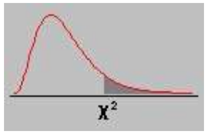
بسم الله الرحمن الرحيم

Chi Square (χ^2) test

PART 2

PROF. DR. WAQAR AL-KUBAISY

August 1 -2023



SPECIFIC LEARNING OUTCOMES

On completion of this lecture, you should be able to:

1.Explain the basis for the use of Chi square tests on qualitative data

2.Explain the **limitations of the Chi square** tests

3.Carry out the Chi square tests

4.Interpret the findings from the Chi square tests of significance

5.Interpret degrees **of freedom and critical** values of Chi square statistics from **Chi square table**

CONTENTS

1.Explanation of the basis for the use of Chi square tests on **qualitative data**

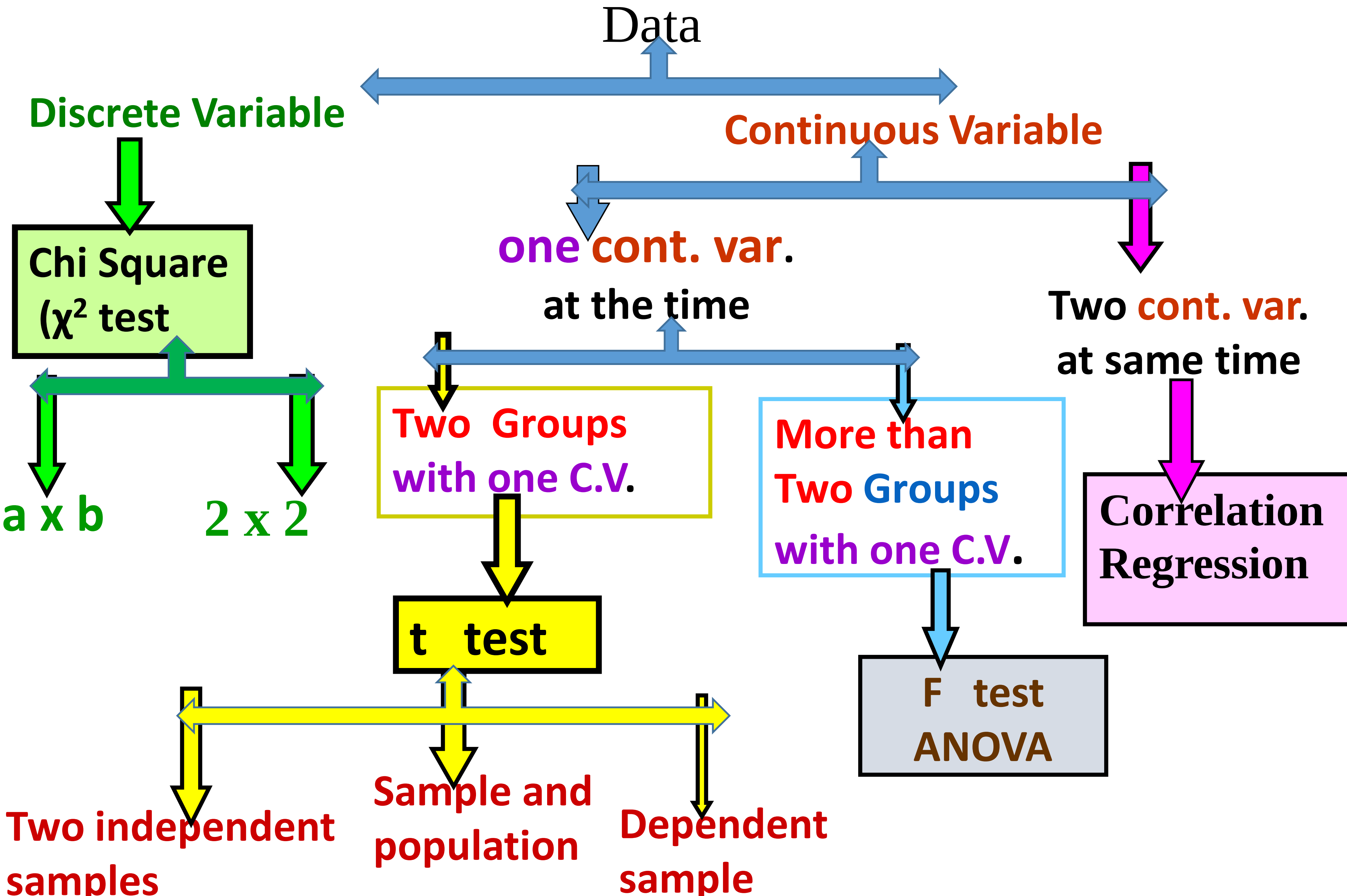
2.Explanation of the limitations of the Chi square tests

3.Calculation of Chi square

4.Chi square table

5.Interpretation of the findings from the Chi square tests of significance

An important thing is the type of the variable concerned.

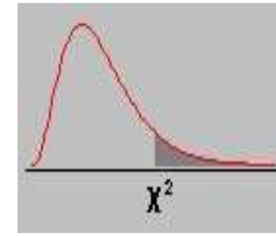


An important thing is the type of the variable concerned.

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

Application of χ^2.

1. **2×2 table .**
2. **$a \times b$ table .**



2 × 2 table

The application of χ^2 is to test the significance association between outcome and certain factor that we are interested in .

Here we have

two groups with
two outcome for each group

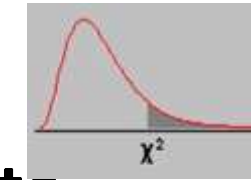
two groups
each group with two outcome for each group
.

In this case we use what we call it 2 × 2 table .

In this case we are going to compare between two proportion of two groups of population .

2 × 2 table

Example



A sample of 671 diseased person were subjected to treatment, 354 individuals of them, were given drug A. Of those given drug A only 240 patients were survived. On the other hand only 212 patients *who's given drug B were survived can we conclude that the effectiveness of treatment differ* between two drugs (A&B) ????.

Let α 0.05

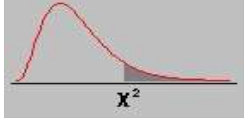
Out come	Drug A	Drug B	Total
Survived	240	212	?????
Died	??????	????	?????
Total	354	?????	671

(**also known as a cross tabulation or crosstab**)

Out come	Drug A	Drug B	Total
Survived	240	212	452
Died	114	105	219
Total	354	317	671

We would like to see if there is a **significance difference in** the **survival rate between the two drugs** . Let α 0.05

$$\text{Total Survival rate} = \frac{452}{671} \times 100 = 67.4\%$$



Survival rate for A $= \frac{240}{354} \times 100 = 67.8\%$

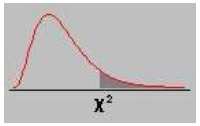
Survival rate for B $= \frac{212}{317} \times 100 = 66.9\%$

There is an **observed difference** in the survival rate between drug A (67.8%) and B (66.9%) .

