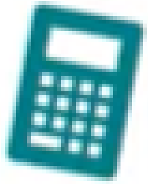
Non parametric tests

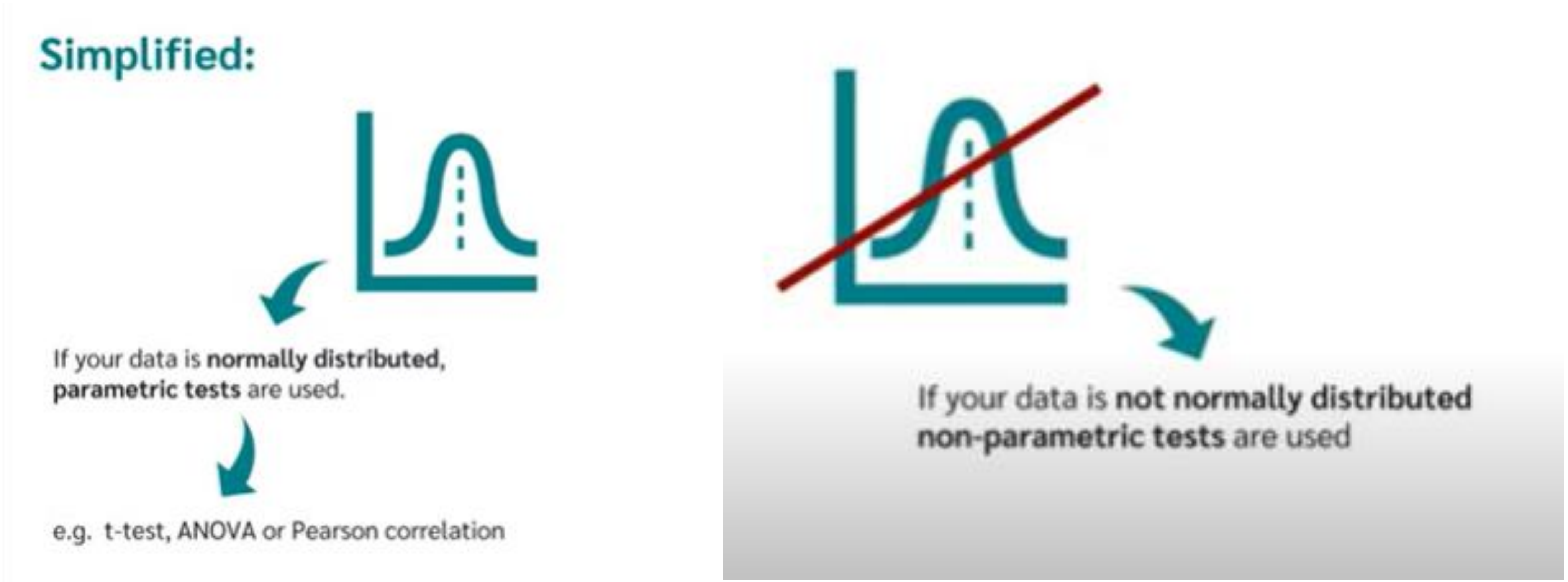
ANISHA JOSEPH

Parametric and non-parametric tests



If you want to calculate a **hypothesis test**, you must first check the **assumptions**.

One of the most common **assumptions** is that the data used must show a **certain distribution**, usually the **normal distribution**.


Simplified:
If your data is normally distributed, parametric tests are used.
e.g. t-test, ANOVA or Pearson correlation
If your data is not normally distributed non-parametric tests are used

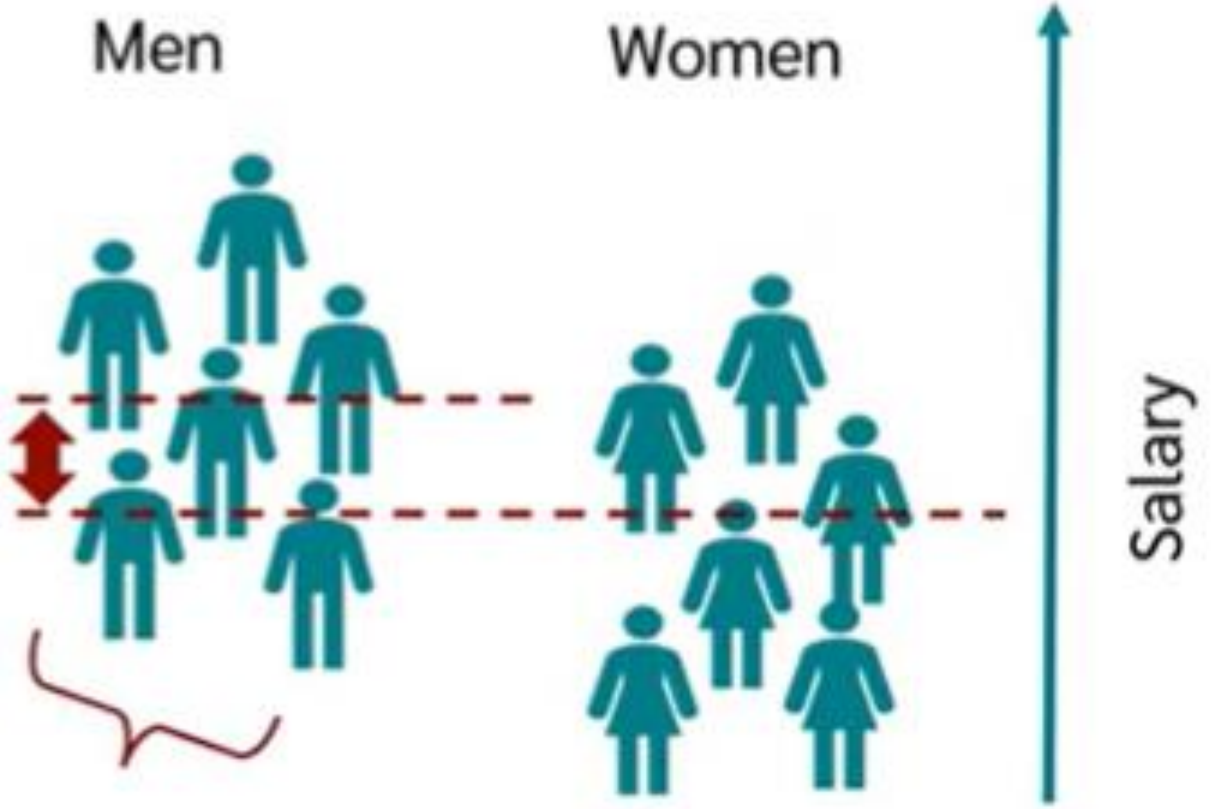
H0 Null hypothesis

You have formulated your null hypothesis

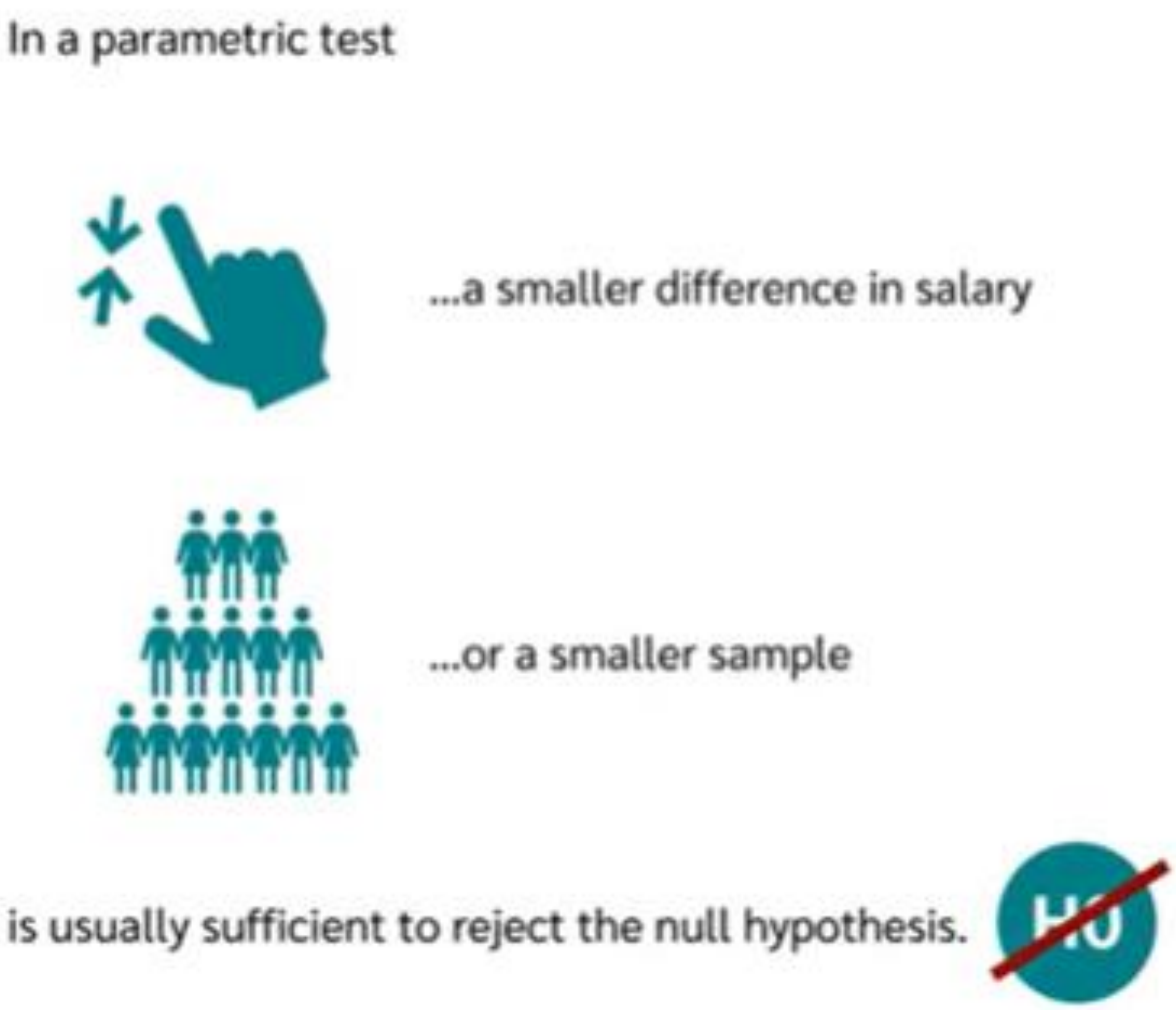
e.g. The salary of men and women does not differ.

