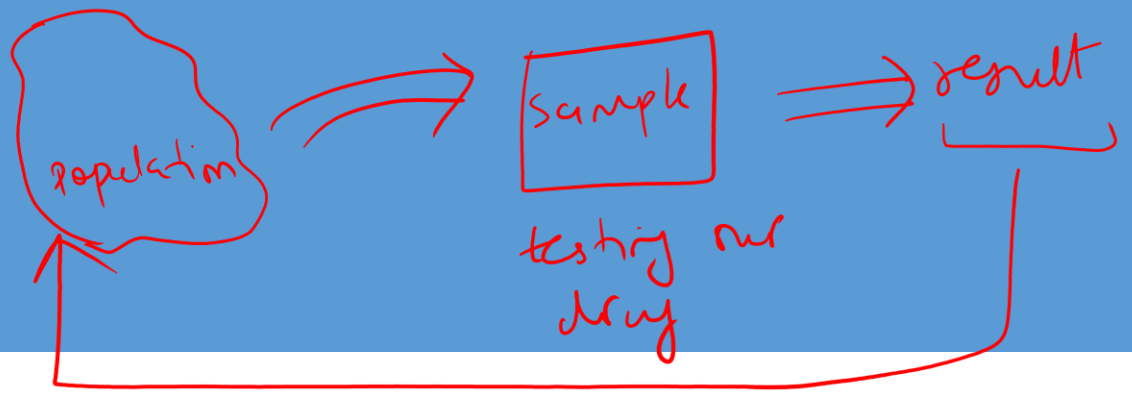
Hypothesis Testing

Author: Rohit Mande
Founder and CEO, inttrvu.ai

Hypothesis Testing



Hypothesis testing is used to test results of survey or experiment
e.g., whether vaccine reduces the chance of getting infected by disease

It is used to know whether results have happened by chance or there is relationship