Is this difference in survival rate due to :

- Drug Effectiveness .
- Chance Factor .

Out come	Drug A	Drug B	Total
Survived	240 (67.5%)	212(66.9%)	452(67.4%)
Died	114	105	219
Total	354	317	671



Data

Data consist of sample of patients divided into two groups, group A and group B .

Survival rate in group treated by drug **A was 67.8 %, and**

Survival rate in group treated by drug B was 66.8% .

Assumption

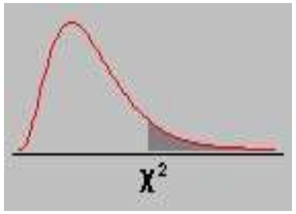
Two independent group of patients given two different type of treatment chosen randomly from normal distribution population .

Formulation of Hypothesis

Ho

HA

Formulation of Hypothesis



Ho

There is no significance difference in the proportion (rate) of survival between two groups .

survival rate group treated by drug A was 67.8% &

survival rate group treated by drug B was 66.9%

There is no significance association between survival rate and type of treatment .

P1 = P2 = P0 .

HA

There is a **significance difference** in the survival **rate** between two type of treatment .

P1¬ ≠ P2 ≠ P0 .

Survival rate is **higher among** group of patients treated by drug A .

Critical region

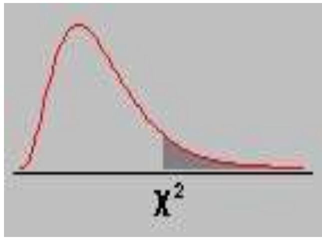
Level of significance 0.95, $\alpha = 0.05$

d.F =

(No. of rows – 1) (No. of column – 1)

= (r – 1) (c – 1)

(2 – 1) (2 – 1) = 1



Outcome	Drug A	Drug B	Total
Survived	240	212	452
Died	114	105	219
Total	354	317	671

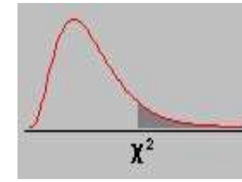
tabulated χ^2 of d.F =1 with α 0.05 = 3.841

Proper test

χ^2 , 2 × 2 table

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

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



$$E = \frac{total\ column \times total\ rows}{Grand\ \ total} \quad for\ each\ cell$$

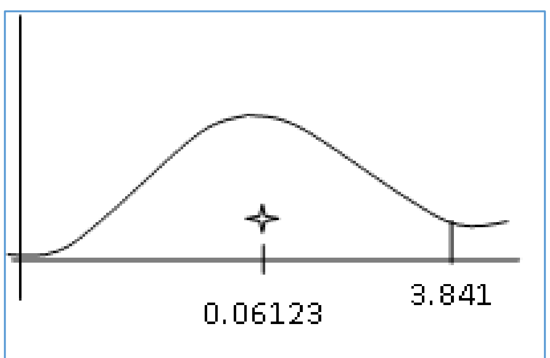
$$E_{240} = \frac{354 \times 452}{671} = 238.5$$

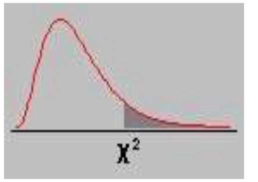
$$E_{114} = \frac{354 \times 219}{671} = 115.5$$

$$E_{212} = \frac{452 \times 317}{671} = 213.5$$

$$E_{105} = \frac{317 \times 219}{671} = 103.5$$

Outcome	Drug A		Drug B		Total
	O	E	O	E	
Survived	240	238.5	212	213.5	452
Died	114	115.5	105	103.5	219
Total	354		317		671





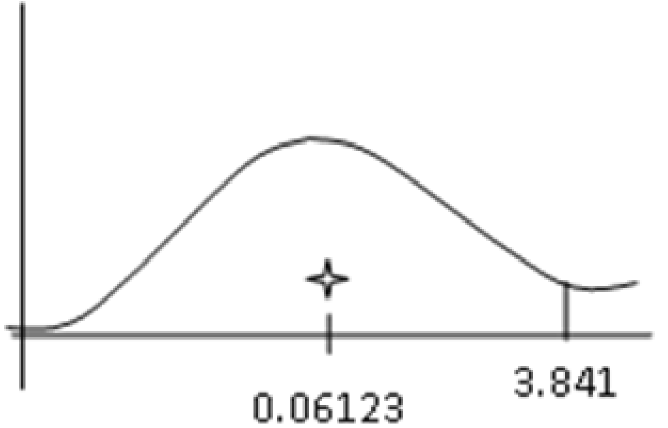
$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(240-238.5)^2}{238.5} + \frac{(114-115.5)^2}{115.5} + \frac{(212-213.5)^2}{213.5} + \frac{(105-103.5)^2}{103.5}$$

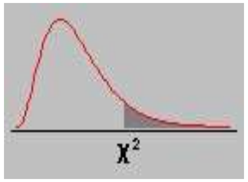
$$= \frac{(1.5)^2}{238.5} + \frac{(1.5)^2}{115.5} + \frac{(-1.5)^2}{213.5} + \frac{(1.5)^2}{103.5} = \frac{2.25}{238.5} + \frac{2.25}{115.5} + \frac{2.25}{213.5} + \frac{2.25}{103.5}$$

$$= 0.009434 + 0.0195 + 0.01056 + 0.02174$$

$$= 0.061234$$



Calculated χ^2 fall in Accept Region → so



We **not reject** (accept) **Ho** .

There is no significance difference in proportion of survival rate between two drugs

P > 0.05

Calculated χ^2 less than tabulated χ^2

chance factor increases,

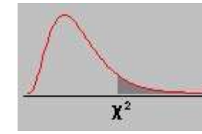
influencing factor decrease

There is no significance effect of drug A to increase survival rate .

P > 0.05

.

P > 0.05 .