Whether the null hypothesis is rejected depends, on other things...

... on the difference in salary



...and on the sample size


In a parametric test
...a smaller difference in salary
...or a smaller sample
is usually sufficient to reject the null hypothesis.
H0

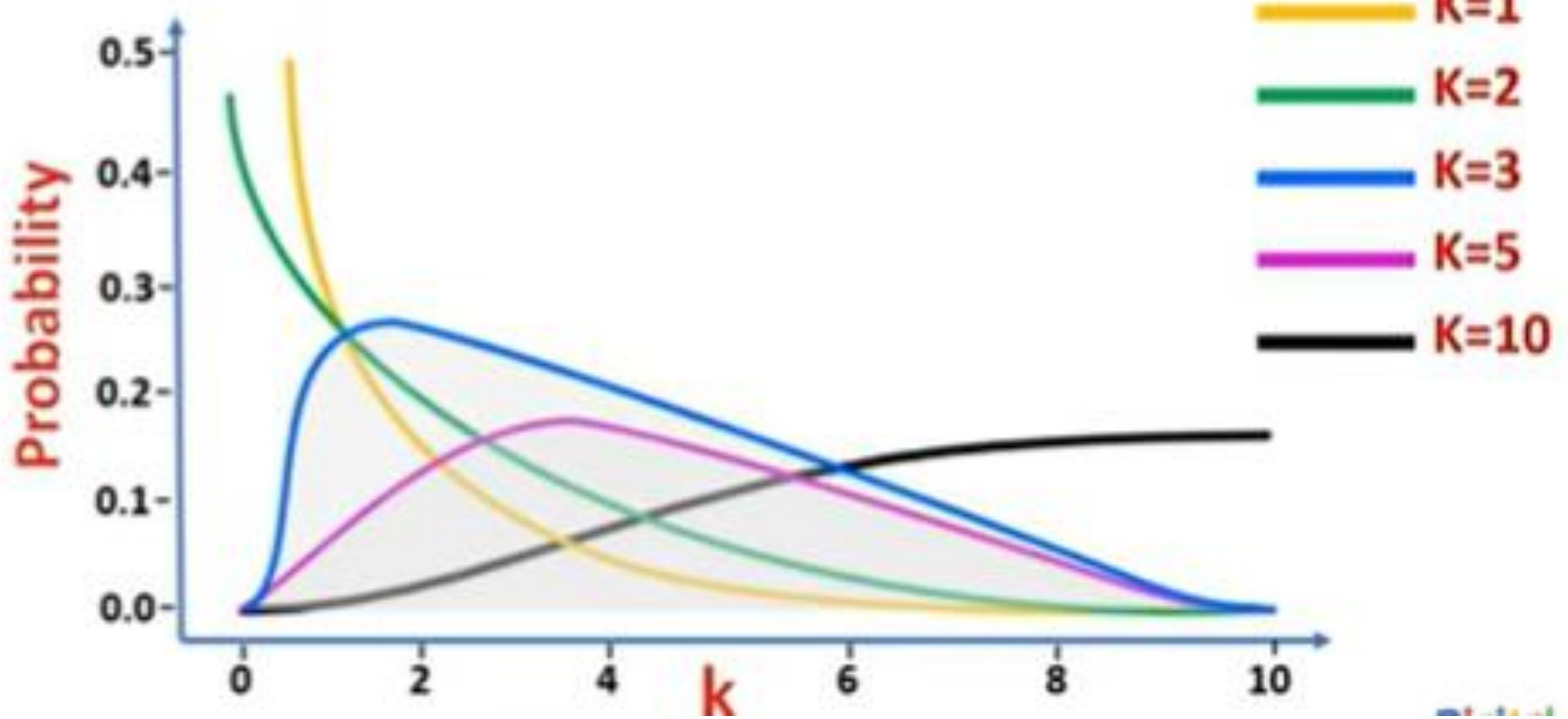
Chi Square Distribution : Goodness of Fit

- Chi-square (χ^2) distribution was discovered by :
- It is a goodness of fit test for a categorical variable
- Think of this a "Square" of standard normal distribution.

$$\chi^2$$

Chi-Square Distributions Shape

The chi-square distributions are a family of distributions that take only positive values and are skewed to the right. A particular chi-square distribution is specified by giving its degrees of freedom.



Chi Square Distribution : Goodness of Fit

$$\chi^2 = \frac{\Sigma(O - E)^2}{E}$$

$$\chi^2 = \Sigma \frac{(\text{Observed - Expected})^2}{\text{Expected}}$$

where the sum is over all possible values of the categorical variable.

χ^2 is pronounced "chi-square"

- **Observed frequencies** – Obtained frequency for each category
- **Expected frequencies** – Hypothesized frequency for each category.

Assume Expected Frequencies >5

Chi Square Distribution : Goodness of Fit

$$\chi^2 = \frac{\Sigma(O - E)^2}{E}$$

$$\chi^2 = \Sigma \frac{(\text{Observed - Expected})^2}{\text{Expected}}$$

where the sum is over all possible values of the categorical variable.

Acceptance Criteria

χ^2 Critical $(\alpha, k-1)$ <= χ^2 Test statistic

- Reject Null hypothesis.

χ^2 Critical $(\alpha, k-1)$ > χ^2 Test statistic

- Fail to reject Null hypothesis.

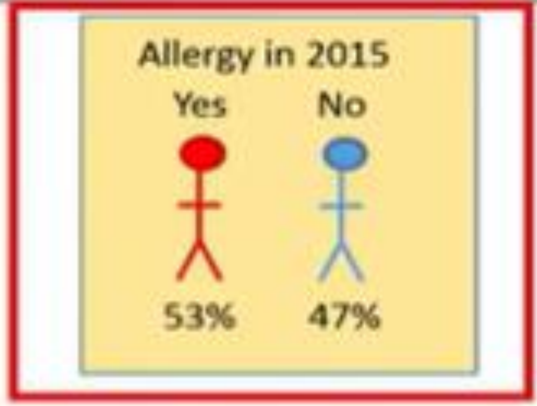
α = Level of Significance

k = Degree of Freedom

Chi Square Distribution : Goodness of Fit

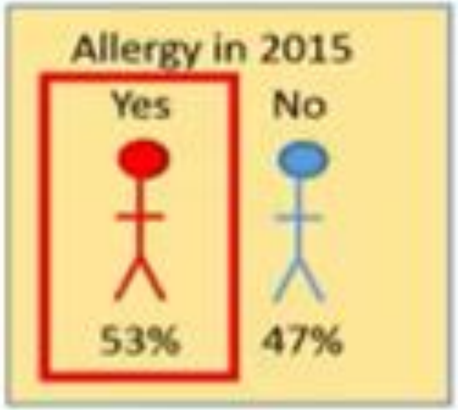
- All chi-square values χ^2 are greater than or equal to zero
- The area under each curve of the chi-square distribution equals one.
- Chi-square distribution is always positively skewed.
- Chi-square distribution is different for each number of degrees of freedom (K)
- Because differences from expected values are squared, the value of χ^2 cannot be negative.

Chi-square goodness of fit test



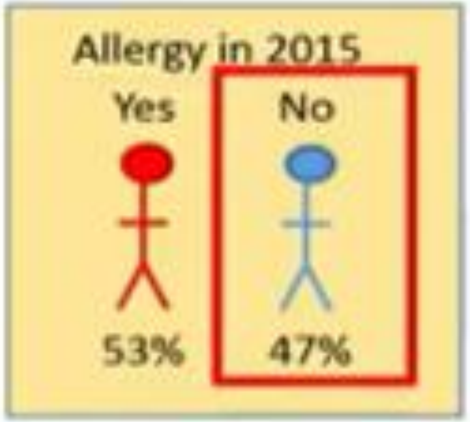
In an extensive study from 2015, a group of investigators analyzed how common allergies were in a certain population.

Chi-square goodness of fit test



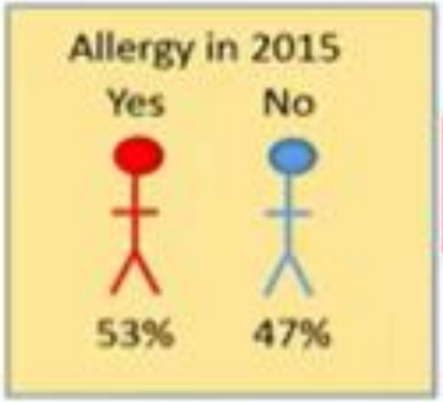
The study, which covered the complete population, found that 53% had some sort of allergy.

Chi-square goodness of fit test



whereas 47% did not have an allergy.