inferential statistics

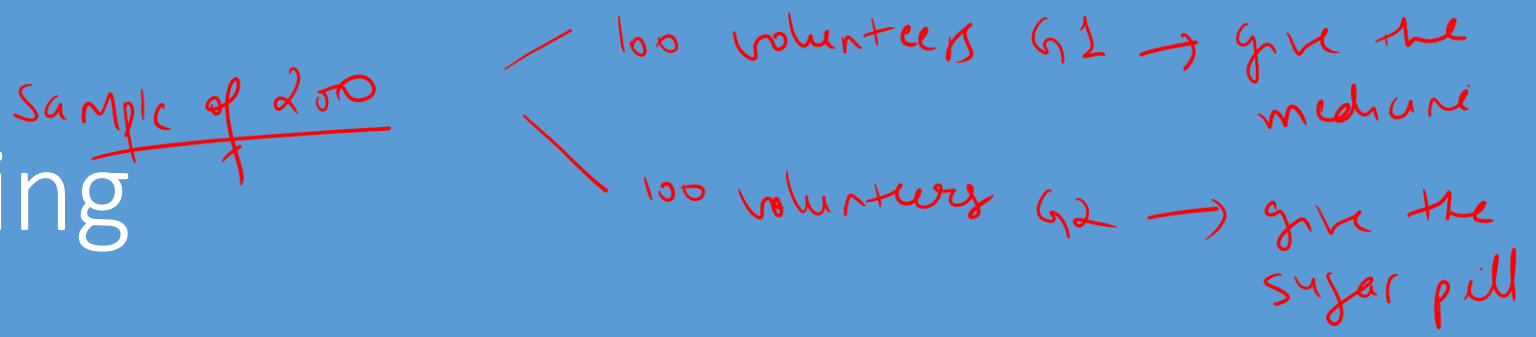
claiming 99% cure for population

assumption or hypothesis

To test the assumption

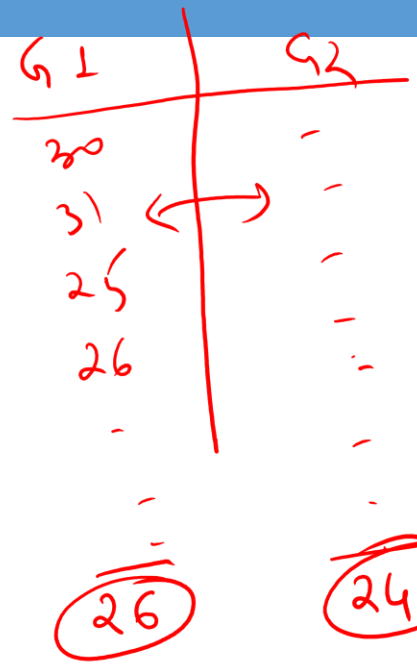
Hypothesis testing

Hypothesis Testing



Null Hypothesis (H_0) :

- Null hypothesis indicates that two possibilities are the same
- It indicates that any observed difference is by chance alone
- Vaccine does not reduce chance of infection



Alternate Hypothesis (H_1) :

- It is the opposite of Null hypothesis
- It indicates that observed difference is not by chance – there is relationship /pattern
- This is the hypothesis researcher are trying to prove
- Vaccine does reduce chance of infection

Hypothesis Testing

probability value

p-value:

- p-value or probability value is a number which indicates how likely it is that the data relationship/pattern) would have occurred by random chance (i.e., null hypothesis is true)
- Level of statistical significance is expressed in terms of p-value which lies between 0 and 1
- Smaller the p-value – smaller the probability of random chance – null hypothesis can be rejected
- Generally, p-value less than 0.05 is statistically significant. It indicates strong evidence against null hypothesis as there is less than 5% chance that the null hypothesis is correct

p_value (probability) → what this value corresponds to?

what this value represents?

Flow of Hypothesis Testing

1) We define our Null hypothesis and design
2) To test the null hypothesis → we perform the experiment
3) We collect the data
4) Calculation of p-value

Step 1 No efficiency in vaccine

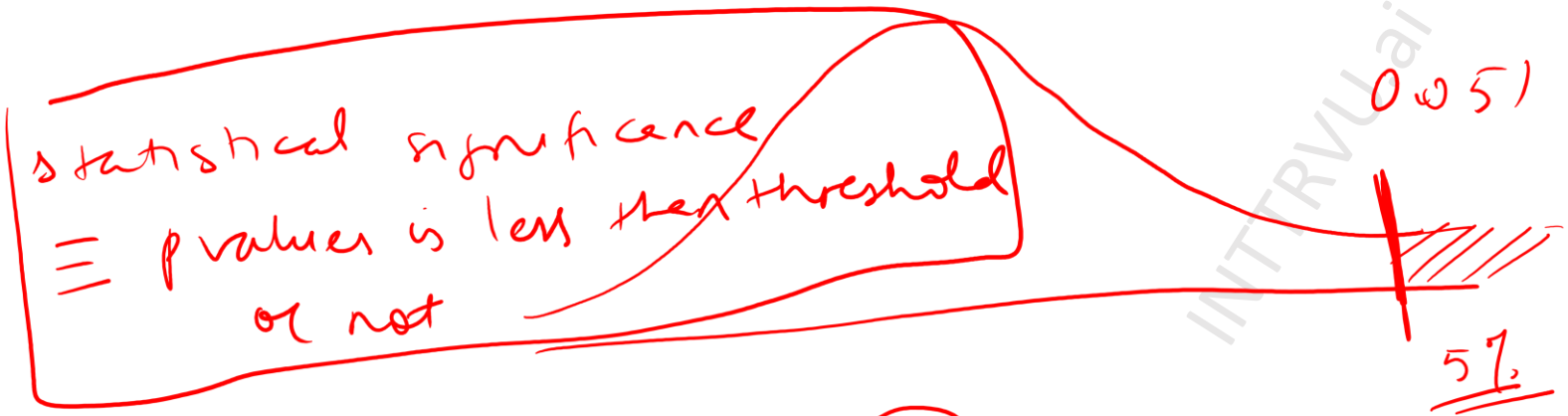
Step 2 We make the groups, decide the number of participants, arrays

Step 3 the volunteers, basic testing
collect the data

4) p-value

Assuming the null hypothesis is true, what is the probability of observing the data that we observed in step 3 ?