Example

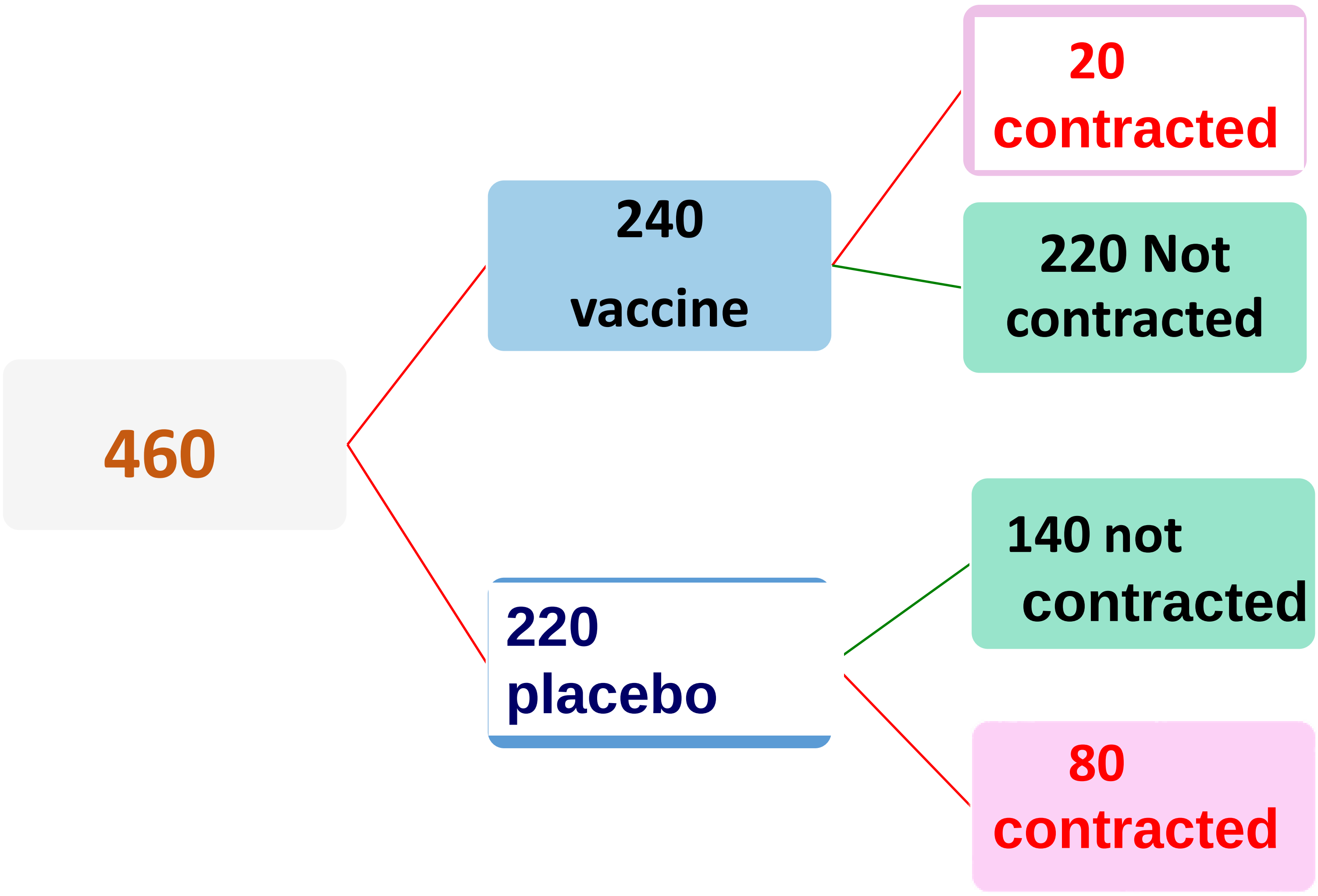
A sample of 460 adult was chosen , 240 were given influenza vaccine while the remaining given placebo Overall 100 persons contracted influenza, of whom 20 were in vaccine group .

we would like to assess the strength of evidence that vaccination affect the probability of contracting disease

is there any evidence that vaccine have an effect on contracting the disease ??

Total 460 → 100 persons contracted influenza

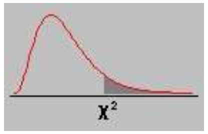
240 vaccinated → 20 contracted influenza



Total 460 → 100 persons contracted influenza

240 vaccinated → 20 contracted influenza

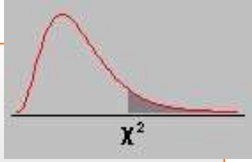
We start by display data in 2X2 table .



- **The exposure is vaccination (the row variable) and**
- **the outcome is contracting influenza (the column variable)**
- **we therefore include row % in the table**

Exposure	Out come +ve	Out come -ve	total
yes			
no			
Total			

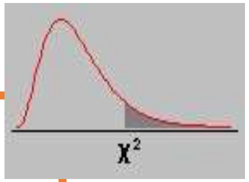
(also known as a cross tabulation or crosstab)



We start by display data in 2X2 table .
The exposure is vaccination (the row variable) and the outcome is contracting influenza (the column variable) we therefore include row % in the table

Given	Contract influenza N %	Not contract influenza N %	Total
Vaccine	**20**	220	**240**
placebo	80	140	220
Total	**100**	360	**460**

Total **460** → **100 persons contracted influenza**
240 vaccinated → **20 contracted influenza**



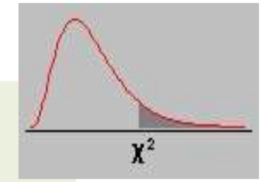
	Contract influenza N (%)	Not contract influenza N (%)	Total
Vaccine	**20 (8.3)**	**220 (91.7)**	**240**
placebo	**80 (36)**	**140 (63.6)**	**220**
Total	**100 (21.7)**	**360 (78.3)**	**460**

Overall persons contracting influenza
100/460= 21.7%

The chi square compare the observed number in each of four categories with the number expected

$$E = \frac{\text{Total row X total column}}{\text{Over all total frequency}}$$

E expected (E) $= \dfrac{\text{total column X total row}}{\text{Grand total}}$



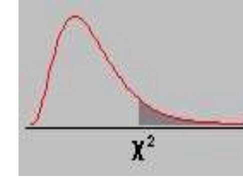
E20 = $\dfrac{240X\ 100=}{460}$

E220 = $\dfrac{240X\ 360=}{460}$

E80 = $\dfrac{220X\ 100=}{460}$

E140 = $\dfrac{220X\ 360=}{460}$

	Contract influenza N (%)	Not contract influenza N (%)	Total
Vaccine	20 (8.3)	220 (91.7)	240
placebo	80 (36)	140 (63.6)	220
Total	100(21.7)	360 (78.3)	460



E expected (E) $= \dfrac{\text{total column X total row}}{\text{Grand total}}$

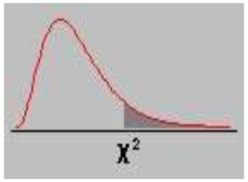
The chi square compare

the observed number in each of four categories

with the number expected

	Contract influenza		Not contract influenza		total
	O	**E**	**O**	**E**	
Vaccine	**20**	**52.2**	**220**	**187.8**	**240**
placebo	**80**	**47.8**	**140**	**172.2**	**220**
Total	**100**		**360**		**460**

Then chi square be calculated by calculating E. frequencies



if there were no difference in the efficacy between vaccine and placebo.

if the vaccine and placebo having same efficiency then we expect to have same proportion in each group that is in the

vaccine group 100/460 X 240= 52.2

in placebo group 100/460 X 220 = 47.8

would have contract influenza. .