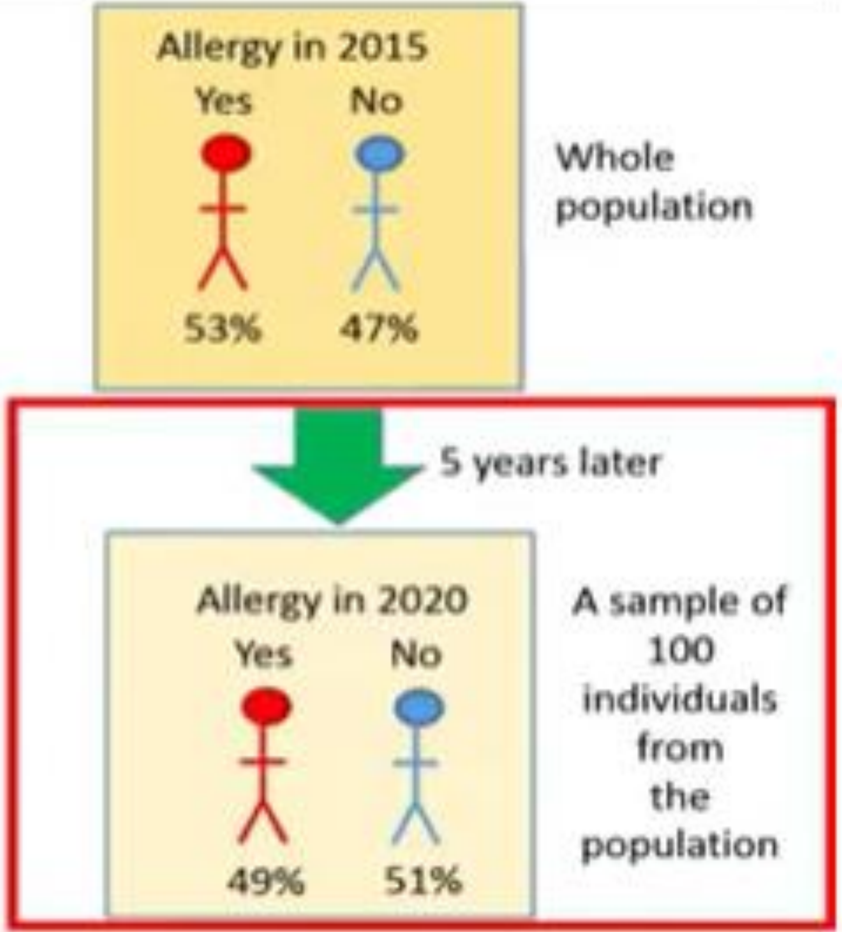
Chi-square goodness of fit test



Whole population

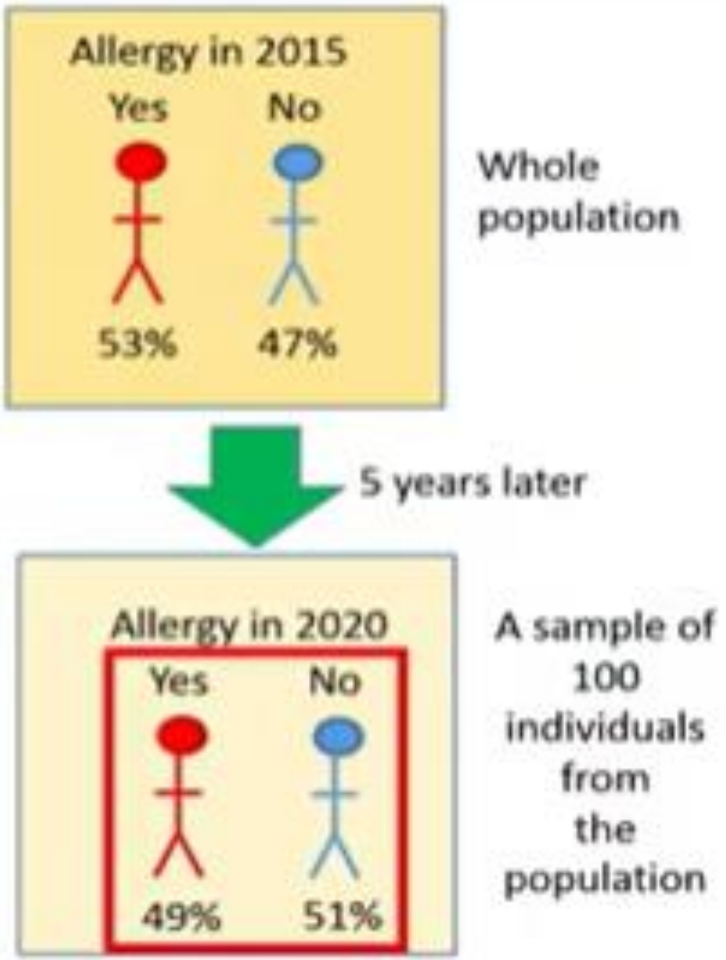
Note that this is not a sample since the whole population was analyzed.

Chi-square goodness of fit test



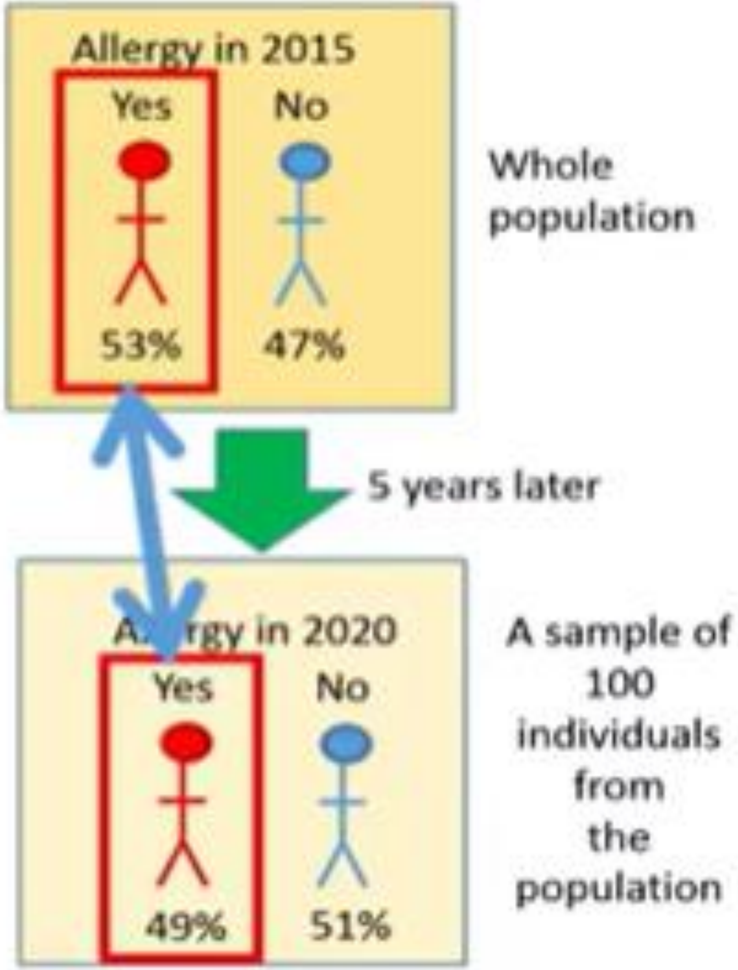
Five years later, the investigators wanted to see if the proportion of allergy in the population had changed since 2015. However, this time, the investigators did not have enough resources to analyze the whole population.

Chi-square goodness of fit test



In this sample, 49 persons, or 49%, had some sort of allergy, whereas 51 individuals did not have an allergy.

Chi-square goodness of fit test

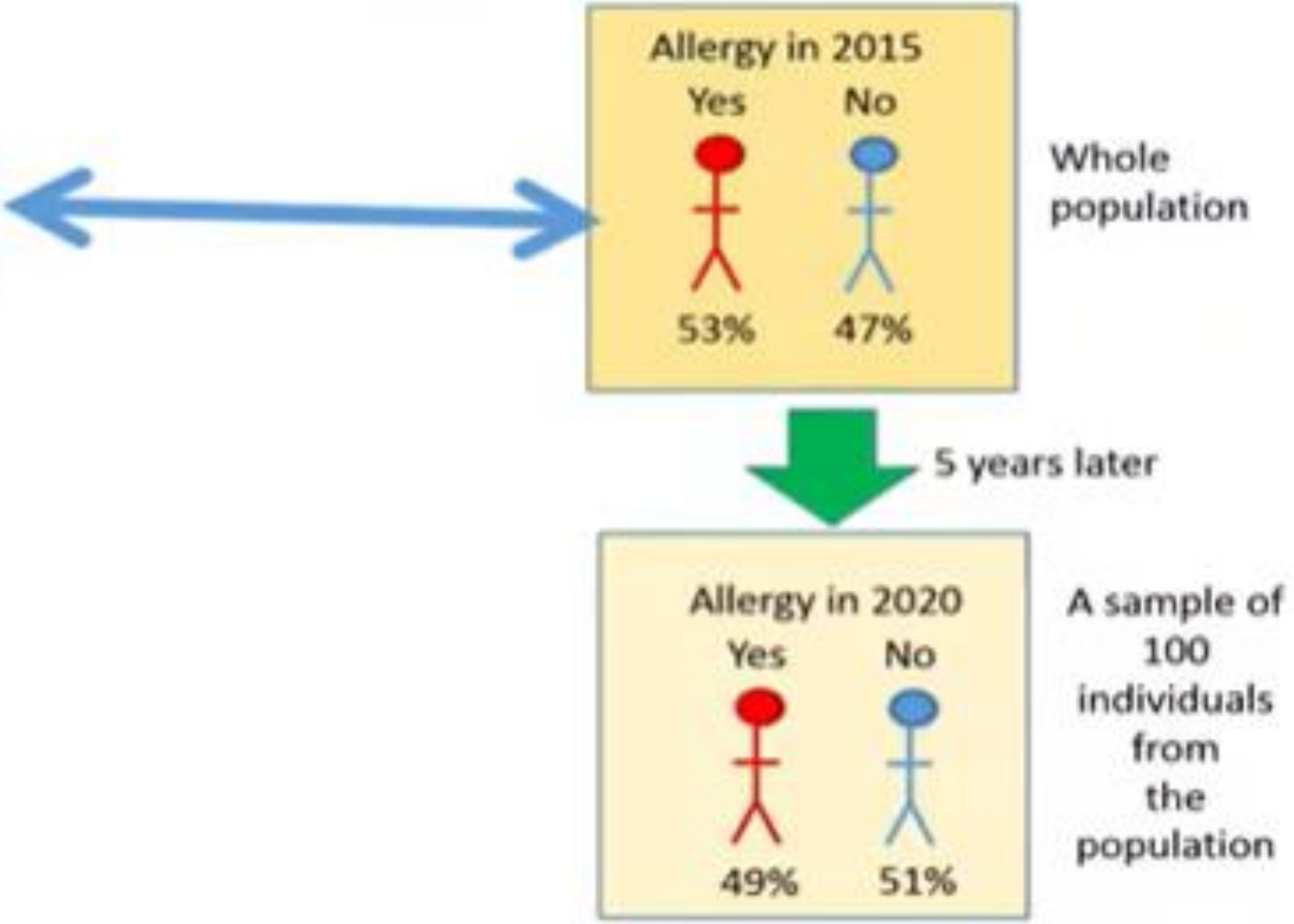


Based on this sample, we like to know if the proportion of allergic people in the population has changed over the five years.

