Chi-Squared :-

"Check all e.gs from slides"

What if we have categorical data and we want to make predictions about that?

→ / tests

e.g (i) Whether gender is related to voting preference?

(ii) Whether smoking is related to lungs cancer?

To answer this we use chi-squared distribution.

Imp ★

if $Z_1 \cdots Z_k$ are independant std normal random variables then. $Z_1^2 + \cdots + Z_k^2$, they have χ^2 dist with k degrees of freedom.

Date

first

the ↑ picture in the slides shows the probability density function (PDF) of chi-squared distribution of various degrees of freedom.

(X²)

Chi squared dist is a continuous probability distribution which is often used in confidence intervals and hypothesis testing for variance.

(i) It is only defined for $x \geq 0$, since it represents squared values (of Z perphaps).

(ii) The shape of the dist depends upon degree of freedom

x-axis → random variable $f_k(x)$

y-axis → Probability density of how likely values of x are given for k.

↳ e.g = k=1.

When k = small the distribution is heavily skewed to the right, with most of the probability density concentrated near zero.

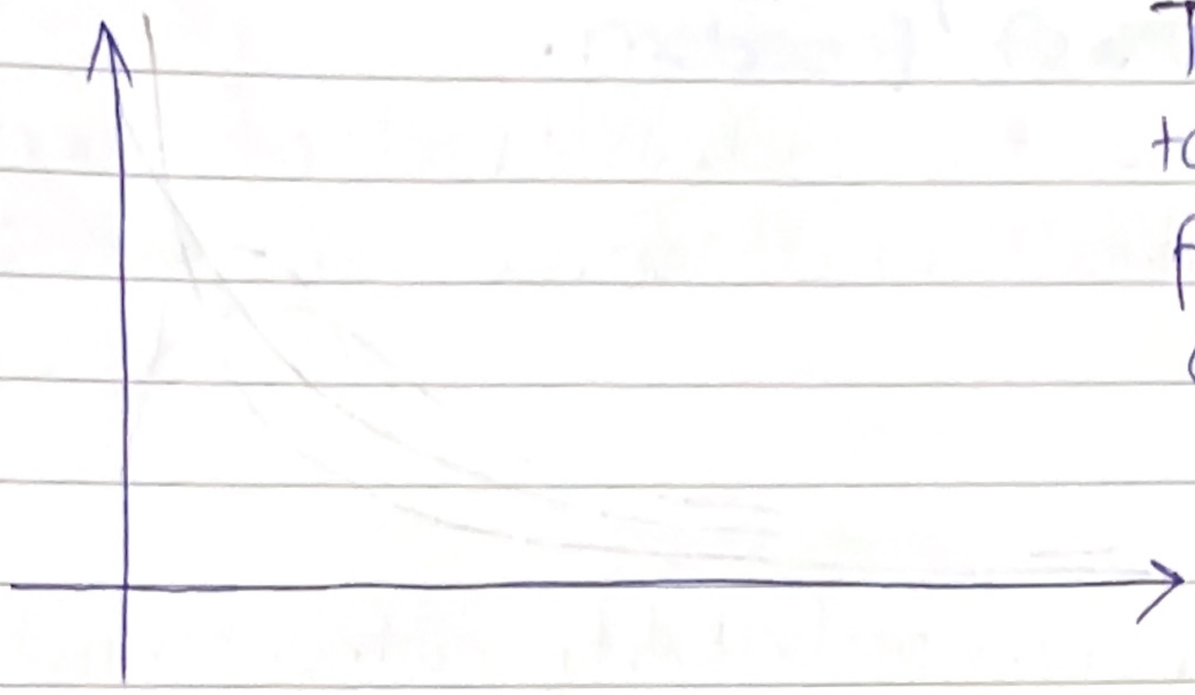
k = increases, the distribution becomes more symmetric and spreads out, resembling a normal distribution.

↳ e.g k=9

* The mean of chi squared dist ⇒ (k) → degree of freedom.

The std-deviation of " ⇒ 2k

how value of k - affects the chi-squared distribution.