H0= 52.2 =47.8

Similarly

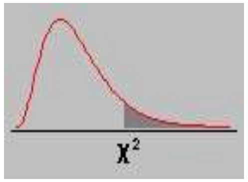
360/ 460 x240=187.8 in vaccine group

360/460 X 220 =172.2 in placebo group

will escape influenza

Then chi square be calculated by calculating E. frequencies

$$X^2 = \sum \frac{(O - E)^2}{E} \qquad d.f. = 1$$



	Contracting Influenza O	Contracting Influenza E	Not contract influenza O	Not contract influenza E	total
Vaccine	20	52.2	220	187.8	240
placebo	80	47.8	140	172.2	220
Total	100		360		460

$$X^2 = \frac{(20 - 52.2)^2}{52.2} + \frac{(80 - 47.8)^2}{47.8} + \frac{(220 - 187.8)^2}{187.8} + \frac{(140 - 172.2)^2}{172.2}$$

$$19.86 + 21.69 + 5.52 + 6.02 = 53.99$$

compare calcul. With tabulated

Critical region

d.F = (C – 1) (r – 1)

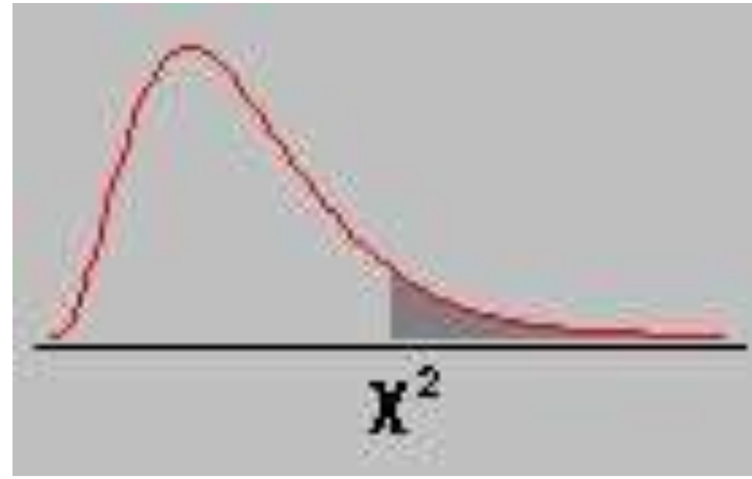
= (2 – 1) (2 – 1) = 1

$\alpha = 0.05$

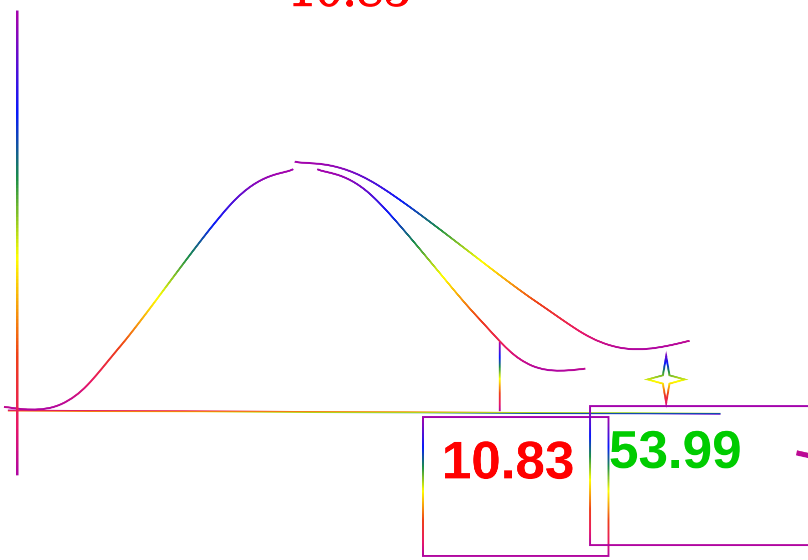
tabulated X^2 = 3.84

6.64

10.83



10.83



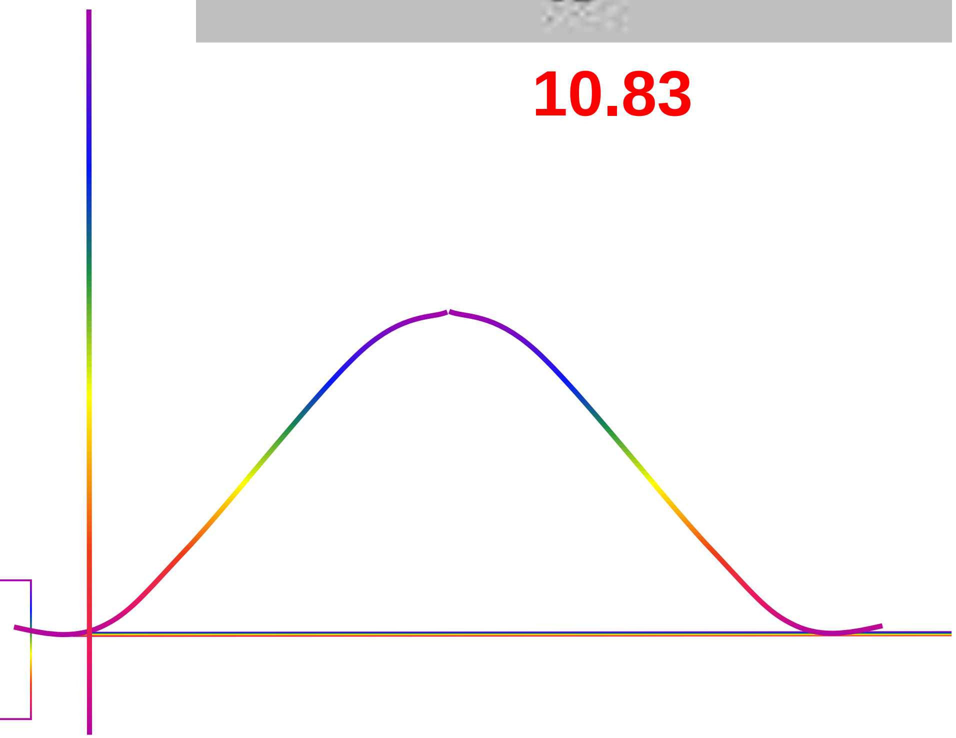
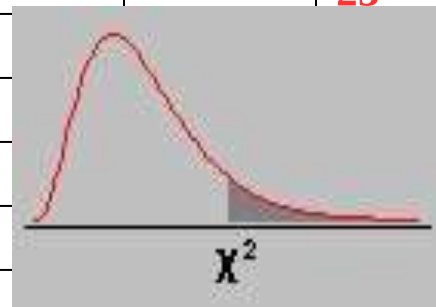


