P-Value

P-value is the prob of getting a sample as or more extreme (having more evidence against H0) than our own sample given the Null hypo (H0) is true.

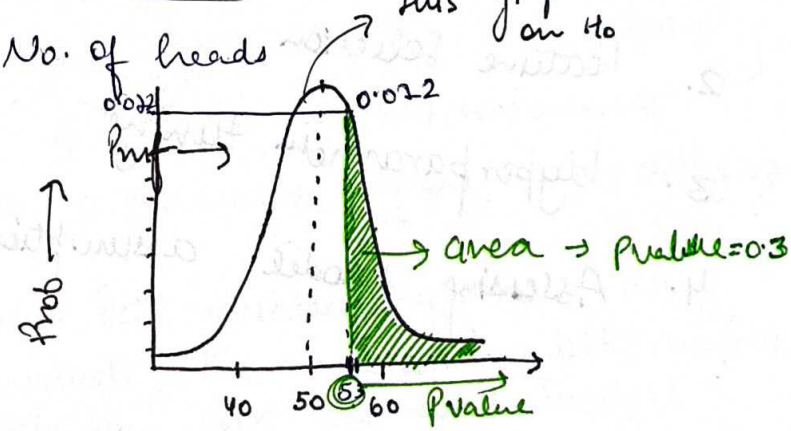
Experiment

↓

1 coin → 100 times toss

binomial distribution (ex head, tail)

No. of heads



$H_0 : P(H) = P(T)$

$H_a : P(H) > P(T)$

again did exp → 100 time toss

53 heads

If exp → 100 coins → 100 time toss

P = 0.3 then

30 times → 53 head came.

Interpreting p-value

with significance value

$\alpha = 0.05 / 0.01$ → p-value $\leq \alpha$

Reject Null Hypo

Without Significance Value

1. Very small p-value (eg $p < 0.01$) indicate strong evidence against the null hypo, suggesting that the observed effect or different is unlikely to have by chance alone.

2. Small p-value ($0.01 < p < 0.05$) indicate moderate evidence against the null hypo, suggesting that the observed effect or diff is less likely to have occured by chance alone.

3. Large p-value (eg.. $0.05 \leq p \leq 0.1$) indicate weak evidence against the null hypo, suggesting that the observed effect or difference might have occured by chance alone, but there is still some level of uncertainity.

4. Very larger p-value (eg $p \geq 0.1$) indicate weak or no evidence against the null hypothesis, suggesting that the observed effect or diff is likely to have occured by chance alone.

P-value in context of Z-test

Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the avg productivity of its employees before implenting the training program. The avg productivity was 50 units per day. After implementing the training program the company measures the productivity of a random sample of 30 employees. The sample has an avg productivity of 53 units per day and the pop std is 4. The company wants to know the new training program has significantly increased productivity.

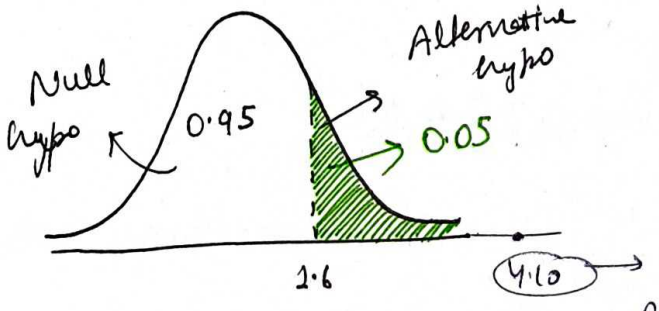
$\mu = 50$ $\quad n = 30$ $\quad \bar{x} = 53$

$\sigma = 4$ $\quad \alpha = 0.05$

$H_0 : \mu = 50$

$H_a : \mu > 50$

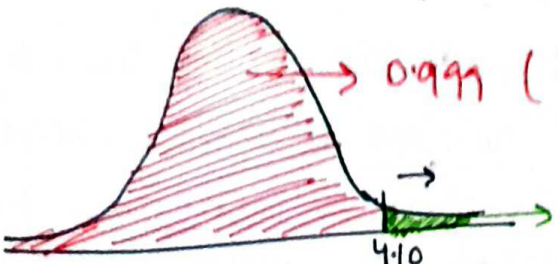
$$Z\text{-Stat} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{53 - 50}{4/\sqrt{30}} = \frac{3}{4} \times \sqrt{30} = 4.10$$



→ reject Null

Problem → not have strength of evidence.

P-value → critical point



P-value < 0.05

0.0001 < 0.05 → reject Null hypo

Q Suppose a snack food company claims that their lays wafer packets contains an avg weight of 50gm per packet To verify this claims a consumer watchdog organisation decides to test a random sample of lays wafer packets. The organisation wants to determine whether the actual avg weight differs significantly from the claimed 50 gms. The organisation collects a random sample of 40 lays wafer packet and measure their weights. They find that sample has an average weigh of 49 grams with a standard deviation of 5 grams.

