـــاتـــور ١٧٤٠ ق

نوفمبر
الأربعاء

٢٢

November
Wednesday
22

Applications of Chi square test

① Chi-square test when expectations are based on normal distribution

② testing the divergence of observed results from expected results when our expectations are based on the hypothesis of equal probability.

③ Chi-square test when our expectations are based on predetermined results.

④ chi square test of independence in contingency tables.

Uses of Chi-square test in Data Science

① Chi-square test is used for feature selection, one common problem related to feature selection determines how relevant the input features are to the predictive output, this test can be used to understand how the output variable depends on the input variable

Chi-square test

test of Goodness of fit

test of independence

في توصيات ال

Week 47

SAT	SUN	MON	TUE	WED	THU	FRI	SAT	SUN	MON	TUE	WED	THU	FRI
				01	02	03	04	05	06	07	08	09	10
11	12	13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30								

326-40

November
2023
Wednesday

(2) Degree of freedom is the maximum number of logically independent values, which may vary in data sample, Degrees of freedom are calculated by subtracting one from the number of items within the data sample.

if we have a data sample consisting of 5 positive integers, the values of the five integers must have an average of six. if 4 items within the data set are {3, 8, 5 and 4}, the fifth number must be 10. Because the first four numbers can be chosen at random, the degree of freedom is four.

(3) T- Distribution is heavily used in the world of statistics, in particular, where the sample size is small (~30 observations) and/or the Population standard deviation is unknown, it is an approximation of normal distribution

F- Distribution is a continous statistical distribution which arises in the testing of whether two observed samples have the same variance.

the F-distribution can be regarded as the equivalent extension of the t distribution when there is more than one variable but small sample sizes.

SAT	SUN	MON	TUE	WED	THU	FRI	SAT	SUN	MON	TUE	WED	THU	FRI
				01	02	03	04	05	06	07	08	09	10
11	12	13	14	15									

November
Wednesday
15

نوفمبر
الأربعاء
١٥

١ جمـاد أول ١٤٤٥ هـ
٥ هـاتـور ١٧٤٠ ق

In Data Science, we can use ~~t~~ F-distribution for

(1) finding the critical values for a confident interval when the data is approximately normally distributed.

(2) finding the corresponding p-value from a statistical test that uses the t-distribution (t-test, regression analysis)

F-distribution is used for

(1) hypothesis testing, to compare data from two or more populations

(2) ~~An~~ Analysis of variance (ANOVA)

[4] Probability Distribution is the mathematical function that gives the probabilities of occurance of different possible outcomes for an experiment.
itis often depicted using graphs or probability tables

[5] there are two types of probability distribution

① Discrete probability distribution:- itis a probability distribution of a categorical or discrete variable.
Common types of Discrete probability distribution:->
Binomial → describes variables with two possible outcomes.
Discrete uniform → Describes events that have equal probabilities.
Poisson → Describes count data, it gives

the probability of an event happening k number of times within a given interval of time or space.

(2) Continuous probability distributions is the probability distribution of a common variable.

Common types :-

Normal distribution :- Describes data with variables that becomes less probable the farther they are from the mean with a bell-shaped probability density function.

Continous uniform :- Describes data for which equal-sized intervals have equal probability.

Log-normal, and Exponential

problem 2:-

From 10 am to 11:30 am, length of interval $= 1.5$

$\Rightarrow X \sim \text{Poisson}(\lambda = 1.5 * 10 = 15)$

$\therefore$ ~~$P(10 < X < 15) \Rightarrow$~~ where X is a Poisson random variable

$\therefore P(10 < X \leq 15) = \sum_{K=11}^{15} P_X(K)$

$$= \sum_{K=11}^{15} \frac{e^{-15} \cdot 15^K}{K!}$$

$$= e^{-15}\left[\frac{15^{11}}{11!} + \frac{15^{12}}{12!} + \frac{15^{13}}{13!} + \frac{15^{14}}{14!} + \frac{15^{15}}{15!}\right]$$

$$= 0.4498$$

MCQ

4) 15, 11, 9, 5, 18, 4, 15, 13, 7

mean = $\frac{15+11+9+5+18+4+15+13+7}{9}$ =

Sorting: ~~5, 4, 5, 9, 11~~ 4, 5, 9, 11, [13], 15, 15, 17, 18

~~4, 5, 7, 9, 11, 13, 15, 15, 18~~

median = 13

mode = 15

اذا كثر تكرارا

answer ⇒ 13, 15

5) True ⇒ X is the value repeated the maximum number of times ⇒ mode

6) True

من الـ properties الـ expectation بتاعت

7) probability of tail = $\frac{1}{2}$ for a single toss

P = 0.5

سنة أولى

$q = 0.5$

$n = 4$, X is binomial random variable

$$P(X=3) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \binom{4}{3} * 0.5^3 * (0.5)^1$$

$$= \frac{4!}{3!(4-3)!} * 0.5^3 * (0.5)$$

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

(8) $E[X^2] - [E(X)]^2$

(9) $\int_0^3 x^2 \, dx = \left[\frac{x^3}{3}\right]_0^3$

$= \frac{27}{3} - 0 = 9$

(10) $Var(Z) = Var(5X - 2Y)$

$= Var(5X) + Var(2Y)$

$= 25\,Var(X) + 4\,Var(X)$

Var(Z) = 25 * 0.2 + 4 * 0.5 = 7

(6) probability distribution

(7) Continuous random variable
continuous يكون فيها float أو

(8) Discrete random variable
int

(9) $\sum X P(X)$

(10) $\int X P(X)$

(11) $P(X) = -0.5$
Because $P(X) < 1$

(12) $E(Z-X) = E(Z) - E(X)$
$= 4 - 2 = 2$

(13) positive probability, because
$0 < P(X) < 1$

(3) Random variable

(4) Discrete random variable

(5) Continuous " "

(6) $\therefore \sum P(x) = 1$ مجموع الاحتمالات = ١

$\therefore \quad k^2 - 8 = 1 \quad , \quad k^2 = 9$

$\therefore \quad k = \sqrt{9} = 3$

(7) $E(x) = \sum x P(x) = 4 * 0.5 = 2$

(8) Discrete distribution

(7) Discrete Distribution

(8) $P(x) = \dfrac{e^{-m} m^x}{x!} \Rightarrow P(0) = \dfrac{e^{-m}}{1} = e^{-m}$