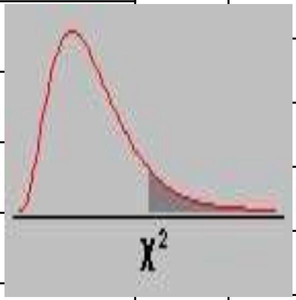
Table of Chi-square statistics

df	P =0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.69	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82



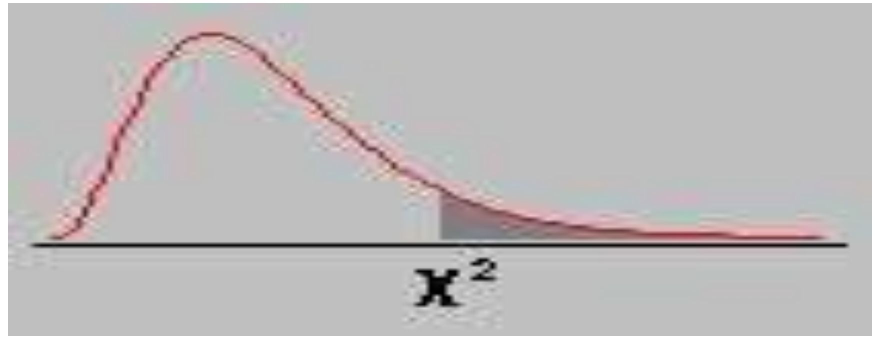
21	32.67	38.93	46.80
22	33.92	40.29	48.27
23	35.17	41.64	49.73
	36.42	42.98	51.18
	37.65	44.31	52.62
	38.89	45.64	54.05
	40.11	46.96	55.48
28	41.34	48.28	56.89
29	42.56	49.59	58.30
30	43.77	50.89	59.70
31	44.99	52.19	61.10
32	46.19	53.49	62.49
33	47.40	54.78	63.87
34	48.60	56.06	65.25
35	49.80	57.34	66.62
36	51.00	58.62	67.99
37	52.19	59.89	69.35
38	53.38	61.16	70.71
39	54.57	62.43	72.06
40	55.76	63.69	73.41

41	56.94	64.95	74.75
42	58.12	66.21	76.09
43	59.30	67.46	77.42
44	60.48	68.71	78.75
45	61.66	69.96	80.08
46	62.83	71.20	81.40
47	64.00	72.44	82.72
48	65.17	73.68	84.03
49	66.34	74.92	85.35
50	67.51	76.15	86.66
51	68.67	77.39	87.97
52	69.83	78.62	89.27
53	70.99	79.84	90.57
54	72.15	81.07	91.88
55	73.31	82.29	93.17
56	74.47	83.52	94.47
57	75.62	84.73	95.75
58	76.78	85.95	97.03
59	77.93	87.17	98.34
60	79.08	88.38	99.63



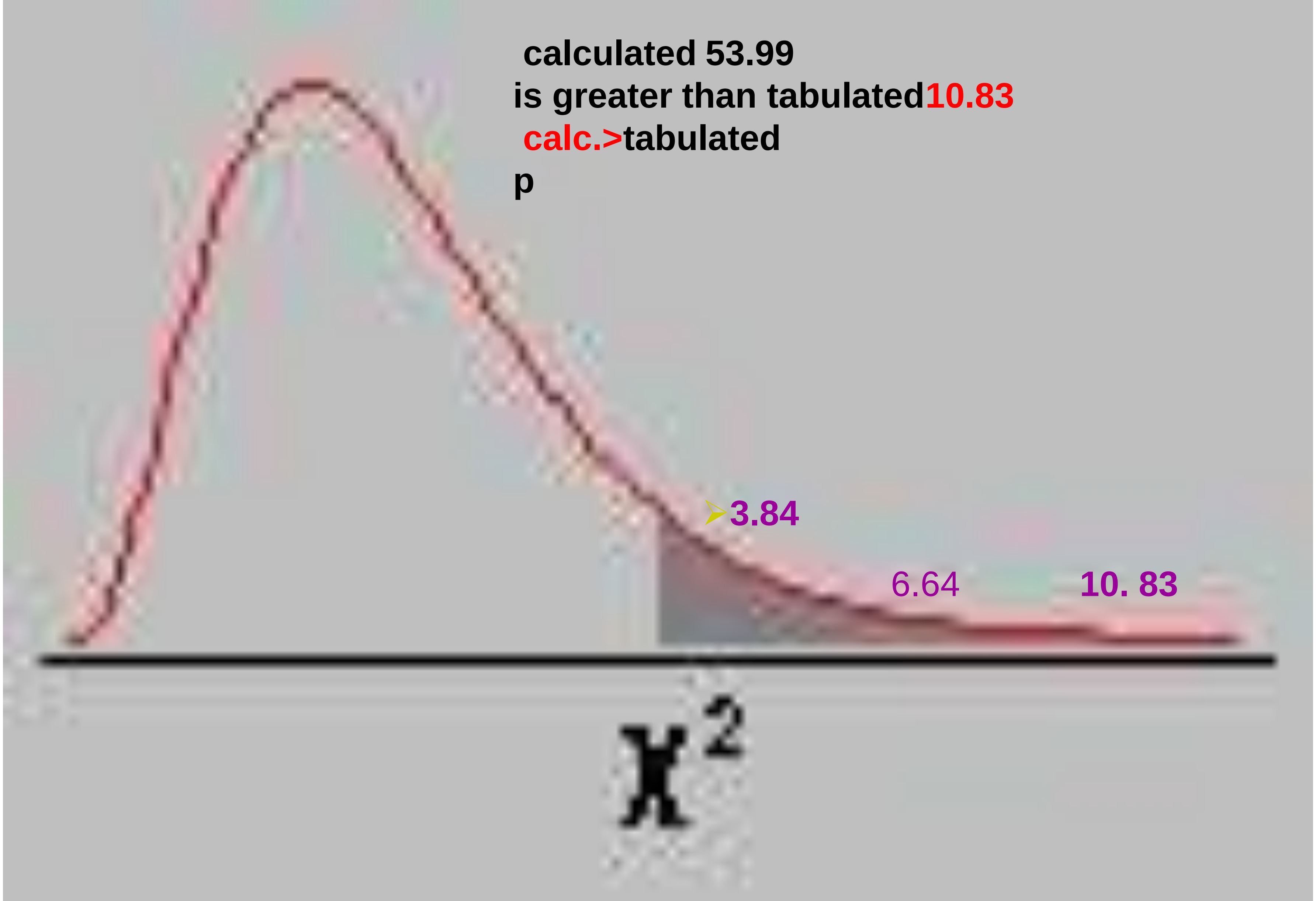
61	80.23	89.59	100.88
62	81.38	90.80	102.15
63	82.53	92.01	103.46
64	83.68	93.22	104.72
65	84.82	94.42	105.97
66	85.97	95.63	107.26
67	87.11	96.83	108.54
68	88.25	98.03	109.79
69	89.39	99.23	111.06
70	90.53	100.42	112.31
71	91.67	101.62	113.56
72	92.81	102.82	114.84
73	93.95	104.01	116.08
74	95.08	105.20	117.35
75	96.22	106.39	118.60
76	97.35	107.58	119.85
77	98.49	108.77	121.11
78	99.62	109.96	122.36
79	100.75	111.15	123.60
80	101.88	112.33	124.84