$\mu = 50$ $n = 40$ $\bar{x} = 49$

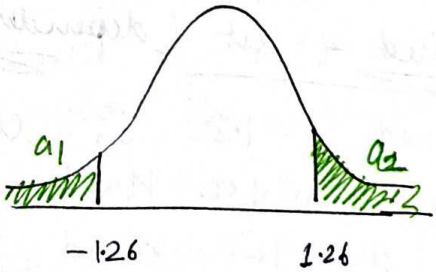
$\sigma = 5$ $\alpha = 0.05$

$H_0 : \mu = 50$

$H_a : \mu \neq 50$

$$z = \frac{49 - 50}{5/\sqrt{40}} = \frac{-\sqrt{40}}{5} = -1.26$$

both side bcz we don't know direction ($\mu > 50$ or $\mu < 50$)



P-value = $a_1 + a_2$

$a_1 = 0.103$

P-value = 0.103 + 0.103 = 0.206

P-value (0.206) > 0.05 → Not reject Null hypo

T-tests

A t-test is a statistical test used in hypo testing to compare the means of two samples or to compare a sample mean to a known pop mean. The t-test is based on the t-distⁿ, which is used when the pop standard deviation is unknown and the sample size is small.

There are three main types of t-tests:

One sample t-test: The one-sample t-test is used to compare the mean of a single sample to a known pop mean. The null hypothesis states that there is no significant diff betⁿ the sample mean and the pop mean, while the alternative hypo states that there is a significant difference. $\bar{x} \rightarrow \mu$ pop std σ

Independent two-sample t-test: The independent two sample t-test is used to compare the means of two independent samples. The null hypo states that there is no significant diff betⁿ the means of the two samples, while the alternate hypo states that there is a significant diff.

Paired t-test (dependent two sample t-test): The paired t-test is used to compare the means of two samples that are dependent or paired such as pre-test and post-test scores for the same group of subjects or measurements taken on

the same subjects under two diff condition. The null hypo states that there is no significant difference betn the means of the paired diff, while the alternative hypo states that there is a significant diff.

Single Sample t-test

A one-sample t-test checks whether a sample mean differ from the pop mean.

Assumption for a single sample t-test

1. Normality: Population from which the sample is drawn is normally distributed
2. Independence: The observation in the sample must be independent
3. Random Sampling
4. Unknown pop std.

Q Suppose a manufacture claims that the average weight of their new chocolate bars is 50 grams, we highly doubt that and want to check this so we draw out a sample of 25 chocolate bars and measured their weight, the sample mean came out to be 49.7 grams and the sample std dev was 1.2 grams. Consider the significance level to be 0.05.

$\mu = 50$ $n = 25$ $\bar{x} = 49.7$ $\alpha = 0.05$

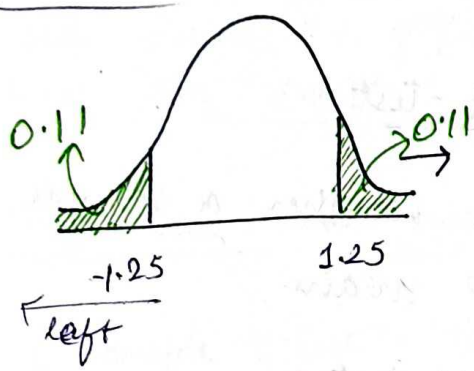
$S = 1.2$

$H_0 : \mu = 50$

$H_{ac} : \mu \neq 50$

assumig it is normal

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{49.7 - 50}{1.2/\sqrt{25}} = -1.25$$



Student 't distbⁿ

for t-value require → (n-1) degree of freedom

→ Cumulative dist^n or one tail or two tail

we don't have this info so we not use t-table

```
from scipy.stats import t
cdf_value = t.cdf(t_value, df)
```

→ area from left

P-value = 0.11 + 0.11 = 0.22

P-value = 0.22

P-value > 0.05 → Not reject Null Hypo

Shapiro-Wilk → use for check whether sample data is Normal distbⁿ or not

→ P-value < 0.05 not normal

Independent 2 sample t-test

An Independent two-sample t-test, also known as an unpaired t-test, is a statistical method used to compare the means of two independent groups to determine if there is a significant diff betn them.

Titanic

male avg > female avg

Assumption for the test:

1. Independence of observations:
2. Normality
3. Equal variance (Homoscedasticity): $(\sigma_A^2 = \sigma_B^2)$

test equal var → F-test or levene

* If this assumption is not true then use Welch's t-test which does not require equal paramt.

Suppose a website owner claims that there is no diff in the avg time spent on their website betn desktop and mobile users. To test this claim, we collect data from 30 mobiles users regarding the time spent on the website in minutes. The sample statistics are as follows:

desktop users = [12, 15, 18, 16, 20, - - - -]

mobile users = [10, 12, 14, 13, 16, - - - - - - - -]

Desktop users

- Sample size (n1) : 30
- Sample mean (mean1) : 18.5 minutes
- Sample standard dev (Std_dev) = 3.5 minutes

Mobile users:

- Sample size (n2) : 30
- Sample mean (mean2) : 14.3 minutes
- Sample std (std-dev2) = 2.7 minutes

We will use a significance (α) of 0.05 for the hypo test.