The Chisquared dist is often skewed to the right for small degree of freedom because, the sum of a very few normal random variables often results in small values near to zero but occassional larger values

As the k increases, the graph becomes; that is why a long tail. more spread out and starts resembling a symmetric, bell shaped curve.

(i) This happens because adding more squared normal variables averages out the variability and reduces the skewness.

Chi-squared distribution uses:-

(i) goodness of fit;
to observe a frequency distribution fits a specific expected distribution, also known as the goodness of fit.

(ii) to test whether the categorical variables are independant or if there is some association b/w them.
(testing independence in contigency tables)

(iii) estimating variance.

Date

Example :- test at a significance level of 0.05 or 5%.

H_0 : The die was fair
thus the probability of each toss from the 90 tosses is equal to
↳ 1/6.
↳ $P(X=x)$ where $x = 1,2,3,4,5,6$.

$\Rightarrow H_A$: The die is not fair.

$\therefore P(X=x) \neq 1/6$.

The expected frequency thus $\Rightarrow (90)\,1/6. \Rightarrow 15$.
Observed frequencies are $\Rightarrow$ 20, 15, 12, 17, 9 and 17.

$$\frac{(20-15)^2}{15} + \left(\frac{(15-15)^2}{15}\right) + \frac{(12-15)^2}{15} + \ldots \frac{(17-15)^2}{15}.$$

$\Rightarrow 5.2$

number of categories.

$df = 6 - 1$
$\Rightarrow 6-1$
$\Rightarrow 5$
$\alpha = 0.05$

$\chi^2 \Rightarrow 11.1$

$11.1 > 5.2$.

therefore we accept the NULL hypothesis.

Example:-

observed frequencies values:	Boy	Girls	Total.
Grades	117	130	247
Popular	50	91	141
Sports	60	30	90
Total.	227	251	478.

Test at a significance level of 5%.

Date

$H_0 \Rightarrow$ The gender and categories are independant.

$H_A \Rightarrow$ " " " are dependant.

Solve for specifics separately.

for boys

$$\frac{(117-117.299)^2}{117.299} + \frac{(50-66.960)^2}{66.960} + \frac{(60-42.741)^2}{42.741}$$

$$\Rightarrow 11.3032.$$

for girls.

$$\frac{(130-129.701)^2}{129.701} + \frac{(91-74.040)^2}{74.040} + \frac{(30-47.259)^2}{47.259}$$

$$\chi^2_c \Rightarrow 11.1999.$$

total Chi-squared.

$$\chi^2 = 11.3032 + 11.1999 \Rightarrow 22.5031.$$

rows ↑

$df \Rightarrow (r-1)(c-1) \Rightarrow (3-1)(2-1) \Rightarrow 2.$

$\alpha = 0.05$

↓ columns.

$\chi^2_c = 5.99$ $\quad 5.99 < 22.199.$

as the Critical Chisquared value is greater that the Chi-squared statistic we reject the NULL hypothesis.

Date

Variance Estimation:-

Chi-squared dist is often used for testing hypothesis about the variance of a population but the estimating variance in it, refers to constructing an interval of confidence where the true population variance can lie.

S^2 → unbiased population variance estimater.

$$S^2 = \frac{1}{n-1} \sum_{i}^{n} (X_i - \overline{X})$$

↙ sample variance.

The sample variance s^2 can be transformed to represent the chi-squared distribution (to account for the smaller sample sizes).

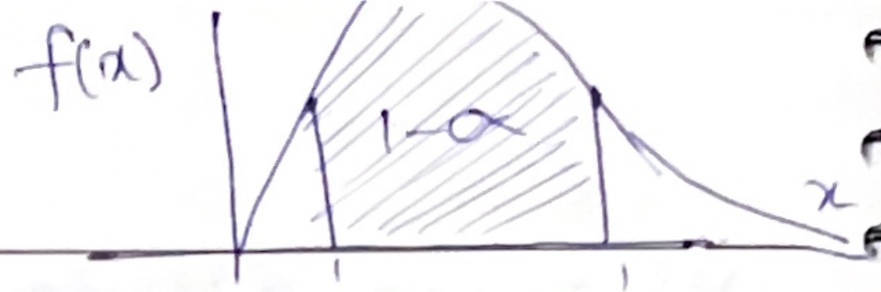
$$\frac{(n-1)s^2}{\sigma^2} \sim X^2_{n-1}$$

or.