81	103.01	113.51	126.09
82	104.14	114.70	127.33
83	105.27	115.88	128.57
84	106.40	117.06	129.80
85	107.52	118.24	131.04
86	108.65	119.41	132.28
87	109.77	120.59	133.51
88	110.90	121.77	134.74
89	112.02	122.94	135.96
90	113.15	124.12	137.19
91	114.27	125.29	138.45
92	115.39	126.46	139.66
93	116.51	127.63	140.90



93	116.51	127.63	140.90
94	117.63	128.80	142.12
95	118.75	129.97	143.32
96	119.87	131.14	144.55
97	120.99	132.31	145.78
98	122.11	133.47	146.99
99	123.23	134.64	148.21
100	124.34	135.81	149.48

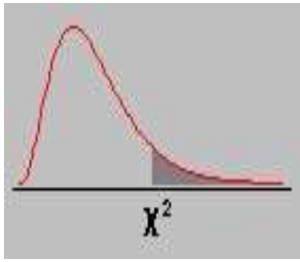
calculated 53.99
is greater than tabulated10.83
calc.>tabulated
p



53.99

p is ?????????

- This mean that
- the probability is less than 0.001
- that this difference is due to chance factor



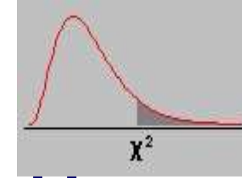
- and more than 99.999 that this difference
- due to vaccine

- Thus there is a <u>strong evidence against</u> null hypotheses that is saying no effect of vaccine on the probability of contracting influenza .

- there is a strong evidence that vaccine is effective

- Therefore it is concluded that vaccine is effective

Continuity Correction

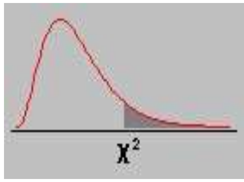


The chi square test for 2X2 table can be improved by using continuity correction we call it Yates continuity correction the formula become

$$X^2 = \sum \left[\frac{(O - E) - 0.5}{E} \right]^2 \quad d.f. = 1$$

Pearson's chi-squared test by subtracting 0.5 from the difference between each observed value and its expected value in a 2 × 2

Resulting in small value for chi square (the value of O –E) ignoring the sig



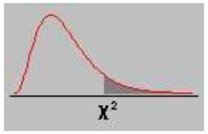
Chi square calculation procedure

- ✓ Calculate the expected values E for each cell
- ✓ Calculate the value O- E for each cell
- ✓ O is the observed
- ✓ Square O-E
- ✓ Divide each squared O- E by E for each cell
- ✓ Sum all of the values in previous step

this result is called test statistic

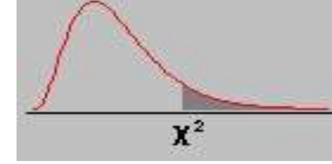
- ✓ identify the critical chi-square obtained
- ✓ from the chi square table.
- ✓ To reject the null hypothesis of equal proportion i.e. of independent variables the value of the test statistics must exceed the critical chi-square obtained from the chi square table.

Example



A sample of 84 mother chosen randomly
20 were smoker who delivered 14 babies with small birth weight (BW) and 6 normal BW.
On the other hand 64 non smoker women deliver 20 small BW babies and 44 normal BW babies
can we conclude that maternal smoking has a relation to small birth weight ?

mother	Small BW	Normal BW	total
Smoker	14	6	20
Non smoker	20	44	64
Total	**34**	**50**	84



Example

A sample of 84 mother chosen randomly 20 were smoker who delivered 14 babies with small birth weight (BW) and 6 normal BW. On the other hand 64 non smoker women deliver 20 small BW babies and 44 normal BW babies can we conclude that maternal smoking has a relation to small birth weight ?

	Small BW	Normal BW	total
Smoker	14 (70%)	6 (30%)	20
Non smoker	20 31.3 %)	44 68.7%	64
Total	34 (40.5%)	50	84

Ho ;

small BW and smoking status during pregnancy are not related in the population.

The Two variables are independent

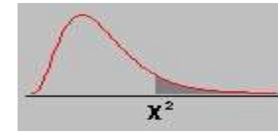
H1:

Small BW and smoking status during pregnancy are **related** in the population .

The Two variable are Dependent

$$Ho = P_1 = P_2 = P_0$$

$$H_A = P_1 \neq P_2 \neq P_0$$

If the two variables are unrelated(H0) 

then there is no reason why the proportion of small BW among smokers should be different to the proportion of small BW among non smokers mothers (H0)

In another ward these two proportions should be equal

P1 =P2   70% = 31.3%

this difference could be due to chance (H0)

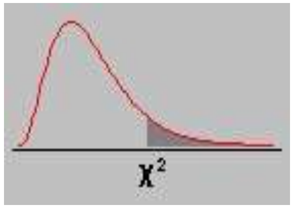
	Small BW		Normal BW		total
Smoker	14	**70%**	6	**30%**	20
Non smoker	20	**31.3 %**	44	**68.7%**	64
Total	34	**40.5%**	50		84

The question is that what proportion would we expect to find if null hypothesis of unrelated variable is true ??