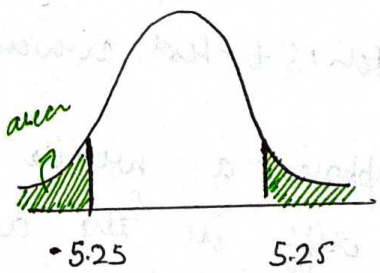
$H_0 : \mu_{desktop} = \mu_{mobile}$

$H_a : \mu_d \neq \mu_m$

1. Shapiro test → To check Normality
2. levene test → To check equal variance (p-value > 0.05 then equal variance)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightarrow \text{this formula for 2 sample test}$$

$$t = \frac{18.5 - 14.3}{\sqrt{\frac{(3.5)^2}{30} + \frac{(2.1)^2}{30}}} = 5.25$$



degree of freedom = $n_1 + n_2 - 2$

```
from scipy.stats import t
cdf-value = t.cdf(t_value, df)
```

p-value < 0.05
(reject Null Hypo)

Paired 2 Sample t-test

A paired two-sample t-test, also known as a dependent or paired-samples t-test is a statistical test used to compare the means of two related or dependent groups.

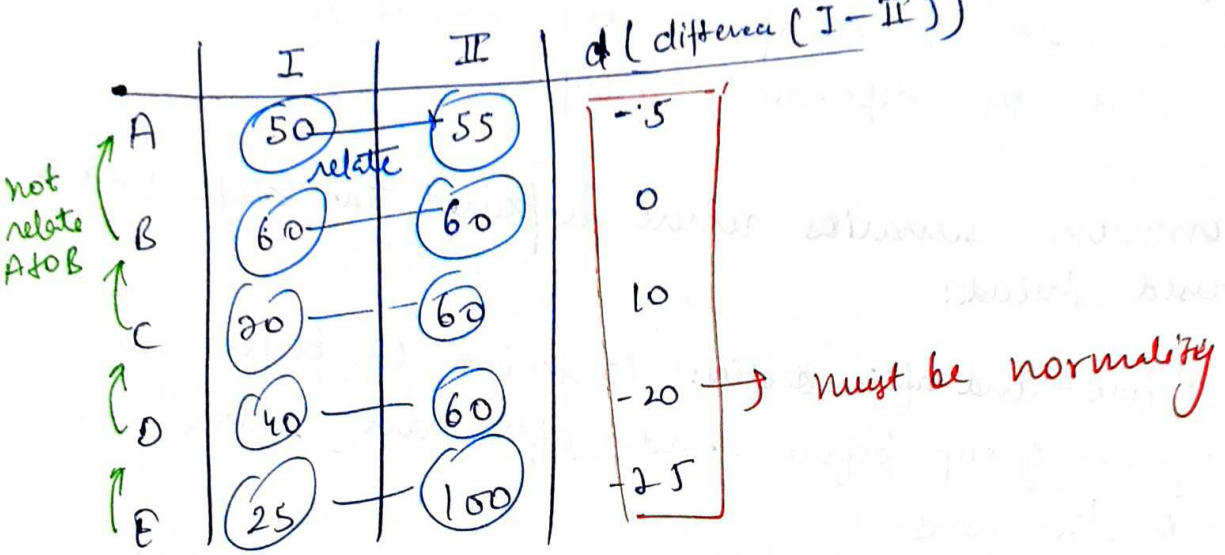
Common scenarios where a paired two sample t-test is used include:

1. Before-and after studies: Comparing the performance of a group before and after an intervention or treatment.
2. Matched or correlated groups: Comparing the performance of two groups that are matched or correlated in some way, such as siblings or pairs of individuals with similar characteristics.

Assumption

1. Paired observations: The two sets of observations must be related or ~~paired~~ paired in some way, such as before-and-after measurements on the same subjects or observations from matched or correlation groups.
2. Normality: histogram, Q-Q plots or statistical test (Shapiro-wilk test)
3. Independence of pairs: Each pair of observation should be independents of other pairs. In other

words, the outcome of one pair should not affect the outcome of another pair. This assumption is generally satisfied by appropriate study design and random sampling.



Let's assume that a fitness center is evaluating the effectiveness of a new 8-week weight loss program. They enroll 15 participants in the program and measure their weights before and after the program. The goal is to test whether the new weight loss program leads to a significant reduction in the participants weights.