Chi-square goodness of fit test

$$\boxed{H0:(p_1, p_2) = (53\%,\ 47\%)}$$

$$H1:(p_1, p_2) \neq (53\%,\ 47\%)$$

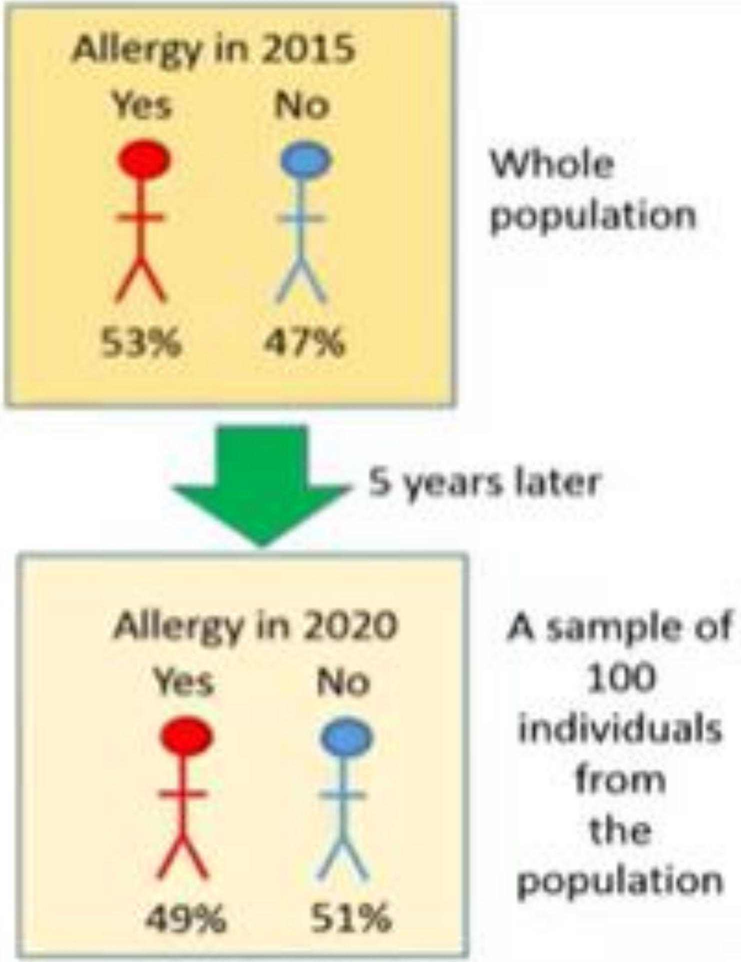


The null hypothesis of this test states that the proportions, or frequencies, of individuals with allergies, has not changed since 2015,

Chi-square goodness of fit test

$$H0: (p_1, p_2) = (53\%,\ 47\%)$$

$$\boxed{H1: (p_1, p_2) \neq (53\%,\ 47\%)}$$



whereas the alternative hypothesis states that there is a change, or that the population proportions in year 2020 are not identical to the population proportions from 2015.

Chi-square goodness of fit test

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$n = 100$$
$$\boxed{p_0 = 0.53}$$
$$\hat{p} = 0.49$$

$$O_1 = 49 \qquad E_1 = 53$$
$$O_2 = 51 \qquad E_2 = 47$$

Chi-square goodness of fit test

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} = \frac{(49-53)^2}{53} + \frac{(51-47)^2}{47} = 0.64$$

$$\boxed{df = k - 1}$$

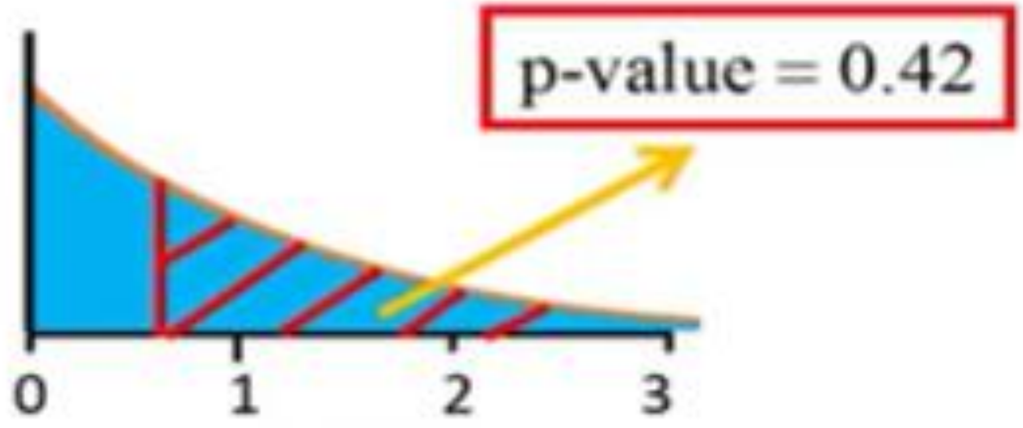
$$O_1 = 49 \qquad E_1 = 53$$

$$O_2 = 51 \qquad E_2 = 47$$

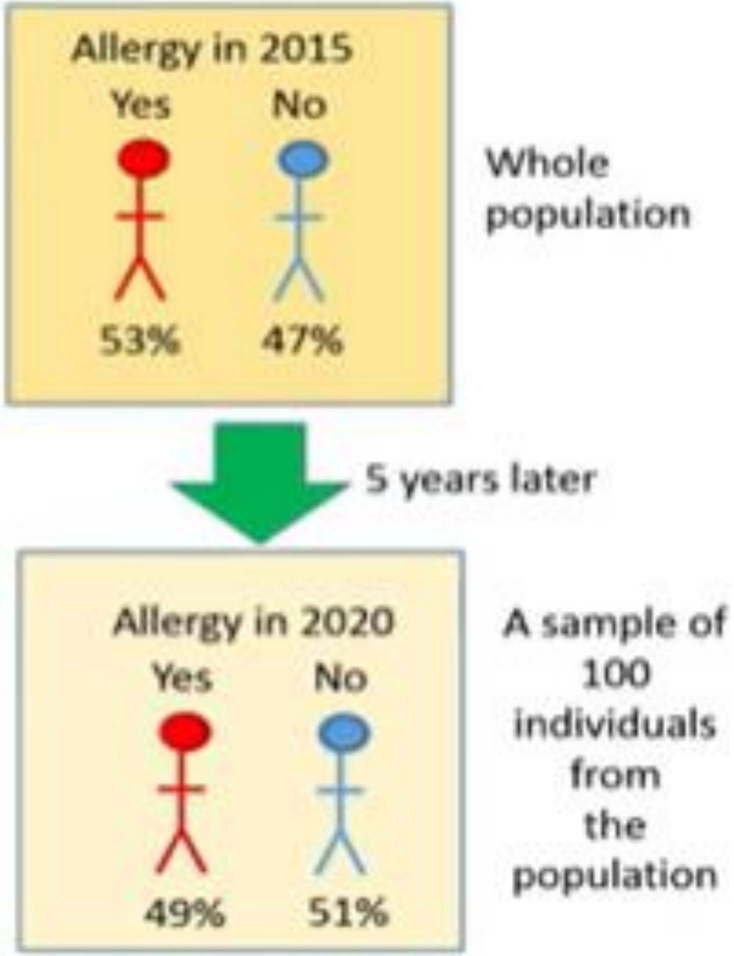
Chi-square goodness of fit test

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(49-53)^2}{53} + \frac{(51-47)^2}{47} = 0.64$$



Chi-square goodness of fit test



$$\boxed{H0:(p_1, p_2) = (53\%,\ 47\%)}$$
$$H1:(p_1, p_2) \neq (53\%,\ 47\%)$$

Since the p-value is greater than the general significance level of 0.05, we do not reject the null hypothesis. We can thereby conclude that we do not have enough evidence to say that the proportion of people with allergy has changed since 2015.

Example 1

Some research shows higher number of flight tickets are brought by males in comparison to Females with ratio of 2:1. Out of 150 tickets, 88 tickets were bought by males and 62 by females. We need to find out if the experimental manipulation causes the change in the results, or we are observing a chance variation.

Example 1

Some research shows higher number of flight tickets are brought by males in comparison to Females with ratio of 2:1. Out of 150 tickets, 88 tickets were bought by males and 62 by females. We need to find out if the experimental manipulation causes the change in the results, or we are observing a chance variation.

$H_o : p_o = p_1 \qquad H_o : p_o \neq p_1$

Ans:

$$\chi^2 = \frac{\Sigma(O-E)^2}{E}$$

Male		Female	
$O_m = 88$	$E_m = \frac{2}{3} * 150 = 100$	$O_f = 62$	$E_f = \frac{1}{3} * 150 = 50$

$$\chi^2 = \frac{(O_m - Em)^2}{E_m} + \frac{(Of - Ef)^2}{E_f}$$

$$\chi^2 = \frac{(88-100)^2}{100} + \frac{(62-50)^2}{50}$$

$$\chi^2 = 1.44 + 2.88 = 4.32$$

Degrees of freedom = n-1 = 2-1 = 1

$$\chi^2_{\text{Critical}\ (\alpha, k-1)} = \chi^2_{(0.05, 2-1)} = 3.841$$

$$\chi^2_{\text{Critical}\ (3.841)} <= \chi^2_{\text{Test statistic}\ (4.32)}$$

Reject Null hypothesis.

Example 2

Non Conformity	Observed Frequency (Oi)	Expected Frequency (Ei)
Crack	27	30
Tear	53	45
Damage	20	25
Total	100	

Example 2

$$\chi^2 = \frac{\Sigma(O-E)^2}{E}$$

Non Conformity	Observed Frequency (Oi)	Expected Frequency (Ei)
Crack	27	30
Tear	53	45
Damage	20	25
Total	100	

Ans:

$H_o : p_i = p_{i,0}$

$H_a : p_o \neq p_{i,0}$ *for at least one* $i = 1,2,3\ldots K$

Degrees of freedom = n-1 = 3-1 = 2

$$\chi^2 = \frac{(27-30)^2}{30} + \frac{(53-45)^2}{45} + \frac{(20-25)^2}{25}$$

$$\chi^2_{\text{Critical }(\alpha,k-1)} = \chi^2_{(0.05,2)} = 5.991$$

$$\chi^2 = 0.3 + 1.422 + 1 = 2.722$$

$$\chi^2_{\text{Critical }(5.991)} > \chi^2_{\text{Test statistic }(2.722)}$$

Fail to reject Null hypothesis

Chi-square test of independence

Test of independence

n=8100

	Cancer	No cancer	Total
Smoker	60	800	860
No smoker	40	7200	7240
% smoker	60%	10%	
Total	100	8000	8100

As an example of a chi-square test of independence, let's say that we have collected 8100 individuals from the population and noted smoking status and if they have cancer or not.