The answer is that

since we got 34 small BW in a total of 84.

34/84= **0.405** **40.5%**



so we expect in smokers group to have ;0.405 x 20=8.1

in nonsmokers 0.405 x 64 = 25.92

An easer way to calculate Expected cell frequency

$$\frac{\text{Total row X total column}}{\text{Over all total frequency}}$$

$$\frac{34 \times 20}{84} = 8.094$$

$$\frac{34 \times 64}{84} = 25.904$$

	Small BW	Normal BW	total
Smoker	14	6	20
Non smoker	20	44	64
Total	34	50	84

$$\text{Expected freq.} = \frac{\text{Total row X total column}}{\text{Over all total frequency}}$$

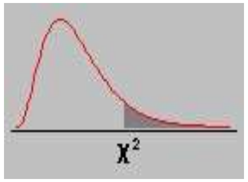
	Small BW O	Small BW E	Normal BW O	Normal BW E	total
Smoker	14	8.1	6	11.9	20
Non smoker	20	25.1	44	30.1	64
Total	34		50		84

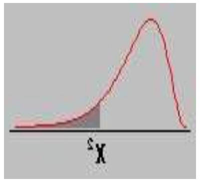
$$\frac{(14-8.1)^2}{8.1} + \frac{(6-11.9)^2}{11.9} + \frac{(20-25.1)^2}{25.1} + \frac{(44-30.1)^2}{30.1}$$

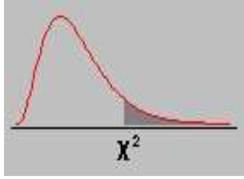
$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

χ^2 = 4,3+2.9+1+6.4 =**14.6**

compare calculated . **χ2 with** tabulated **χ2**







Critical region

d.F $= (C - 1)\,(r - 1)$

$= (2 - 1)\,(2 - 1) = 1$

$\alpha = 0.05$

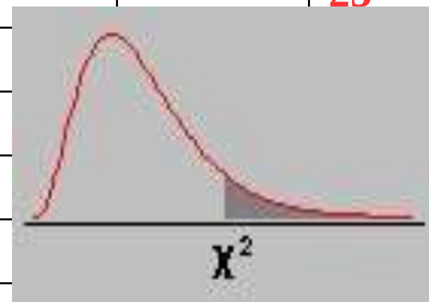
tabulated $\chi^2 = 3.84$

6.64

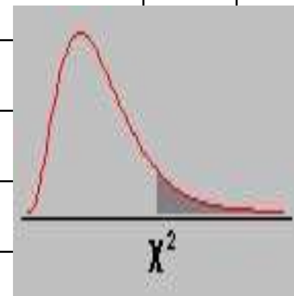
10.83

Table of Chi-square statistics

df	P =0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.69	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82
20	31.41	37.57	45.32
21	32.67	38.93	46.80
22	33.92	40.29	48.27
23	35.17	41.64	49.73
[illegible]	36.42	42.98	51.18
[illegible]	37.65	44.31	52.62
[illegible]	38.89	45.64	54.05
[illegible]	40.11	46.96	55.48
28	41.34	48.28	56.89
29	42.56	49.59	58.30
30	43.77	50.89	59.70
31	44.99	52.19	61.10
32	46.19	53.49	62.49
33	47.40	54.78	63.87
34	48.60	56.06	65.25
35	49.80	57.34	66.62
36	51.00	58.62	67.99
37	52.19	59.89	69.35
38	53.38	61.16	70.71
39	54.57	62.43	72.06
40	55.76	63.69	73.41

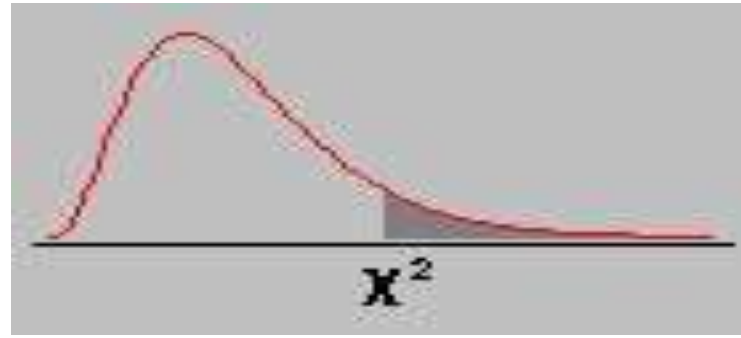


41	56.94	64.95	74.75
42	58.12	66.21	76.09
43	59.30	67.46	77.42
44	60.48	68.71	78.75
45	61.66	69.96	80.08
46	62.83	71.20	81.40
47	64.00	72.44	82.72
48	65.17	73.68	84.03
49	66.34	74.92	85.35
50	67.51	76.15	86.66
51	68.67	77.39	87.97
52	69.83	78.62	89.27
53	70.99	79.84	90.57
54	72.15	81.07	91.88
55	73.31	82.29	93.17
56	74.47	83.52	94.47
57	75.62	84.73	95.75
58	76.78	85.95	97.03
59	77.93	87.17	98.34
60	79.08	88.38	99.62