If p-value threshold is 0.05 or 5%



	G_1	G_2
P_1	25	26
P_2	25	25
P_3	27	29
⋮	⋮	⋮
P_{10}	28	30
avg Temp	26.7	27.8
	Sugar	Tablet

ΔT $27.8 - 26.7 = 1.1 \Rightarrow$

what is the probability there could be a ΔT of 1.1 given that Tablet has no efficacy ? <5%

to test the relationship / dependency between two categorical variables.

social experiment to understand dependency b/w gender and social media exiting

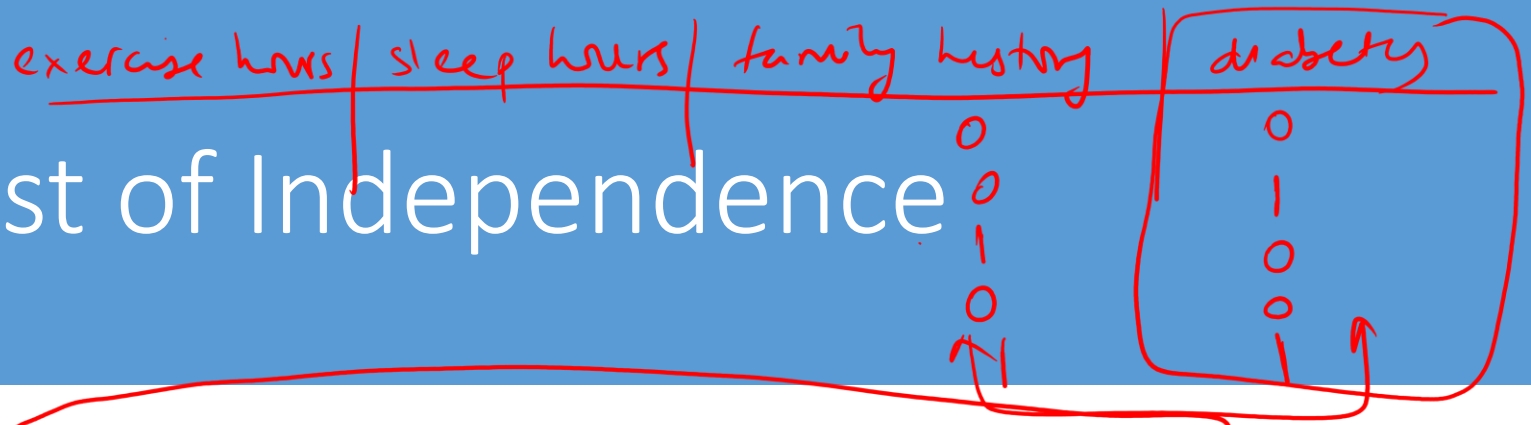
	Gender	Exit status from social media
1	M	Yes
2	F	No
1	M	Yes
2	M	No
2	F	No
2	F	No
1	M	Yes
1	M	Yes

Chi-Square test is one of the methods to test the hypothesis

1) Null hypothesis = exit from social media doesn't depend on gender

2) and 3) design, perform and collect the data

Chi Square Test of Independence



- It is useful for checking relationship between two categorical variables
- It can be used for selecting categorical features in classification model
- It is used to determine whether there is statistically significant difference between expected frequency and observed frequency

E.g., To determine whether Gender feature is important in determining whether a user will continue to be a part of social network group or will exit it

feature / label

10%

Chi Square Test of Independence

1. Define Null Hypothesis

- Null hypothesis:

  Two variables are independent.