Before the program:

[80, 92, 75, - - - - - - - - - - - - - -]

After the program

[78, 93, - - - - - - - - - - - - - -]

Significance level (α) = 0.05

H_0: $\mu_{\text{before avg weight}} = \mu_{\text{after avg weight}}$

H_1: $\mu_{\text{before}} > \mu_{\text{after}}$

Name	wt before	wt after	diff
A	-	-	-
B	-	-	-
C	-	-	-
:	:	:	:
k			

① Normal check
② $\bar{x}_{diff}$ → Find
③ S_{diff} → Find

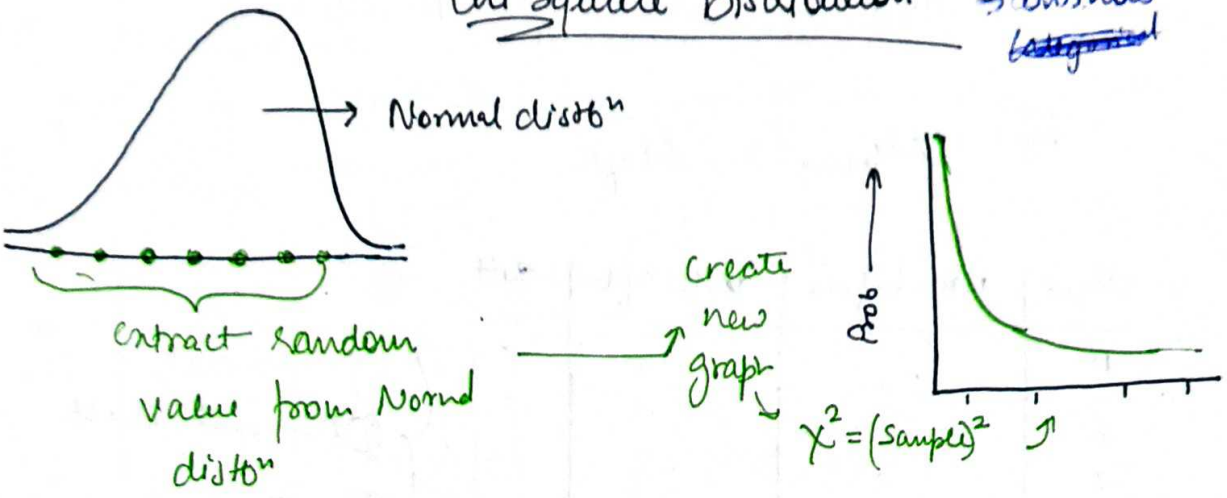
$$t = \frac{\bar{x}_{diff} - \mu_{diff}}{S_{diff} / \sqrt{n}}$$

assumption is there's no diff so we take as a 0.

Check code

$$t = \frac{\bar{x}_{diff} - 0}{S_{diff} / \sqrt{n}}$$

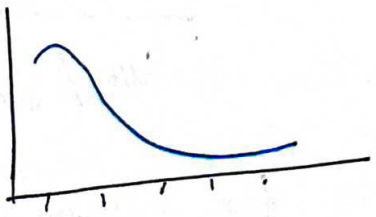
Chi Square Distribution → ~~Categorical~~ ordinal



Now take multiple sample value from Normal distⁿ

Sample value (1) Sample value (2)

$\chi^2 = (\text{Samp1})^2 + (\text{Sample2})^2$



~~Sample~~ degree of freedom (extract data from pop) = df

df ↑↑ = graph → Normal distⁿ

(no. of Sample var)

$\chi^2 = Z_1^2 \Rightarrow df = 1$

$\chi^2 = Z_1^2 + Z_2^2 \Rightarrow df = 2$

$\chi^2 = Z_1^2 + Z_2^2 + Z_3^2 \Rightarrow df = 3$

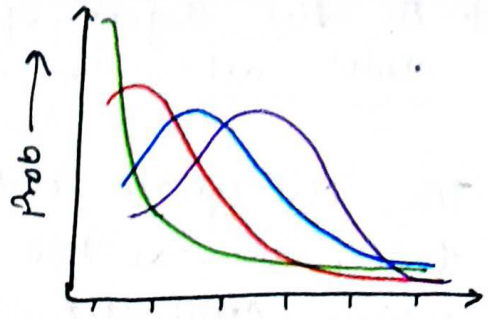
$\chi^2 = \sum_{i=1}^{k} Z_k^2$ $\boxed{df = k}$

$\chi_1^2 = z_1^2 \Rightarrow df = 1$

$\chi_2^2 = z_1^2 + z_2^2 \Rightarrow df = 2$

$\chi_3^2 = z_1^2 + z_2^2 + z_3^2 \Rightarrow df = 3$

$\chi_4^2 = z_1^2 + z_2^2 + z_3^2 + z_4^2 \Rightarrow df = 4$



The chi-square distbⁿ, also written as χ^2 distbⁿ is a continous prob distbⁿ that is widely used in statistical hypothesis testing, particularly in the content of goodness-of-fit-tests and test for independence in contigency tables. It arises when the sum of the squares of independent standard normal random variable follows this distribution.

The chi-square distbⁿ has a single parameter, the degree of freedom (df), which influences the shape and spread of the distbⁿ. The degree of freedom are typically associated with the numn. of independt var or constraints in a statistical problem.