Chi-square test of independence

Test of independence

n=8100

H0: There is **no** association between smoking and cancer

H1: There is an association between smoking and cancer

	Cancer	No cancer	Total
Smoker	60	800	860
No smoker	40	7200	7240
% smoker	60%	10%	
Total	100	8000	8100

Chi-square test of independence

Test of independence

n=8100

	Cancer	No cancer	Total
Smoker	60	800	860
No smoker	40	7200	7240
% smoker	60%	10%	
Total	100	8000	8100

We can see that out of the ones who have cancer, 60% are smokers,

Chi-square test of independence

Test of independence

n=8100

	Cancer	No cancer	Total
Smoker	60	800	860
No smoker	40	7200	7240
% smoker	60%	10%	
Total	100	8000	8100

whereas only 10% are smokers in the group who do not have cancer.

$$E_{1,1} = \frac{R_1}{N} \cdot \frac{C_1}{N} \cdot N$$

$$df = (r-1)(c-1)$$

Chi-square test of independence

Test of independence

n=8100

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 260$$

$$\boxed{p < 0.001}$$

	Cancer	No cancer	Total
Smoker	60	800	860
No smoker	40	7200	7240
% smoker	60%	10%	
Total	100	8000	8100

The p-value is smaller than the general significance level of 0.05, which means that we can reject the null hypothesis and conclude that there is an association between smoking and cancer.

Chi-square test of independence

Although it is tempting to conclude that smoking causes cancer, the chi-square test only tells us that there is an association, noting about causality.

Chi-squared tests

Test for independence

120 people are surveyed for their preferred social media platform.

Is there enough evidence to suggest social media preference is independent of sex?

Observed frequencies

	Male	Female	TOTAL
Facebook	15	20	35
Instagram	30	35	65
Tik Tok	5	15	20
TOTAL	50	70	120

Chi-squared tests

Test for independence

120 people are surveyed for their preferred social media platform.

Is there enough evidence to suggest social media preference is independent of sex?

Observed frequencies

	Male	Female	TOTAL
Facebook	15	20	35
Instagram	30	35	65
Tik Tok	5	15	20
TOTAL	50	70	120

Expected frequencies

	Male	Female	TOTAL
Facebook	14.6	20.4	35
Instagram	27.1	37.9	65
Tik Tok	8.3	11.7	20
TOTAL	50	70	120

H_0: Social media preference is independent of sex

H_1: Social media preference is NOT independent of sex

Chi-squared distribution

Test for independence

120 people are surveyed for their preferred social media platform.

Is there enough evidence to suggest social media preference is independent of sex?

Observed frequencies

	Male	Female	TOTAL
Facebook	15	20	35
Instagram	30	35	65
Tik Tok	5	15	20
TOTAL	50	70	120

Expected frequencies

	Male	Female	TOTAL
Facebook	14.6	20.4	35
Instagram	27.1	37.9	65
Tik Tok	8.3	11.7	20
TOTAL	50	70	120

H_0: Social media preference is independent of sex

H_1: Social media preference is NOT independent of sex

Test statistic

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

df = (r-1)(c-1)

Test for independence

120 people are surveyed for their preferred social media platform.

Is there enough evidence to suggest social media preference is independent of sex?

Observed frequencies

	Male	Female	TOTAL
Facebook	15	20	35
Instagram	30	35	65
Tik Tok	5	15	20
TOTAL	50	70	120

Expected frequencies

	Male	Female	TOTAL
Facebook	14.6	20.4	35
Instagram	27.1	37.9	65
Tik Tok	8.3	11.7	20
TOTAL	50	70	120

H_0: Social media preference is independent of sex

H_1: Social media preference is NOT independent of sex

Test statistic

$$\chi^2 = \sum \frac{(o-e)^2}{e} \sim \chi^2_2 \quad (\textit{As all expected frequencies are} > 5)$$

$$= \frac{(15-14.6)^2}{14.6} + \frac{(30-27.1)^2}{27.1} + \cdots$$

	Male	Female	TOTAL
TOTAL	50	70	120

	Male	Female	TOTAL
TOTAL	50	70	120

preference is independent of sex

Test statistic

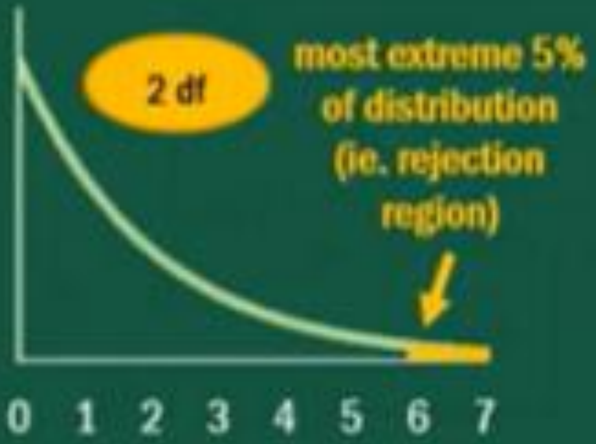
$$\chi^2 = \sum \frac{(o-e)^2}{e} \sim \chi^2_2 \quad \textit{(As all expected frequencies are > 5)}$$

$$= \frac{(15-14.6)^2}{14.6} + \frac{(30-27.1)^2}{27.1} + \cdots = 2.84$$

Decision rule

Reject if $\chi^2 > 5.99$

=CHISQ.INV(0.95,2)



Conclusion

Do not reject H_0 *as* $\chi^2 < 5.99$.
Not enough evidence to suggest sex and social media preference are dependent (at the 5% level of significance)

Kolmogorov-Smirnov (K-S) Test

This test is used in situations where a comparison has to be made between an observed sample distribution and theoretical distribution.

In this lecture, we'll learn how to conduct a test to see how well a hypothesized distribution function $F(x)$ fits an empirical distribution function $F_n(X)$.

What is Empirical distribution function?

Given an observed random sample $X_1, X_2, \ldots, X_n$, an **empirical distribution** function $F_n(x)$ is the fraction of sample observations less than or equal to the value x.

ascending

More specifically, if $y_1 < y_2 < \cdots < y_n$ are the order statistics of the observed random sample, with no two observations being equal, then the empirical distribution function is defined as:

$$F_n(x) = \begin{cases} 0 & ; \ x < y_1 \\ \frac{k}{n} & ; \ y_k \le x \le y_{k+1}, k = 1,2,\ldots,n-1 \\ 1 & ; \ x \ge y_n \end{cases}$$

Example: A random sample of 8 people yields the following counts the number of times they read in the past months:

2 1 0 2 6 6 4 7

Calculate the empirical distribution function $F_n(x)$.

Solution: The empirical distribution function is

	x	0	1	2	4	6	7
pdf	$f(x)$	1/8	1/8	2/8	1/8	2/8	1/8
cdf	$F_n(x)$	1/8	2/8	4/8	5/8	7/8	8/8

$\longrightarrow F_n(x)$

Kolmogorov-Smirnov's approximation to Null Distribution

Kolmogorov-Smirnov test statistics D_n is defined as

$$D_n = \sup_x |F(x) - F_0(x)|$$

Hypothesis
(3 Steps Rule)

Step 1:

Define the Hypothesis

Null Hypothesis -

$$H_0: F_0(x) = F_n(x)$$

Alternative Hypothesis -

$$H_1: F_0(x) \neq F_n(x)$$

assumes no difference between the observed (empirical) $F_0(x)$ and theoretical $F_n(x)$ distribution,

value of test statistic

$$D = \max|F_0(x) - F_n(x)|$$

Step 2:

Compute test statistics

Step 3:

Conclusion:

Step 3: Conclusion:

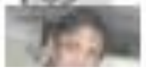
The critical value of D is found from the the K-S table values for one sample test.