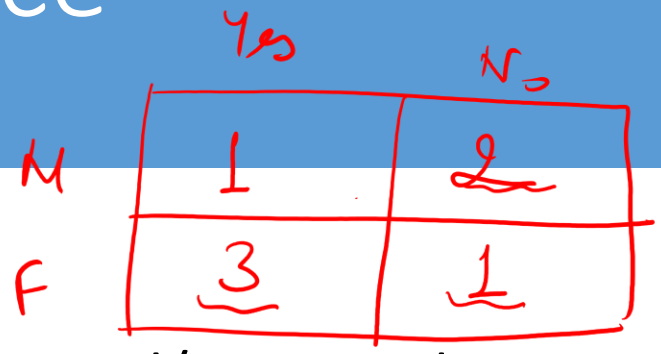
  Gender has no role in determining whether a user will continue to be a part of social network group

- Alternate hypothesis:

  Two variables are dependent.

  Whether a user will continue exploring social network group is dependent on gender of user

Chi Square Test of Independence

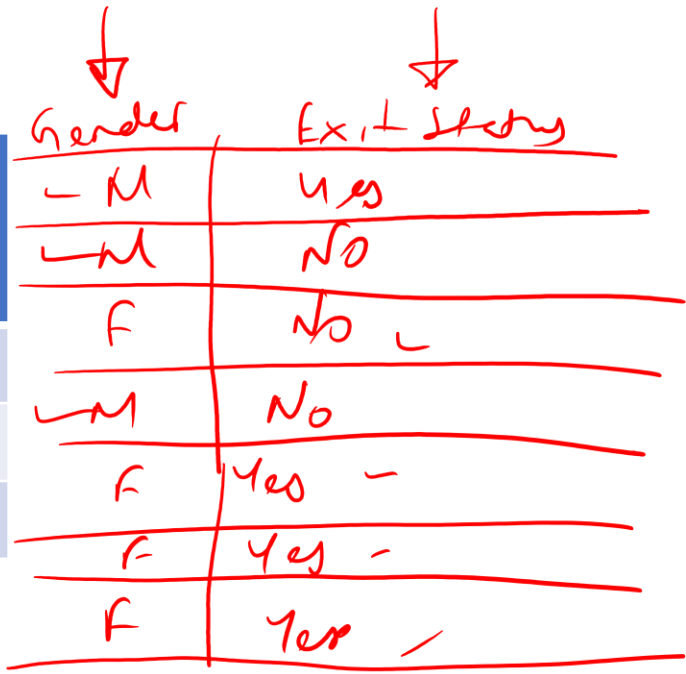


2. Create Contingency table

Suppose we have data for 400 users which have used and exited/continued using social network group in the past

observed data / frequency

Exited → / Gender ↓	Yes	No	Total
Female	120	60	180
Male	105	115	220
Total	225	175	400



Chi Square Test of Independence

Expected frequency

$O_1 - E_1$
$O_2 - E_2$
$O_3 - E_3$
$O_4 - E_4$

3. Calculate Expected Frequencies

E1 = Expected Frequency for 'Gender=Female' and 'Exited = Yes'

P1 = P(Yes) * P(Female)
= (225/400) * (180/400)
= 0.253

E1 = P1 * total population = 0.253 * 400 = ~101

Exited / Gender	Yes	No	Total
Female	**120** (E_1)	60 (E_2)	180
Male	105 (E_3)	115 (E_4)	220
Total	225	175	400

E_1 = (female) and (Yes)
= P(female) P(Yes) Total population
= $\frac{180}{400} \times \frac{225}{400} \times 400$

Similarly expected values for other cells can be calculated

	Yes	No
Female	E_1	E_2
Male	E_3	E_4

Assumption: No connection b/w gender and exit status

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P((\text{Gender} = \text{female}) \cap (\text{exit status} = \text{Yes})) = P(\text{Female})\ P(\text{exit} = \text{Yes})$$

$$\text{Prob.} = \frac{180}{400} \times \frac{225}{400}$$

$$E_1 = \text{prob} \times 400 = \frac{180}{400} \times \frac{225}{400} \times 400 = 101$$