$$\underline{(n-1)s^2} = \sum_{i=1}^{n} \underline{(x_i - X)^2}$$
$$\sigma^2 \qquad\qquad \sigma^2.$$

Now to define the confidence interval for true population mean

$$P\left\{ X^2_{1-\alpha/2} \frac{(n-1)s^2}{\sigma} \leq X^2_{\alpha/2} \right\} = 1-\alpha.$$



taking reciprocal.

$$\frac{1}{\chi^2_{1-\alpha/2}} \leq \frac{\sigma^2}{(n-1)s^2} \leq \frac{1}{\chi^2_{\alpha/2}}$$

Multiplying $(n-1)s^2$.

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$

$$\therefore \left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right] \checkmark$$

——— X ——— X ———

Date

Chi-Squared Test.

ii) Goodness of fit :-

In such type of test we want to analyze how well our data is being mapped to the Expected data.
↳ observed

Consider the example by Joshua Emmanuel.

Expected probabilities of different outcomes :-

$p_1 = 0.28$ (being white)

$p_2 = 0.25$ (" black)

$p_3 = 0.16$ (" Silver)

$p_4 = 0.31$ (" others)

H_0 = the given expected values are true.

H_A = atleast one pp is not the same as specified in the H_0.

No. of categories = 4 ✓

No. of degree of freedom = 4 - 1 ✓
$\Rightarrow 3$.

Chi test = $\sum \frac{(f_{obse} - f_{expt})^2}{f_{expt}}$

$\alpha = 0.05$

Using the chisquared table

Chi Critical Value = 7.815

Now rejection region $\Rightarrow \chi^2 > 7.815$.

for that just ÷ categories there could be a senario where you with 4 ⟶ (more below) ← may be given observed frequency and asked if they are equally distributed.

Date

Now using This Chi-squared test formula.

$$\Rightarrow \sum_{i}^{k} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}}$$

simple space.
↗ (140.)
total no of objects

$E_i = np_i$ ↗ expected frequency or probability

O	E	O−E	(O−E)²	(O−E)²/E
39	140(28)	-0.2	0.04	0.001
29	140(0.25)	-60	36.00	1.028
24	140(0.16)	1.6	2.56	0.114
28	140(0.31)	4.6	21.16	0.4876
140	140			$\chi^2 = 1.6315$

hence we accept the NULL hypothesis.

After dividing you will know how much percentage each category will have, form a expected table from that and solve. as the same.

EXAMPLE FROM SLIDES ↓ :- $n = 6+84 = 790$ tosses

	Observed	Expected.	O-E	$(O-E)^2$	$(O-E)^2/E$
DICE1	20	90(1/6)			
DICE2	15	99(1/6)			
DICE3	12	〃			
DICE4	17	〃			
DICE5	9	〃			
DICE6	17	〃			

Remaining steps are similar.

Date

(iii) Chi-Squared test of Independance

USES
- ↳ (i) Cross Tabulation
- ↳ (ii) Contingency
- ↳ (iii) 2-way table
- ↳ (iv) 2-categorical table.

can be ranked ↑ Nominal/ordinal

We aim to describe if the two categories are independant or not using the Chi squared test.

Consider the example by Joshua Emmanuel

$H_0 \Rightarrow$ No relationship b/w the categories.

$H_A \Rightarrow$ There is a relationship.

Now lets make table to draw χ^2 for each category.

YES

	O	E = (row total × colum total)/sample size → 490×809/1440	(O−E)²	(O−E)²/E
under 30	312	275.28		
30 → 50	277	256.18		
Over 50.	224	277.53		

Similarly for No and SUM

$df = (r-1)(c-1) \Rightarrow 2$

$\alpha = 0.05$

$\chi^2_{Yes} + \chi^2_{NO} = 37.26$ ← compare.