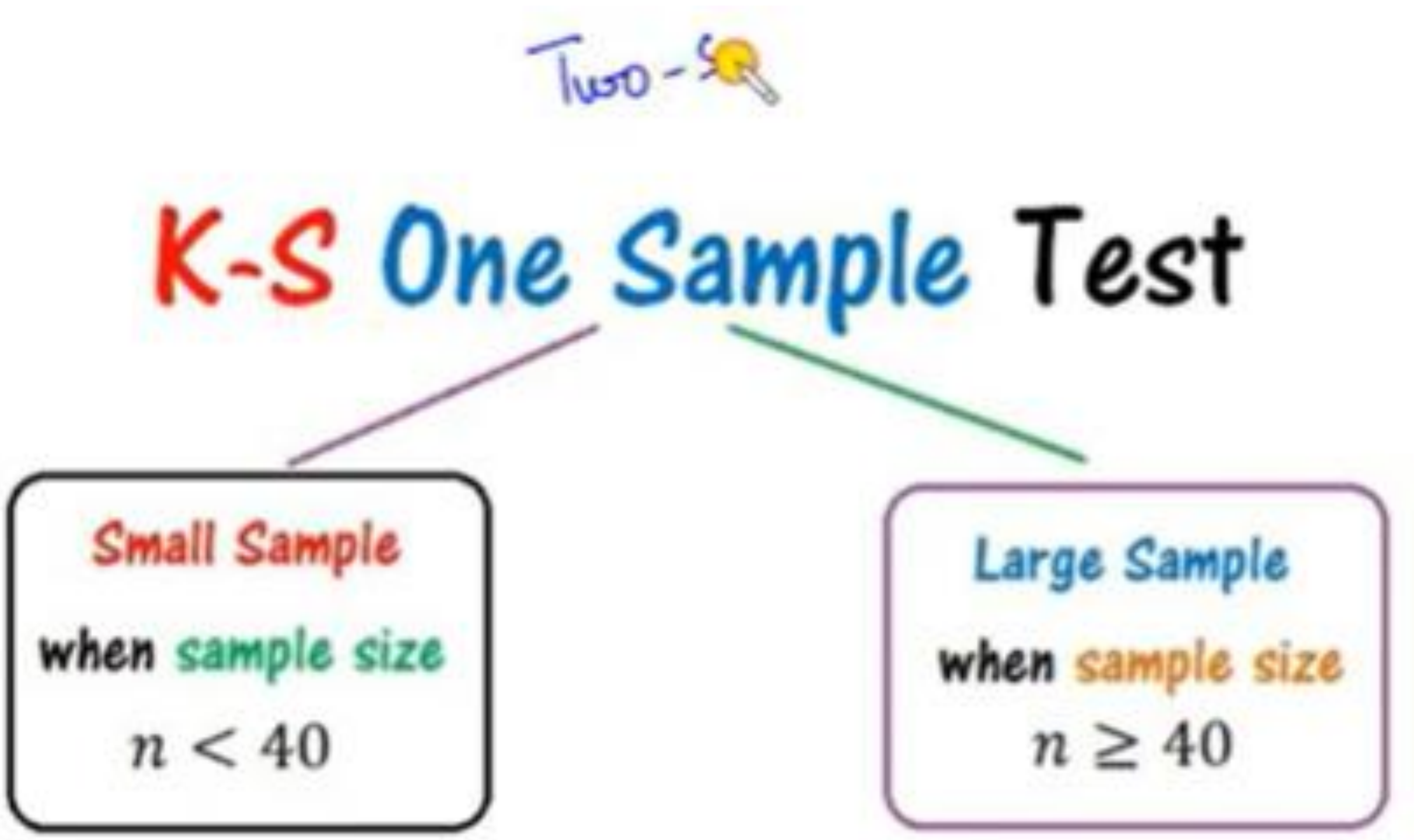
Reject H_0
if $D >$ Critical value

Accept H_0:
if $D <$ critical value

Critical Values of One-Sample K-S Test Statistics D

	Alpha						Alpha				
n	0.20	0.10	0.05	0.02	0.01	n	0.20	0.10	0.05	0.02	0.01
1	0.900	0.950	0.975	0.990	0.995	21	0.226	0.259	0.287	0.321	0.344
2	0.684	0.776	0.842	0.900	0.929	22	0.221	0.253	0.281	0.314	0.337
3	0.565	0.636	0.708	0.785	0.829	23	0.216	0.247	0.275	0.307	0.330
4	0.493	0.656	0.624	0.689	0.734	24	0.212	0.242	0.269	0.301	0.323
5	0.447	0.509	0.563	0.627	0.669	25	0.208	0.238	0.264	0.295	0.317
6	0.410	0.468	0.519	0.577	0.617	26	0.204	0.233	0.259	0.290	0.311
7	0.381	0.436	0.483	0.538	0.576	27	0.200	0.229	0.254	0.284	0.305
8	0.358	0.410	0.454	0.507	0.542	28	0.197	0.225	0.250	0.279	0.300
9	0.339	0.387	0.430	0.480	0.513	29	0.193	0.221	0.246	0.275	0.295
10	0.323	0.369	0.409	0.457	0.489	30	0.190	0.218	0.242	0.270	0.290
11	0.308	0.352	0.391	0.437	0.468	31	0.187	0.214	0.238	0.266	0.285
12	0.296	0.338	0.375	0.419	0.449	32	0.184	0.211	0.234	0.262	0.281
13	0.285	0.325	0.361	0.404	0.432	33	0.182	0.208	0.231	0.258	0.277
14	0.275	0.314	0.349	0.390	0.418	34	0.179	0.205	0.227	0.254	0.273
15	0.266	0.304	0.338	0.377	0.404	35	0.177	0.202	0.224	0.251	0.269
16	0.258	0.295	0.327	0.366	0.392	36	0.174	0.199	0.221	0.247	0.265
17	0.250	0.286	0.318	0.355	0.381	37	0.172	0.196	0.218	0.244	0.262
18	0.244	0.279	0.309	0.346	0.371	38	0.170	0.194	0.215	0.241	0.258
19	0.237	0.271	0.301	0.337	0.361	39	0.168	0.191	0.213	0.238	0.255
20	0.232	0.265	0.294	0.329	0.352	40	0.165	0.189	0.210	0.235	0.252
						$n > 40$ approx.	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	1.63




Two - S
K-S One Sample Test
Small Sample
when sample size
$n < 40$
Large Sample
when sample size
$n \geq 40$

Example: In a study done from various streams of a college 60 students, with equal number of students drawn from each stream, the intention of the students to join the Adventure Club of college noted after the interviewed are listed below.

	B.Sc.	B.A.	B.Com	M.A.	M.Com
No. in each class	5	9	11	16	19

It was expected that 12 students from each class would join the Adventure Club. Using the K-S test to find if there is any difference among student classes with regard to their intention of joining the Club.

Example: In a study done from various streams of a college 60 students, with equal number of students drawn from each stream, the intention of the students to join the Adventure Club of college noted after the interviewed are listed below.

	B.Sc.	B.A.	B.Com	M.A.	M.Com
No. in each class	5	9	11	16	19

It was expected that 12 students from each class would join the Adventure Club. Using the K-S test to find if there is any difference among student classes with regard to their intention of joining the Club.

Solution:

Step I: Define the Hypothesis:

H_0: There is no difference among students of different streams with respect to their intention of joining the club.

Step 2:

Compute observed (F_0) and theoretical (F_n) distributions.

Test statistic D is

$$D = \max|F_0(X) - F_n(X)|$$

$$= \frac{11}{60}$$

$$= 0.183$$

E T

Streams	No. of students interested in joining		$F_0(X)$ (cdf)	$F_n(X)$ (cdf)	$D = \lvert F_0(X) - F_n(X)\rvert$
	Observed	Theoretical			
B.Sc.	5	12	5/60	12/60	7/60
B.A.	9	12	14/60	24/60	10/60
B.COM.	11	12	25/60	36/60	11/60
M.A.	16	12	41/60	48/60	7/60
M.COM.	19	12	60/60	60/60	0/60
Total	n=60				

Step 3: Conclusion:

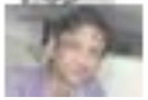
The critical value of D is found from the the K-S table values for one sample test.

Reject H_0 if $D >$ Critical value

Accept H_0: if $D <$ critical value

Critical Values of One-Sample K-S Test Statistics D

	Alpha						Alpha				
n	0.20	0.10	0.05	0.02	0.01	n	0.20	0.10	0.05	0.02	0.01
1	0.900	0.950	0.975	0.990	0.995	21	0.226	0.259	0.287	0.321	0.344
2	0.684	0.776	0.842	0.900	0.929	22	0.221	0.253	0.281	0.314	0.337
3	0.565	0.636	0.708	0.785	0.829	23	0.216	0.247	0.275	0.307	0.330
4	0.493	0.656	0.624	0.689	0.734	24	0.212	0.242	0.269	0.301	0.323
5	0.447	0.509	0.563	0.627	0.669	[illegible]	0.208	0.238	0.264	0.295	0.317
6	0.410	0.468	0.519	0.577	0.617	26	0.204	0.233	0.259	0.290	0.311
7	0.381	0.436	0.483	0.538	0.576	27	0.200	0.229	0.254	0.284	0.305
8	0.358	0.410	0.454	0.507	0.542	28	0.197	0.225	0.250	0.279	0.300
9	0.339	0.387	0.430	0.480	0.513	29	0.193	0.221	0.246	0.275	0.295
10	0.323	0.369	0.409	0.457	0.489	30	0.190	0.218	0.242	0.270	0.290
11	0.308	0.352	0.391	0.437	0.468	31	0.187	0.214	0.238	0.266	0.285
12	0.296	0.338	0.375	0.419	0.449	32	0.184	0.211	0.234	0.262	0.281
13	0.285	0.325	0.361	0.404	0.432	33	0.182	0.208	0.231	0.258	0.277
14	0.275	0.314	0.349	0.390	0.418	34	0.179	0.205	0.227	0.254	0.273
15	0.266	0.304	0.338	0.377	0.404	35	0.177	0.202	0.224	0.251	0.269
16	0.258	0.295	0.327	0.366	0.392	36	0.174	0.199	0.221	0.247	0.265
17	0.250	0.286	0.318	0.355	0.381	37	0.172	0.196	0.218	0.244	0.262
18	0.244	0.279	0.309	0.346	0.371	38	0.170	0.194	0.215	0.241	0.258
19	0.237	0.271	0.301	0.337	0.361	39	0.168	0.191	0.213	0.238	0.255
20	0.232	0.265	0.294	0.329	0.352	40	0.165	0.189	0.210	0.235	0.252
						$n > 40$ approx.	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	1.63



Step 2: Compute observed (F_0) and theoretical (F_n) distributions.

Test statistic D is

$$D = \max|F_0(X) - F_n(X)|$$
$$= \frac{11}{60}$$
$$= 0.183$$

Step 3: Since $n > 40$,

so large sample K-S test.

The critical value of D at 5% significance level is given by

$$D_{0.05} = \frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{60}} = 0.175$$

Since 0.183 > 0.175

hence we reject the null hypothesis H_0 and conclude that there is a difference among students of different streams in their intention of joining the Club.

Example: Consider the data points 1.41, 0.26, 1.97, 0.33, 0.55, 0.77, 1.46, 1.18

Is there any evidence to suggest that the data were not randomly sampled from a uniform (0, 2) distribution?

Example: Consider the data points 1.41, 0.26, 1.97, 0.33, 0.55, 0.77, 1.46, 1.18

Is there any evidence to suggest that the data were not randomly sampled from a uniform (0, 2) distribution?

Solution:

Step I: Define the Hypothesis:

$$H_0: F(x) = F_0(x)$$

$$H_1: F(x) \neq F_0(x)$$

Where $F(x)$ is the unknown cdf from which our data were sampled

and $F_0(x)$ is the cdf of the uniform distribution (0, 2).

The pdf of uniform distribution over (0, 2) is

$$f(x) = \frac{1}{2}\,; 0 < x < 2$$

Thus, it's cdf is

$$F_0(x) = \begin{cases} 0\,; & x < 0 \\ \frac{x}{2}\,; & 0 < x < 2 \\ 1\,; & x \geq 2 \end{cases}$$

And empirical cdf satisfies

$$F_n(x_k) = \frac{k}{8} \text{ for } k = 1,2,..,8$$

Step 2:

Compute observed (F_0) and theoretical (F_n) distributions.

Test statistic D is

$$D = \max|F_0(X) - F_n(X)|$$

k	Items	$F_n(x_k)$	$F_0(x_k)$	$D = \lvert F_0(x_k) - F_n(x_k)\rvert$
1	0.26	1/8		
2	0.33	2/8		
3	0.55	3/8		
4	0.77	4/8		
5	1.18			
6	1.41			
7	1.46			
8	1.97			

Step 2:

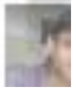
Compute observed (F_0) and theoretical (F_n) distributions.