$E_2 = P(\text{female}) \times P(\text{No}) \times \text{Population size}$

$= \frac{180}{400} \times \frac{175}{400} \times 400$

$E_3 = P(\text{male}) P(\text{Yes}) \times \text{Pop. size}$

$= \frac{220}{400} \times \frac{225}{400} \times 400$

$E_4 = P(\text{male}) \times P(\text{No}) \times \text{Pop. size}$

$= \frac{220}{400} \times \frac{175}{400} \times 400$

Find a way to condense all 4 differences together

$diff1 = O_1 - E_1$

$diff2 = O_2 - E_2$

$diff3 = O_3 - E_3$

$diff4 = O_4 - E_4$

Chi Square Test of Independence

2 x 2
i = 1 to 4
50 x 100
i = 1 to 5000

chi square statistic

4. Calculate chi-square statistic

The formula for Chi Square is

$$x_c^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$\sum$ = sum

$$\chi_c^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

chi square value / statistic

degree of freedom

i = 1 to n denotes total number of possibilities (number of cells in contingency table)

c = degrees of freedom

= (number of rows in contingency table -1) * (number of columns in contingency table -1)

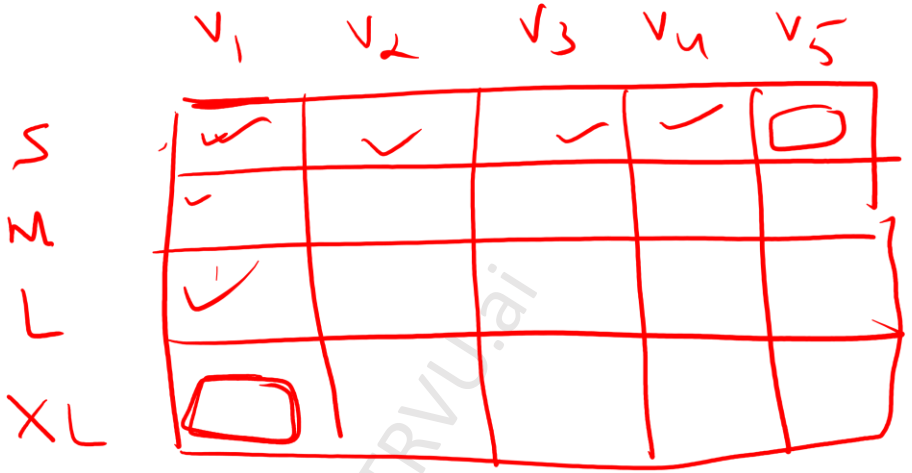
O = Observed Frequencies

E = Expected Frequencies

degree of freedom

χ^2 → degree of freedom

$C = (\text{number of rows} - 1) * (\text{number of columns} - 1)$



3

1 4 5 10 10

5 − 1 = 4

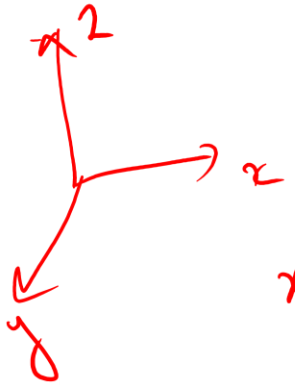
10 15 5

8

2 20

3 − 1

30



$x + y + 2 = 10$

Chi Square Test of Independence

$(2-1)^* (2-1) = 1$

5. Calculate p-value

$\chi_c^2 = 8.5$
$c-3$

Refer chi-square table to get p-value for given score and degrees of freedom

In this case if p-value < 0.05 Null Hypothesis can be rejected with greater than 95% confidence

It will indicate that whether a user will continue exploring social network group is dependent on gender of user

Gender of user is important feature in predicting whether user will exit the group or not

Degrees of freedom (*df*)	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666

Example of chi-square table

Classification model

Diabetes prediction

Social Task

Weight	Numerical	Label
	70	0
	80	1
	130	0
	140	0
	60	1
	50	0

Step size

Step size depends on gender?