Some key properties of the chi-square distbⁿ are:

a. It is continous distbⁿ, defined for non-negative values.
b. It is positively skewed, with the degree of skewness decreasing as the degrees of freedom increase.
c. The mean of the chi-square distbⁿ is equal to its degrees of freedom and its variance is equal to twice the degree of freedom.

d. As the degree of freedom increase, the chi-square distbⁿ approaches the normal distribution in shape.

The chi-square distbⁿ is used in various statistical tests, such as the chi-square goodness-of-fit test, which evaluates whether an observed frequency distbⁿ fits an expected theoritical distbⁿ, and the chi-square test for independence, which checks the association between categorical variables in a contigency table.

Chi Square Test

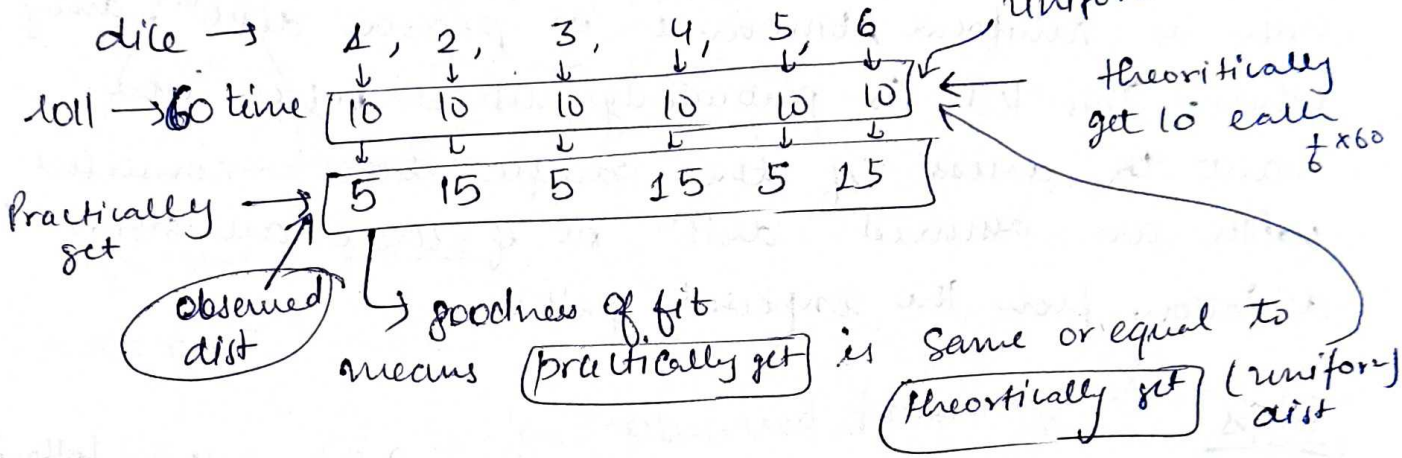
The chi-Square test is a statistical hypo test used to determine if there is a significant association between categorical variables or if an observed distbⁿ of categorical idata differs from an expected theoretical distbⁿ. It is based on the chi-square (χ^2) distbⁿ and it is commonly applied in two main scenarios:

~~t-Test~~

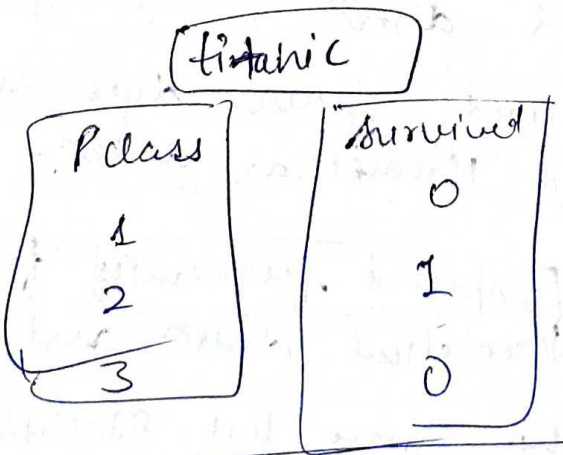
X^2 – Square

- Goodness of fit test
  - 1 categorical col
- test for independence
  - 2 Categorical diff col

1. **Chi-Square Goodness-of-Fit Test:** This test is used to determine if the observed distbⁿ of a single categorical variable matches an expected theoretical distribution. It is often applied to check if the data follows a specific prob distbⁿ, such as the uniform or binomial distbⁿ.



2. **Chi-Square Test for Independence:** (Chi-square Test for Association) This test is used to determine whether there is a significant association betⁿ two categorical variables in a sample.



* Check using chi square for Independence → Pclass and Survived are dependent or each other or not?

Goodness of fit Test

The Chi-square goodness-of-fit test is a statistical hypo test used to determine if the observed distⁿ of a single categorical variable matches an expected theoretical distⁿ. It helps to evaluate whether the data follows a specific prob distⁿ, such as uniform, binomial or poisson distⁿ, among others. This test is particularly useful when you want to assess if the sample data is consistent with an assumed distⁿ or if there are significant deviations from the expected pattern.