Test statistic D is

$$D = \max|F_0(X) - F_n(X)|$$

k	Items	$F_n(x_k)$	$F_0(x_k)$	$D = \lvert F_0(x_k) - F_n(x_k)\rvert$
1	0.26	0.125	0.130	0.005
2	0.33	0.250	0.165	0.085
3	0.55	0.375	0.275	0.100
4	0.77	0.500	0.385	0.115
5	1.18	0.625	0.590	0.035
6	1.41	0.750	0.705	0.045
7	1.46	0.875	0.730	0.145
8	1.97	1	0.985	0.015

$$F_0(x) = \begin{cases} 0 \ ; & x < 0 \\ \frac{x}{2} \ ; & 0 < x < 2 \\ 1 \ ; & x \geq 2 \end{cases} \quad ; \quad F_n(x_k) = \frac{k}{8}$$



Step 2:

Compute observed (F_0) and theoretical (F_n) distributions.

Test statistic D is

$$D = \max|F_0(X) - F_n(X)|$$

$$= 0.145$$

Step 3: Since $n = 8 < 40$, so small sample K-S test.

The critical value of D at 5% significance level with $n = 8$ is given by

$$D_{0.05} = 0.46$$

As $0.145 < 0.46$

hence we fails to Reject the null hypothesis and conclude that the data were sampled from uniform(0,2) distribution.

K-S Two Samples Test

When instead of one, there are two independent samples then K-S two sample test can be used to test the agreement between two cumulative distributions.

K-S Two Samples Test

When instead of one, there are two independent samples then K-S two sample test can be used to test the agreement between two cumulative distributions.

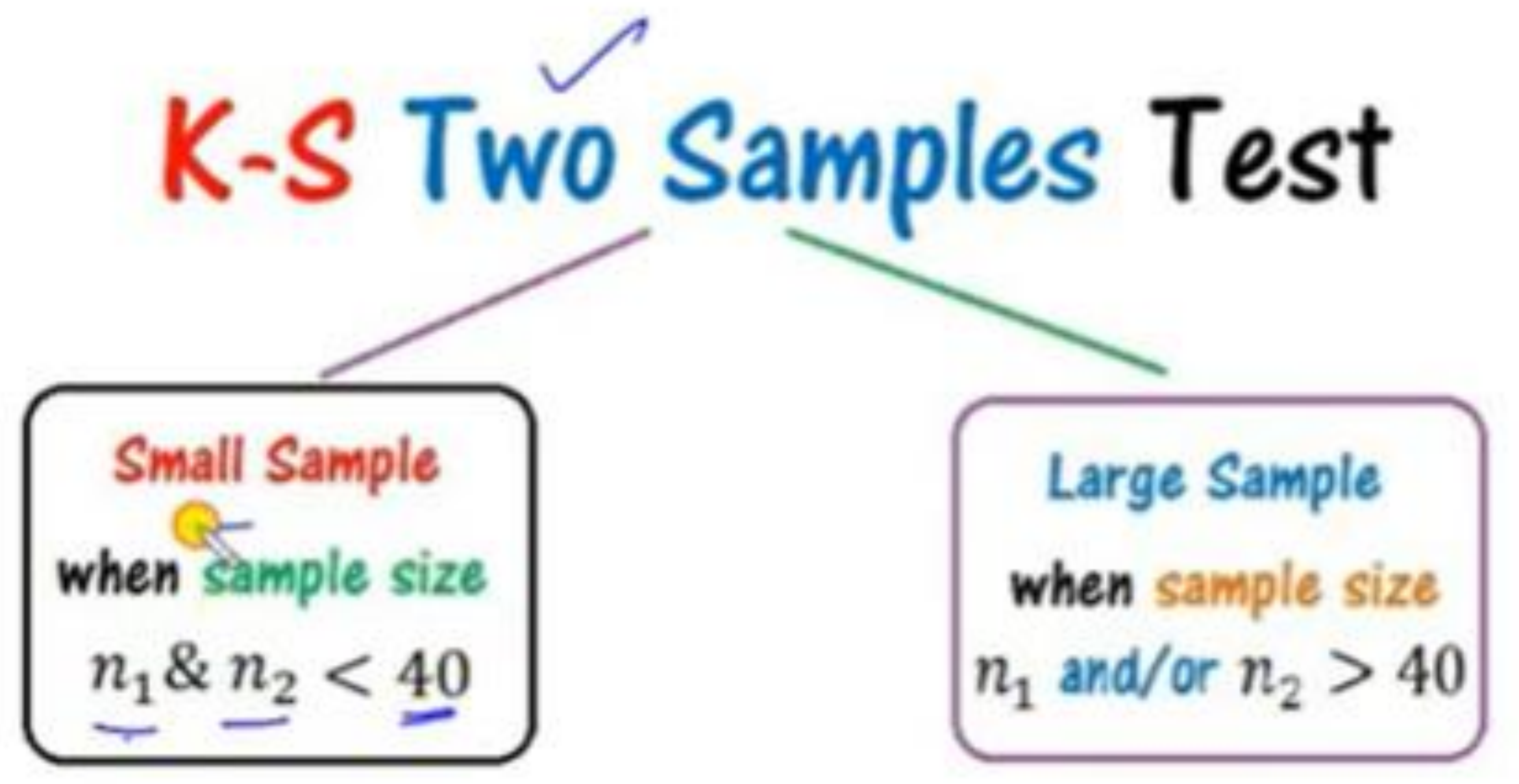
The null hypothesis H_0 states that there is no difference between the two distributions.

The D-statistic is calculated in the same manner as the K-S One Sample Test.

$$D = \max\left|F_{n_1}(X) - F_{n_2}(X)\right|$$

where n_1: observations from first sample

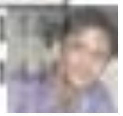
n_2: observations from second sample


K-S Two Samples Test
Small Sample
when sample size
n_1 & n_2 < 40
Large Sample
when sample size
n_1 and/or n_2 > 40

For Small Sample sizes critical value obtained from the table

Table gives critical D-values for $\alpha = 0.05$ (upper value) and $\alpha = 0.01$ (lower value) for various sample sizes. * means you cannot reject H_0 regardless of observed D.

$n_2 \backslash n_1$	3	4	5	6	7	8	9	10	11	12
1	* *	* *	* *	* *	* *	* *	* *	* *	* *	* *
2	* *	* *	* *	* *	* *	16/16 *	18/18 *	20/20 *	22/22 *	24/24 *
3	* *	* *	15/15 *	18/18 *	21/21 *	21/24 24/24	24/27 27/27	27/30 30/30	30/33 33/33	30/36 36/36
4		16/16 *	20/20 *	20/24 24/24	24/28 28/28	28/32 32/32	28/36 32/36	30/40 36/40	33/44 40/44	36/48 44/48
5			* *	24/30 30/30	30/35 35/35	30/40 35/40	35/45 40/45	40/50 45/50	39/55 45/55	43/60 50/60
6				30/36 36/36	30/42 36/42	34/48 40/48	39/54 45/54	40/60 48/60	43/66 54/66	48/72 60/72
7					42/49 42/49	40/56 48/56	42/63 49/63	46/70 53/70	48/77 59/77	53/84 60/84
8						48/64 56/64	46/72 55/72	48/80 60/80	53/88 64/88	60/96 68/96
9							54/81 63/81	53/90 70/90	59/99 70/99	63/108 75/108
10								70/100 80/100	60/110 77/110	66/120 80/120
11									77/121 88/121	72/132 86/132
12										96/1 84/1



For Large Sample sizes critical value is

$$c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

where

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

Example: Consider a survey on two different universities of the Postgraduate students on the topic of their wiliness to join the Research Funding Project on Artificial Intelligence. The following results obtained

University 1	3	2	3	5	8	9	8	8
University 2	2	8	2	4	4	3	6	

Determine whether the samples for University 1 and University 2 come from the same distribution?

Example: Consider a survey on two different universities of the Postgraduate students on the topic of their wiliness to join the Research Funding Project on Artificial Intelligence. The following results obtained

University 1	3	2	3	5	8	9	8	8
University 2	2	8	2	4	4	3	6	

Determine whether the samples for University 1 and University 2 come from the same distribution?

Solution:

Step 1: Define the Hypothesis:

$H_0: F_1(x) = F_2(x)$, i.e., they came from the same distribution

$H_1: F_1(x) \neq F_2(x)$ i.e., they came from the different distribution

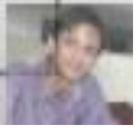
Step 2: Compute F_1 and F_2

Test statistic D is

$$D = \max|F_1(x) - F_2(x)|$$

count

Item	Univ 1 student	Univ 2 student	CDF $F_1(x)$	CDF $F_2(x)$	$D = \lvert F_1 - F_2 \rvert$
2	1	2	1/8	2/7	
3	2	1	3/8	3/7	
4	0	2	3/8	5/7	
5	1	0	4/8	5/7	
6	0	1	4/8	6/7	
8	3	1	7/8	1	
9	1	0	1	1	
Total	$n_1 = 8$	$n_2 = 7$			



Step 2: Compute F_1 and F_2

Test statistic D is

$$D = \max|F_1(x) - F_2(x)|$$

$$= 0.357143$$

Item	Univ 1 student	Univ 2 student	CDF $F_1(x)$	CDF $F_2(x)$	$D = \lvert F_1 - F_2 \rvert$
2	1	2	1/8	2/7	0.160714
3	2	1	3/8	3/7	0.053571
4	0	2	3/8	5/7	0.339286
5	1	0	4/8	5/7	0.214286
6	0	1	4/8	6/7	0.357143
8	3	1	7/8	1	0.125
9	1	0	1	1	0
Total	$n_1 = 8$	$n_2 = 7$			

Step 2: Compute F_1 and F_2

Test statistic D is

$$D = \max|F_1(x) - F_2(x)|$$
$$= 0.357143$$

Step 3: As $n_1, n_2 < 40$, so small two samples test

The critical value of D at 5% significance level with

$$n_1 = 8,\ n_2 = 7$$

is given by $\frac{40}{56} = 0.714285$

Since $0.357143 < 0.714285$

this implies that the null hypothesis is not rejected.

Hence, we MAY conclude that there is no significant difference between the distribution for the two samples.

Sign Test

It is the simplest of the entire non-parametric test.

As the name suggests, it is based on the signs (plus or minus) of the deviations rather than the exact magnitude of the variable values.

34 39 . .

Single Sample Sign Test

It is used to test the hypothesis concerning the median for one population.

Suppose we want to test the hypothesis that $\underbrace{\text{median } (\eta)}$ of a population has a specified value, say η_0, i.e.,

$$H_0: \eta = \eta_0 \xrightarrow{} \text{specified value}$$

Vs $H_1: \eta \neq \eta_0$ (Two-tailed) $H_1: \eta > \eta_0$ (Right-tailed)

$H_1: \eta < \eta_0$ (Left-tailed)

A

Procedure:

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the given population with median $\eta = \eta_0$ (under H_0).