Welch t-test

Gender	Step size
M	71 cm
F	60 cm
.	.
.	.

weight of diabetic patients	weight of healthy
80	70
	130
60	140
	50

Arg 1

Arg 2

Std 1

Std 2

Welch t-test

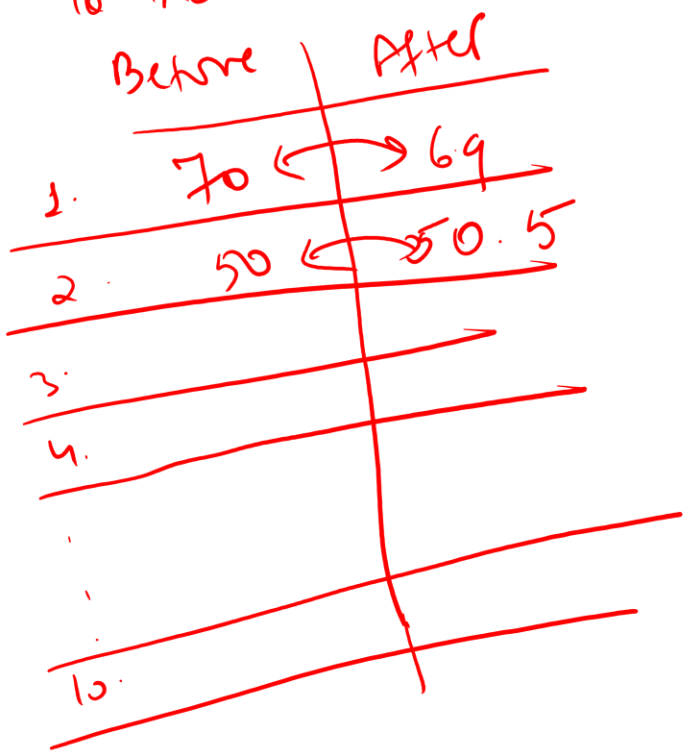
- It can be used for selecting numerical features in classification model
- It used to determine whether means of two populations is different by chance or with high confidence
- Welch t-test is used when two populations are of different size and equal/unequal variance which is usually the case in real world datasets
- This test is called 'unpaired' or 'independent samples' t-test because two sample populations being compared are non-overlapping (different samples)

Example:

- Suppose we want to know whether BMI is useful feature in predicting whether a person will have diabetes or not then welch t–test can be used.
- Here two population to be compare will be BMI of population of people with diabetes and BMI of population of people without diabetes

T- test

- paired
  - G1 ⇒ acted medicine
  - to individuals



- Non paired
  - Example ① weight comparison for diabetic vs healthy
  - ② step size vs gender

Welch t-test

Avg. BMI of healthy individuals =
Avg BMI of diabetic patients.

1. Define Null Hypothesis

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

step 1 Calculate mean
step 2 Calculate deviation from mean
step 3 Square and take their average
step 4 Take a square root

➢Null hypothesis:

Two sample populations have equal mean.

BMI of people having diabetes has equal mean as BMI of people not having diabetes

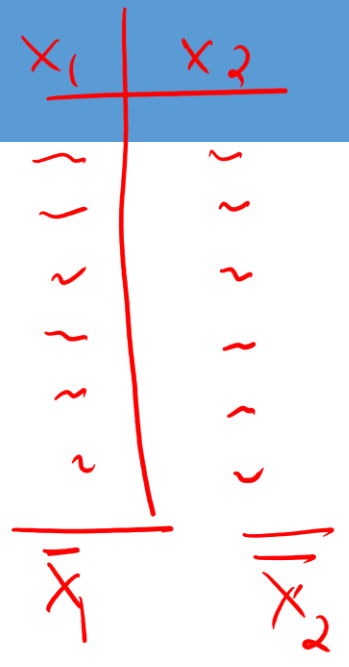
➢Alternate hypothesis:

Two sample populations have different mean.