Steps

The Chi-Square Goodness-of-fit test involves the following steps:

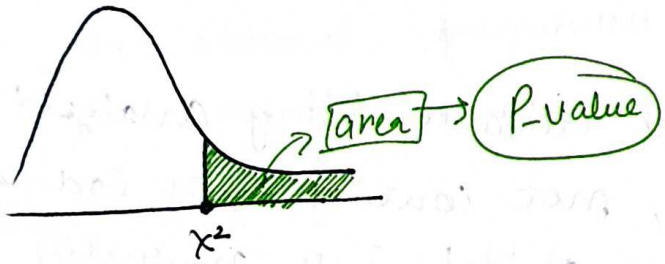
- Define the null hypothesis (H0) and the alternative hypo (H1):
  - → H0: The observed data follows the expected theoritical distⁿ.
  - → H_1: The observed data does not follow the expected theoretical distⁿ.
- Calculate the expected frequencies for each category based on the theoretical distⁿ and the sample size.
- Compute the chi-square test statistic (χ^2) by comparing the observed and expected frequencies. The test statistic is calculated as:

$$\chi^2 = \sum \left(\frac{\text{Observed}_i - \text{Expected}_i}{\text{Expected}_i} \right)$$

Dice →	1	2	3	4	5	6
Observed →	5	15	5	5	15	15
Theoritical →	10	10	10	10	10	10

(expected frequency like uniform distb)

$$\chi^2 = \frac{(5-10)^2}{10} + \frac{(15-10)^2}{10} + \cdots + \frac{(15-10)^2}{10}$$



$df = n - 1 = 6 - 1 = 5$

$\alpha = 0.05$

→ Where O_i is the observed frequency in category i, and E_i is the expected frequency in category i, and the summation is taken over all categories.

→ Determine the degree of freedom (df), which is typically the number of categories minus one ($df = k - 1$), where k is the number of categories.

→ Calculate the p-value for the test statistic using the chi-square distbn with the calculated degree of freedom.

→ Compare the test statistic to the critical value or the p-value.

Assumption

1. **Independence:** The observation in the sample must be independent of each other. This means that the outcome of one observation should not influence the outcome of another observation.

2. **Categorical data:** The variable being analyzed must be categorical, not continous or ordinal. The data should be divided into mutually exclusive and exhaustive categories.

3. **Expected frequency:** Each category should have an expected frequency of at least 5. This guidline helps ensure that the chi-square distbⁿ is a reasonable approximation for the distbⁿ of the test statistic. Having small expected frequencies can lead to an inaccurate estimation of the chi-square distribution potentially increasing the likelihood of a Type I error (incorrectly rejecting the null hypo) or a Type II error (incorrectly failig to reject the null hypo).

4. **Fixed distribution:** The theoritical distbⁿ being compared to the observed data should be specified before the test is conducted. It is essential to avoid choosig a distribution based on the observed data, as doing so can lead to biased results.

Example 1

Suppose we have a six-sided fair die, and we want to test if the die is indeed fair, we roll the die 60 times and record the num. of times each side comes up. We'll use the chi-square Goodness-of-fit test to determine if the observed frequencies and consistent with a fair die (i.e. uniform distⁿ of the sides).

Observed freq.

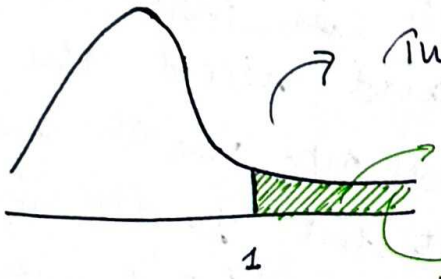
→ side 1: 12 times
→ side 2: 8 times
→ side 3: 11 times
→ side 4: 9 times
→ side 5: 10 times
→ side 6: 10 times

H_0: die is fair → uniform
H_1: die is not fair.

	1	2	3	4	5	
Observed =	12	8	11	10	10	
Expected =	10	10	10	10	10	→ uniform

$$\chi^2 = \frac{(12-10)^2 + (8-10)^2 + (11-10)^2 + (9-10)^2}{10} = \frac{10}{10} = 1$$

$\chi^2 = 1$ → Chi-square $df = n - 1 = 5$



This graph is made accordig to $df = 5$

p-value → (area) $\alpha = 0.05$

p-value will come greater than α

So, reject H_0.

Example 2

Suppose a marketing team at a retail company wants to understand the distbn of visits to their website by day of the week. They have a hypo that visits are uniformly distributed across all days of the week, meaning they expect an equal num. of visits on each day. They collected data on website visits for four weeks and want to test if the observed distbn matches the expected uniform distbn.