Subtract, η_0 from each $X_i's$ and write

1) Plus sign (+) if the deviation is positive.
2) Negative sign (-) if the deviation is Negative,
3) Zero (0) if the deviation is zero.

Alternatively, Replace each observation

- exceeding η_0 with plus (+) sign,
- less than η_0 with minus (-) sign, and
- Equal to η_0 with zero (0).

By the definition of Median, we have

$$P(X > Median) = P(X < Median) = \frac{1}{2}$$

Thus, under H_0 $(\eta = \eta_0)$:, we have

$$P(X > \eta_0) = P(X < \eta_0) = \frac{1}{2}$$

Hence, if H_0 is true, then the number of + signs should be approximately equal to the - signs.

If the difference in the number of plus (+) and minus (-) signs is due to chance variations (or fluctuations of sampling), then we fails to reject the H_0.

Notation:

After Discarding Zeros

T^{+} = Number of Positive Sign

T^{-} = Number of Negative Sign

$T \ = \min(T^{+}, T^{-})$

Example: For the null hypothesis, Median $(\eta) = 5$, compute the values of T^+, T^-, T for the following observations:

8 9 3 5 4 11

Example: For the null hypothesis, Median $(\eta) = 5$, compute the values of T^+, T^-, T for the following observations:

8 9 3 5 4 11

Solution: Null hypothesis: Median $(\eta) = 5$

Subtract 5 from each observations and writing the signs as

+ + − 0 − +

Discard Zero and get

T^+ = Number of positive signs

$= 3$

T^- = Number of negative signs

$= 2$

Thus,

$T = \min(T^+, T^-)$

$= \min(3,2)$

$= 2$

Single Sample Sign test

(Small Samples, $n \leq 25$)

Procedure:

The following steps are summarized:

Step 1:

Set the Hypothesis:

H_0, H_1

Step 2:

compute T^+, T^-

$T = \min(T$

Step 3:

Test statistics T

Step 4:

Critical region:

Procedure:

The following steps are summarized:

Step 1: Set the Hypothesis:

Null Hypothesis

$$H_0: \eta = \eta_0$$

Alternative hypothesis

$H_1: \eta < \eta_0$ (Left-tailed)

$H_1: \eta > \eta_0$ (Right-tailed)

$H_1: \eta \neq \eta_0$ (Two-tailed)

Step 2: compute T^+, T^-

Subtract η_0 from each observations

Discard zeros and hence

compute T^+, T^- values.

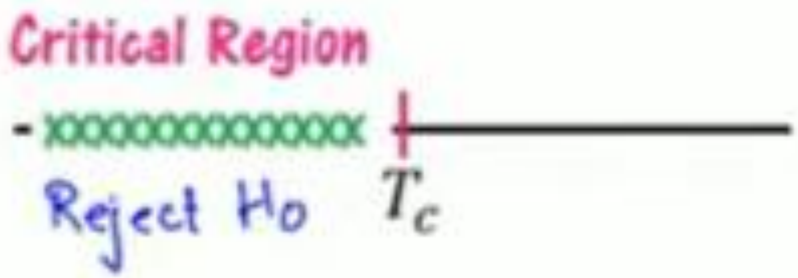
Step 3: Test statistics:

$$T = \min(T^+, T^-)$$

Step 4: Critical region:

Define the critical region as $T \le T_c$.

Where T_c is the critical value of T at given level of significance for one-tailed or two-tailed.



If calculated value (obtained from Step 3)

$$T \le T_c$$

Then REJECT H_0, otherwise,

H_0 may be regarded as TRUE.

Example: Following are the responses to the question "How many hours do you study before a major Statistics test?"

6 5 1 2 2 5 7 5 3 7 4 7

Use the sign test to test the hypothesis at the 5% level of significance that the median number of hours a student studies before a test is 3. Given that the critical value of sign test for $n = 11$ at 5% level of significance for two-tailed test is 1.

Example: Following are the responses to the question "How many hours do you stuay before a major Statistics test?"

6 5 1 2 2 5 7 5 3 7 4 7

Use the sign test to test the hypothesis at the 5% level of significance that the median number of hours a student studies before a test is 3. Given that the critical value of sign test for $n = 11$ at 5% level of significance for two-tailed test is 1.

Solution: Since the sample size is small $(n \leq 25)$.

Step 1: Null Hypothesis:

$$H_0 : \eta = 3$$

Alternative Hypothesis:

$$H_1 : \eta \neq 3 \text{ (Two-tailed)}$$

Example: Following are the responses to the question "How many hours do you study before a major Statistics test?"

6 5 1 2 2 5 7 5 3 7 4 7

Use the sign test to test the hypothesis at the 5% level of significance that the median number of hours a student studies before a test is 3. Given that the critical value of sign test for $n = 11$ at 5% level of significance for two-tailed test is 1.

Solution: Since the sample size is small $(n \leq 25)$.

Step 1: Null Hypothesis:

$$H_0 : \eta = 3$$

Alternative Hypothesis:

$$H_1 : \eta \neq 3 \text{ (Two-tailed)}$$

Step 2: Subtract 3 from each observation and writing the signs as

$$+ + - - - + + + 0 + + +$$

Discard Zero and get

T^{+} = Number of positive signs

$= 8$

T^{-} = Number of negative signs

$= 3$

Step 3: Test statistics

$$T = \min(T^{+}, T^{-})$$

$$= 3$$

Discard Zero and get

T^{+} = Number of positive signs

$= 8$

T^{-} = Number of negative signs

$= 3$

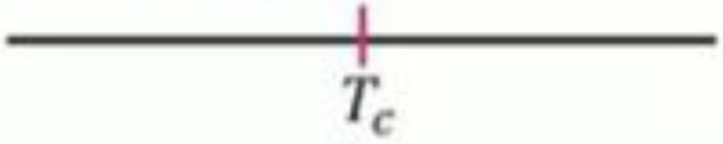
Step 3: Test statistics

$$T = \min(T^{+}, T^{-})$$

$$= 3$$

Step 4: Critical Region

Critical Region



$$n = T^{+} + T^{-}$$

$$= 11$$

Discard Zero and get

T^+ = Number of positive signs

$= 8$

T^- = Number of negative signs

$= 3$

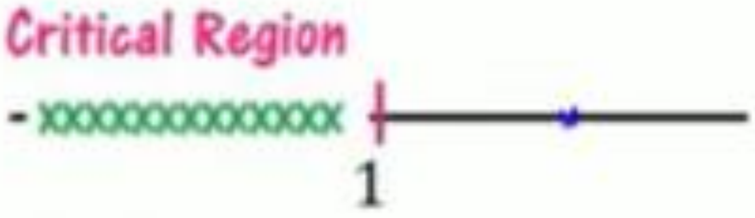
Step 3: Test statistics

$$T = \min(T^+, T^-)$$

$$= 3$$

Step 4: Critical Region

the critical value of sign test for $n = 11$ at 5% level of significance for two-tailed test is 1.



Thus, the critical region is $T \leq 1$

Since calculated value of $T = 3 > 1$,

so $\boxed{\text{we fail to REJECT } H_0.}$

Discard Zero and get

T^+ = Number of positive signs

$= 8$

T^- = Number of negative signs

$= 3$

Step 3: Test statistics

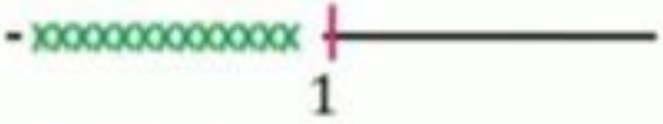
$$T = \min(T^+, T^-)$$

$$= 3$$

Step 4: Critical Region

the critical value of sign test for $n = 11$ at 5% level of significance for two-tailed test is 1.

Critical Region



Thus, the critical region is $T \leq 1$

Since calculated value of $T = 3 > 1$,

so $\boxed{\text{we fail to REJECT } H_0.}$

Therefore, we conclude that the median number of an hour of study before a test is 3 hours.

Example: A teacher claims that the median time to do a particular type of Statistics problems is at most 3 minutes, but her students believe that the median time is more than 3 minutes. A random sample of 10 students completed the problem in the following times (in minutes)

2.5	2	4	4.5	4	2.5	4.5	3	3.5	5

Use the sign test with 5% level of significance to test the teacher's claim. Given that the critical value of sign test for $n = 9$ at 5% level of significance for one-tailed test is 1.

Solution:

Example: A teacher claims that the median time to do a particular type of Statistics problems is at most 3 minutes, but her students believe that the median time is more than 3 minutes. A random sample of 10 students completed the problem in the following times (in minutes)

2.5 2 4 4.5 4 2.5 4.5 3 3.5 5

Use the sign test with 5% level of significance to test the teacher's claim. Given that the critical value of sign test for $n = 9$ at 5% level of significance for one-tailed test is 1.

Solution: The sample size is small $(n \leq 25)$.

Step 1: Null Hypothesis:

$$H_0: \eta \leq 3$$

Alternative Hypothesis:

$$H_1: \eta > 3 \text{ (Right-tailed)}$$

Step 2: Subtract 3 from each observation and writing the signs as

$$- \; - \; + \; + \; + \; - \; + \; 0 \; + \; +$$

$T^+ = 6$ T^-

Discard Zero and get

$T^+ =$ Number of positive signs

$= 6$

$T^- =$ Number of negative signs

$= 3$

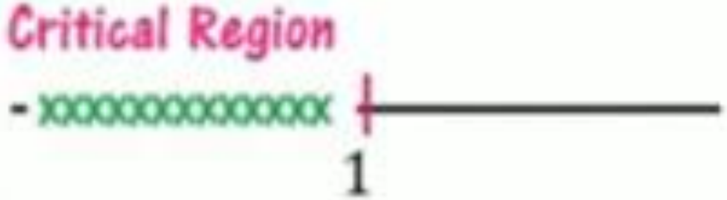
Step 3: Test statistics

$$T = \min(T^+, T^-)$$

$$= 3$$

Step 4: Critical Region

the critical value of sign test for $n = 9$ at 5% level of significance for two-tailed test is 1.



Thus, the critical region is $T \leq 1$

Since calculated value of $T = 3 > 1$,

so $\boxed{\text{we fail to REJECT } H_0}$.

Therefore, we conclude that the teacher claims that the Median time is at most 3 minutes MAY be regarded as TRUE.