BMI of people having diabetes doesn't have equal mean as BMI of people not having diabetes

Welch t-test

X_1 = BMI of diabetic patients
X_2 = BMI of healthy individual



2. Calculate t statistic

$S_{X_1} = \frac{S_1}{\sqrt{N_1}}$ ← standard deviation of X_1

$S_{X_2} = \frac{S_2}{\sqrt{N_2}}$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2}}$$

$$S_{\bar{X}_i} = \frac{S_i}{\sqrt{N_i}}$$

$\bar{X}_i$ is the sample mean of population i (with diabetes population 1 and without diabetes population 2)

$S_{\bar{X}_i}$ is the standard deviation of the sample population X_i → averages of

Degrees of freedom is calculated using formula and its dependent on $S_{\bar{X}_i}$ and N_i (number of samples in each population)

S_1 = standard deviation of sample X_1

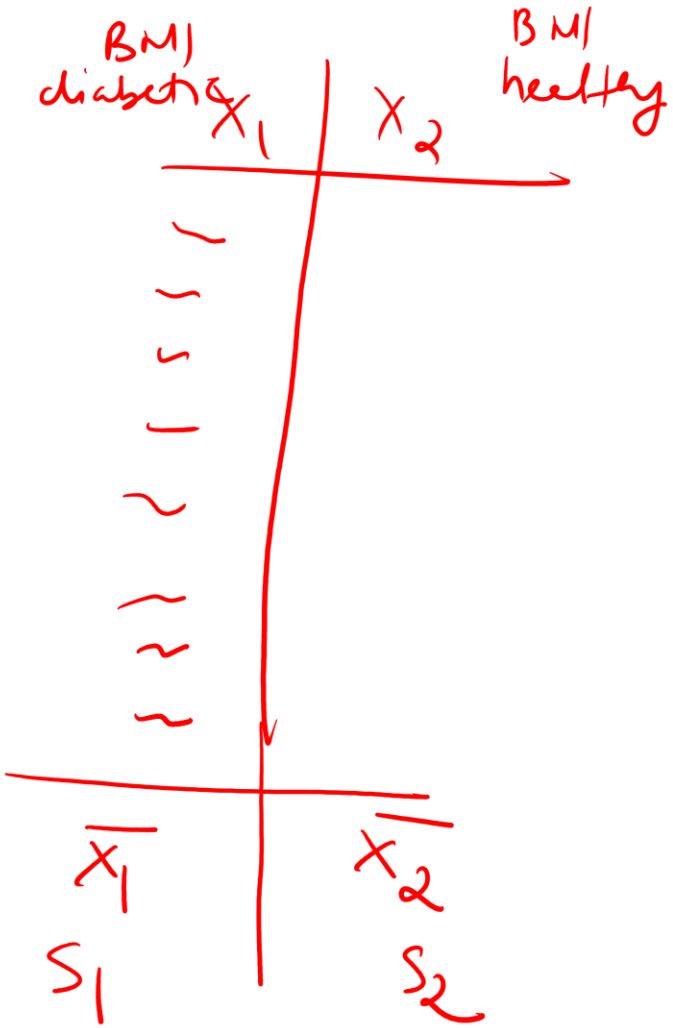
$$S_{\overline{X_1}} = \frac{S_1}{\sqrt{N_1}}$$

N_1 = number of samples in X_1

S_2 = standard deviation of sample X_2

$$S_{\overline{X_2}} = \frac{S_2}{\sqrt{N_2}}$$

N_2 = # samples in X_2



$$df = \frac{\left(\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}\right)^2}{\frac{\left(\frac{S_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{S_2^2}{N_2}\right)^2}{N_2 - 1}}$$

Welch t-test

Based on t-statistic and ~~feature~~ degree of freedom

3. Accept or reject null hypothesis

Refer welch t-test table to get p-value for given t-statistic and degrees of freedom

If p-value < 0.05 then Null hypothesis of equal mean can be rejected

It will indicate that BMI of people having diabetes has different mean than BMI of people not having diabetes with more than 95% confidence (not by random chance)

BMI is useful feature for predicting diabetes

Questions?

INTTRVU