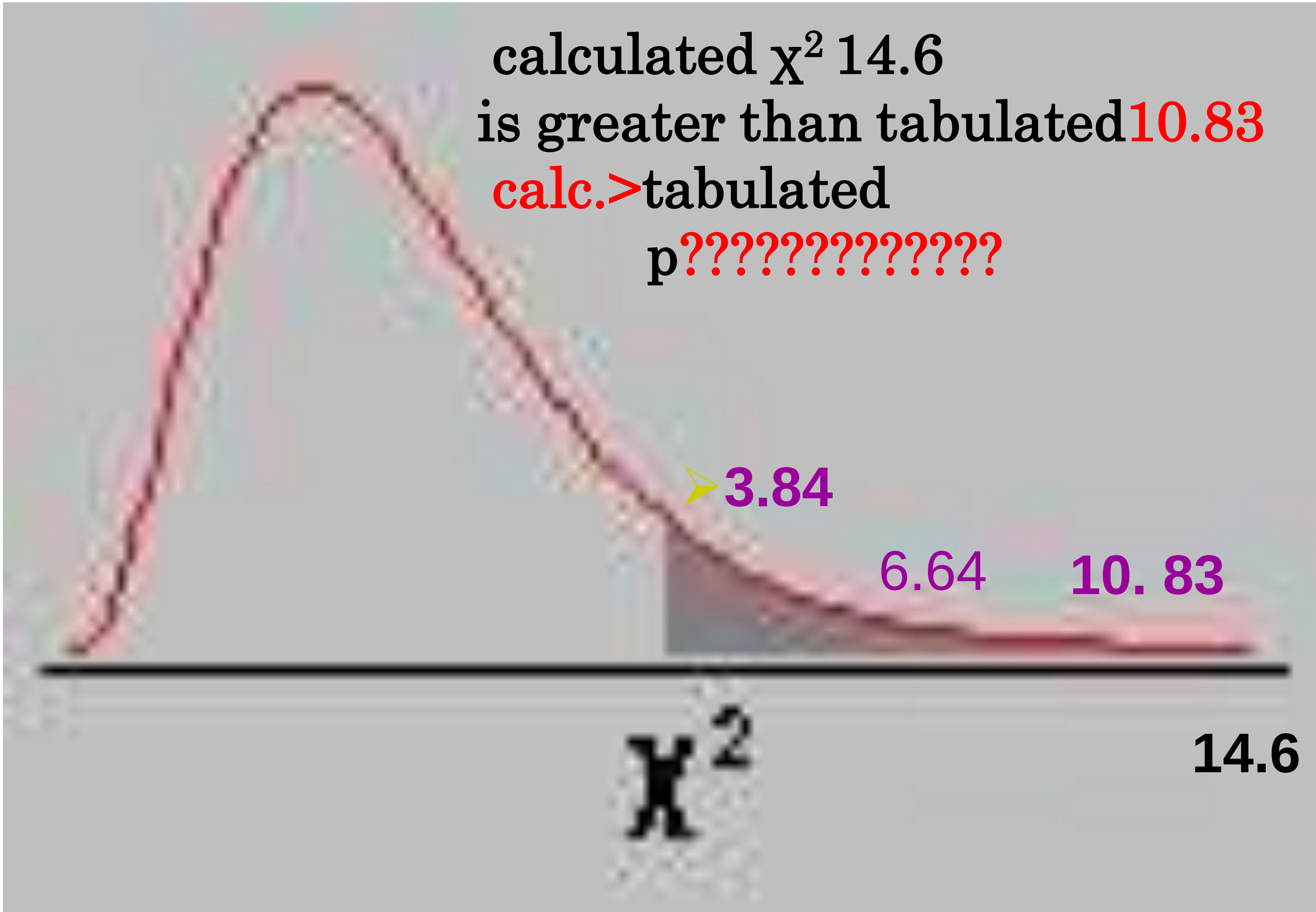


61	80.23	89.59	100.88
62	81.38	90.80	102.15
	82.53	92.01	103.46
	83.68	93.22	104.72
	84.82	94.42	105.97
	85.97	95.63	107.26
	87.11	96.83	108.54
68	88.25	98.03	109.79
69	89.39	99.23	111.06
70	90.53	100.42	112.31
71	91.67	101.62	113.56
72	92.81	102.82	114.84
73	93.95	104.01	116.08
74	95.08	105.20	117.35
75	96.22	106.39	118.60
76	97.35	107.58	119.85
77	98.49	108.77	121.11
78	99.62	109.96	122.36
79	100.75	111.15	123.60
80	101.88	112.33	124.84

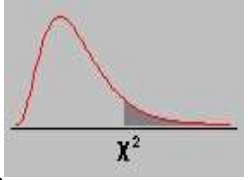
81	103.01	113.51	126.09
82	104.14	114.70	127.33
83	105.27	115.88	128.57
84	106.40	117.06	129.80
85	107.52	118.24	131.04
86	108.65	119.41	132.28
87	109.77	120.59	133.51
88	110.90	121.77	134.74
89	112.02	122.94	135.96
90	113.15	124.12	137.19
91	114.27	125.29	138.45
92	115.39	126.46	139.66
93	116.51	127.63	140.90



93	116.51	127.63	140.90
94	117.63	128.80	142.12
95	118.75	129.97	143.32
96	119.87	131.14	144.55
97	120.99	132.31	145.78
98	122.11	133.47	146.99
99	123.23	134.64	148.21
100	124.34	135.81	149.48


calculated χ2 14.6
is greater than tabulated 10.83
calc.>tabulated
p?????????????
3.84
6.64
10. 83
χ2
14.6

p is ?????????



- This mean that
- the probability is less than 0.001 that this difference is due to chance factor
  - And more than 99.999 that this difference due to smoking
- Thus there is a strong evidence against null hypotheses that is saying no effect of smoking on the probability of LBW.
- there is a strong evidence that LBW is related to smoking
- Therefore it is concluded that smoking is risk

p is ?????????

P > 0.05	**P > 0.01**	**P > 0.001**
p < 0.05	**p < 0.01**	**p < 0.001**

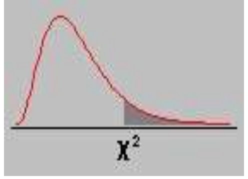
You can answer

if **p-value** associated with chi square is **less than 0.05 or** less than 0.01 you **reject** null hypoth.

And conclude that

- the two variable are **not independent** or

- there is a **statistically significant difference** in the **proportions**

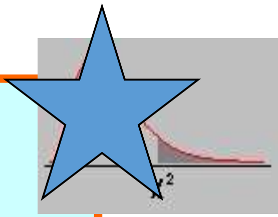


Continuity Correction

The chi square test for 2X2 table can be improved by using continuity correction we call it **Yates continuity correction** the formula become

$$X^2 = \sum \frac{\left[(O - E) - 0.5 \right]^2}{E} \qquad d.f. = 1$$

Resulting in small value for chi square

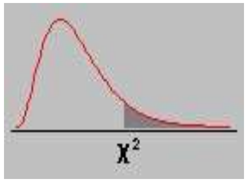


$$X_{\chi^2} = \frac{((14-8.1)-0.5)^2}{8.1} + \frac{((6-11.9)-0.5)^2}{11.9} +$$

$$\frac{((20-25.1)-0.5)^2}{25.1} + \frac{((44-30.1)-0.5)^2}{30.1}$$

$P < 0.001$

Thank You



Validity of χ^2

When the expected numbers are very small the chi square test is not good enough

We recommended other test (Exact Test)

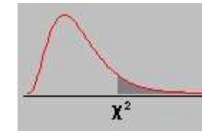
Thus $\chi 2$ is valid

- when the overall total is more than 40 , regardless the expected values

and

- when the overall total between 20 and 40 provided that all expected values are at least 5

Application of χ2



2 × 2 table .

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

r × c table .