Observed frequencies (no. of website visits per day of the week for four weeks):

- Monday: 420
- Tuesday: 380
- Wednesday: 410
- Thursday: 400
- Friday: 410
- Saturday: 430
- Sunday: 390

	Mon	tue	wed	thru	fri	sat	sun
Obs →	420	380	410	400	410	430	390
	405	405	405	405	405	405	405

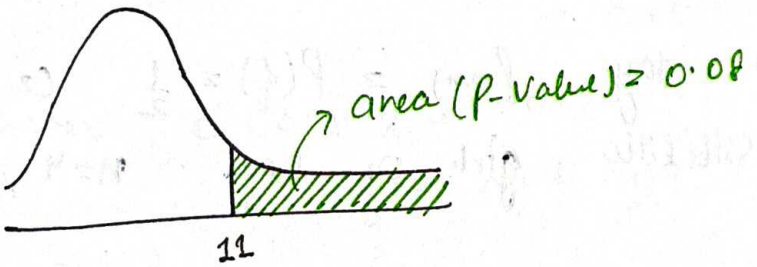
Expected = $\frac{420 + 380 + 410 + 400 + 410 + 430 + 390}{7} = 405$

H_0 : Uniform distribn

H_1 : Not uniform distrib^n

$$X^2 = \sum \frac{(Oi - Ei)}{Ei} = \frac{(420-405)^2 + (380-405)^2 + \cdots}{405}$$

$\chi^2 = 11$, $\quad df = n - 1 = 7 - 1 = 6$



Example 3

A survey of 800 families in a village with 4 children each revealed the following distrib^n:

# boys :	0	1	2	3	4
# families :	32	178	290	236	64

Is that data consistent with the result that male and female births are equally probable?

Survey → Village → 800 families

↓

↳ 4 children

Son

$P(S)_{\text{son}} = P(d)_{\text{daughter}} = \frac{1}{2}$

H_0: $P(m)_{\text{male}} = P(f)_{\text{female}}$

H_a: $P(m) \neq P(f)$

* In this example we use binomial distn not uniform distn. In binomial we use p and q where $q = 1-p$.

We can say $P(m) = P(f) = \frac{1}{2}$ ∵ there's two possibilities girls or boy. $n=4$, $P=\frac{1}{2}$

	0	1	2	3	4
Obs	32	178	290	286	64

$${}^{n}C_x \, p^x (1-p)^{n-x}$$

↳ Formula for binomial

no of boys ⇒ x = 1, 2, 3, 4 = 4 boys 0 is not possible

For 0 ⇒ $P(0) = {}^4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = \frac{1}{16}$ → this is prob but we want number so

$\frac{1}{16} \times 800 = 50$ → Number

For 1 ⇒ $P(1) = {}^4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = \frac{1}{4}$

$\frac{1}{4} \times 800 = 200$

For 2 ⇒ $4C_2 \left(\frac{1}{2}\right)^4 = \frac{6}{16} \times 800 = \boxed{300}$

For 3 ⇒ $4C_3 \left(\frac{1}{2}\right)^4 = \boxed{200}$

For 4 = $\boxed{50}$ → after solve

$$\chi^2 = \frac{(32-50)^2}{50} + \frac{(178-200)^2}{200} + \frac{(290-300)^2}{300} +$$

$$\frac{(236-200)^2}{200} + \frac{(64-50)^2}{50} = 18.93$$

$df = 5-1 = \boxed{4}$

p-value using chi-square is
= 0.00081

which means. $0.00081 < \alpha(0.05)$

reject Null Hypo

Test for Independence

The chi-square test for independence, also known as the chi-square test for association, is a statistical test used to determine whether there is a significant association betn two categorical variables in a sample. It helps to identify if the occurance of one variable is dependent on the occurance of the other variable, or if they are independent of each other.

The test is based on comparing the observed frequencies in a contigency table (a table that displays the frequency distribution of the variable with the frequency that would be expected under the assumption of independent between the two variables.

Pclass ↔ Survived

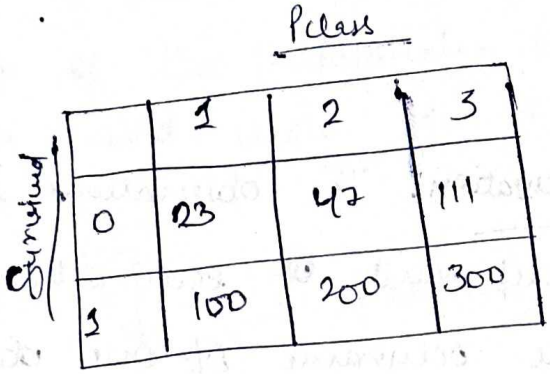
↳ Check whether these col are ~~not related~~ dependent to each other or not.

Steps

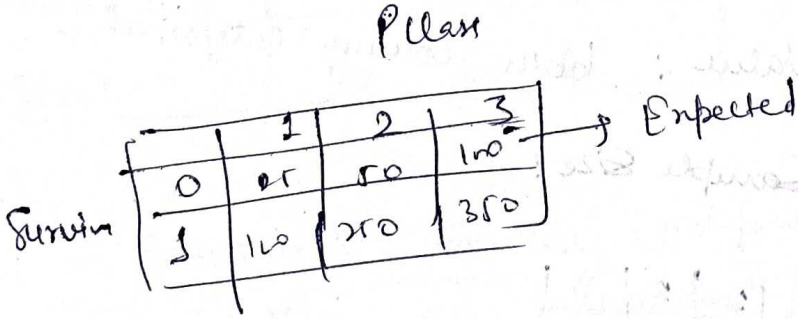
1. State the null hypo (H0) and alternative hypo (H1):
   - H0: There is no association betn the two categorical variables (they are independent).
   - H1: There is an association betn the two

categorical variables (they are dependent).

2. Create a contigency table with the observed frequencies for each combination of the categories of the two variables.



3. Calculate the expected frequencies for each cell in the contigency table assuming that the null hypothesis is true (i.e the variables are independent)



4. Compute the chi-square test statistic:

$$\chi^2 = \sum [(O_i - E_i) / E_i]$$

5. Determine the degree of freedom ⇒ dof = no. of row - 1) * (no. of col - 1)

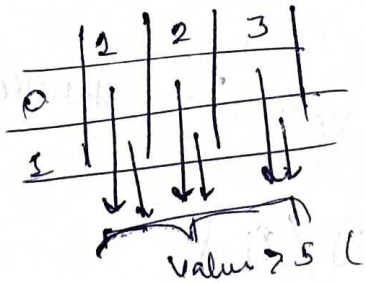
6. Obtain the critical value or p-value using the chi-square distbⁿ table or a statistical software/ calculator with the given degree of freedom and significance level ($\alpha = 0.05$).

Assumptions

1. **Independence of observation:** The observations in the sample should be independent of each other. This means that the occurance of one observation should not affect the occurance of another observation. In practice, this usually implies that the data should be collected using a simple random sampling method.

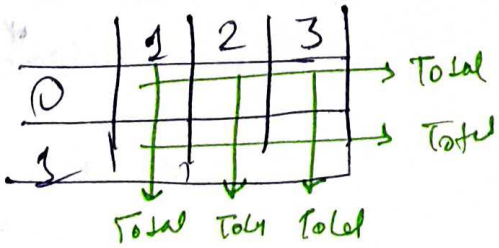
2. Categorical Value : both column categorical

3. Adequate Sample Size :



Value ≥ 5 (every cell value greater than 5)

4. Fixed Margin total



Example 1

A researcher wants to investigate if there is an association betn the level of education (categorical variable) and the preference for a particular type of exercise (categorical variable) among a group of 150 individuals. The researcher collects data and create the following contigency table.

Observed ↓ Exercise Type

Education	Yoga	Running	Swimming	Total
High school	15	20	10	45
Bachelor's	20	30	15	65
Master's or PHD	5	15	20	40
Total	40	65	45	150

Education ⟷ Exercise (independent)

H0: They are independent

H1: They are ASSOciated

expected ↓

	Yoga	Run	Swim
High	$\frac{45 \times 40}{150}$	$\frac{65 \times 45}{150}$	$\frac{45 \times 45}{150}$
Bach	—	—	—
PhD	—	—	—

High school → $\frac{45}{150}$ × $\frac{40}{150}$ ← yoga, × 150

Prob

multiply 150 bcz humme number chahiye prob nhi.

expected

	Yoga	Gym	Swim
High	12	19	13.5
Ball	17	28	20
Pool	10	17	12

$$\sum \frac{(O_i - E_i)}{E_i} \Rightarrow \left(\frac{15-12}{15}\right)^2 + \left(\frac{20-19}{20}\right)^2 + \dots$$

$$\chi^2 = 9.95$$

$$df = (3-1)(3-1) = 4$$

P Value = 0.04

P Value $\leq$ 0.05 (reject Null Hypo)

Applications in Machine Learning

1. **Feature Selection:** Chi-square test can be used as a filter-based feature selection method to rank and select the most relevant categorical features in a dataset. By measuring the association betn each categorical feature and the target variable, you can eliminate irrelevant or redundant features, which can help improve the performance and efficiency of machine learning models.

2. **Evaluation of classification models:** For multiclass classification problems, the chi-square test can be used to compare the observed and expected class frequencies in the confusion matrix. This can help assess the goodness of fit of the classification model, indicating how well the model's prediction align with the actual class distbn.

3. **Analysing relationships betn categorical features:**

4. **Discretization of continous variables:** When converting continous variable into categorical variables (binning), the chi-square test can be used to determine the optimal number of bins or intervals that best represent the relationship betn the continous variable and the target variable.

5. Variable selection in decision trees: Some decision tree algo, such as the CHAD (Chi-squared Automatic Interaction Detection) algo, use the Chi-square test to determine the most significant splitting variables at each node in the tree. The helps construct more effective and interpretable decision trees.