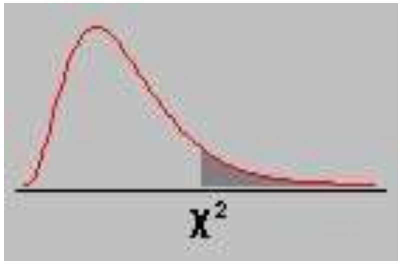
بسم الله الرحمن الرحيم

LXI

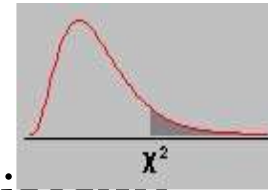


Chi Square (χ2) test

@ July 31- 2023

- **Prof. Dr. Waqar AL-Kubaisy**

SPECIFIC LEARNING OUTCOMES



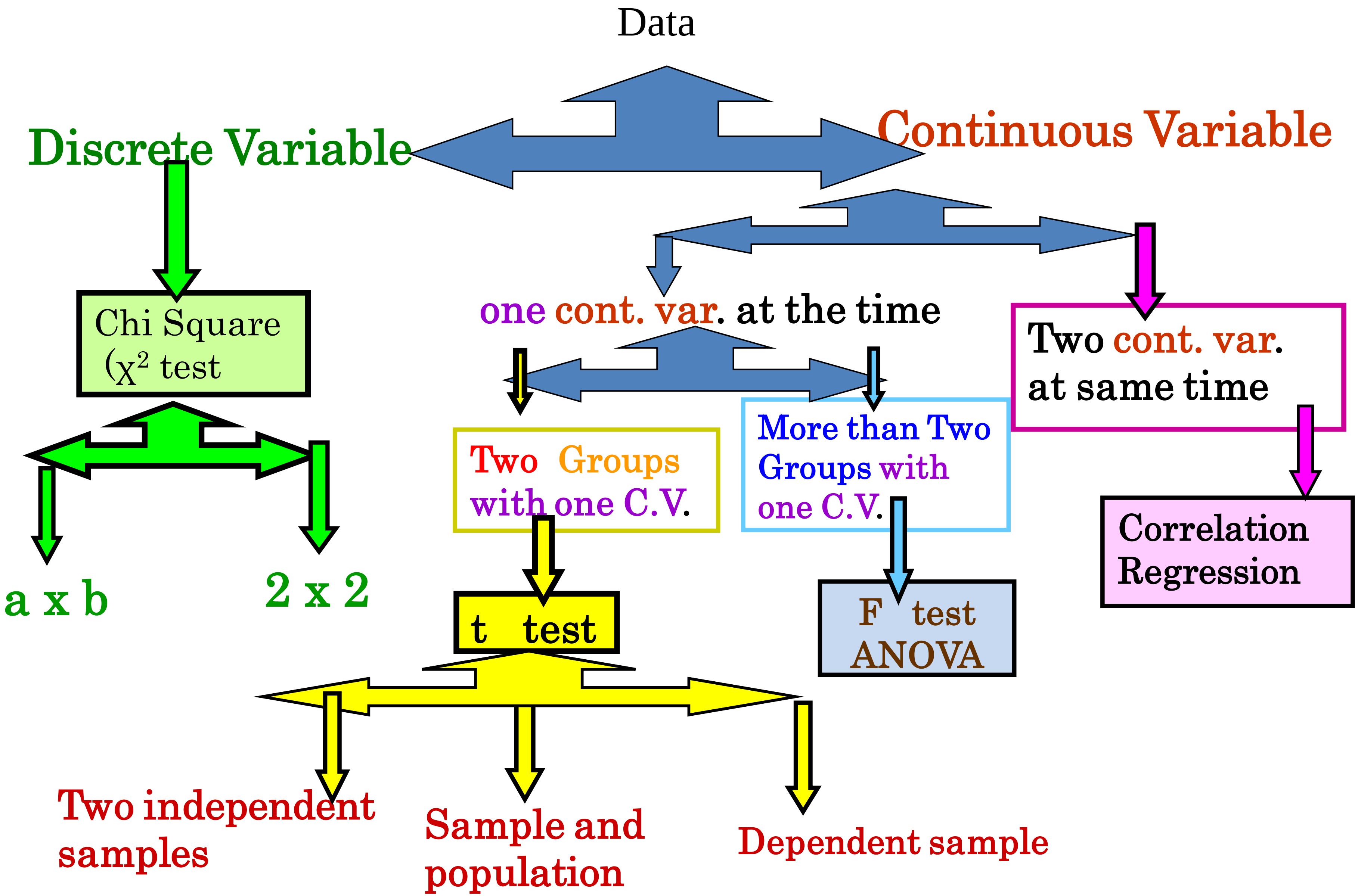
On completion of this lecture, you should be able to:

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.



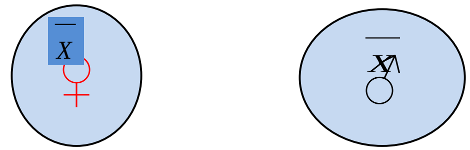
An important thing is the type of the variable concerned.

when the data measurement is continuous

t test be applied

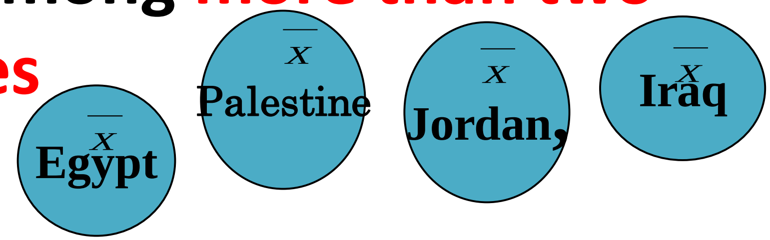
to test significance difference between two means

Body weight,



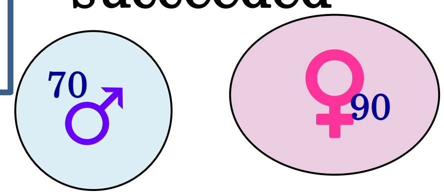
F test be applied

to test significance difference among more than two means Body weight adult males

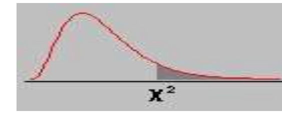


Numbers of students who were succeeded

succeeded



An important thing is the type of the variable concerned.



The data we have here is only enumerative data or **counting data .**

Counting No. of individuals falling in one category, class, group or another

30

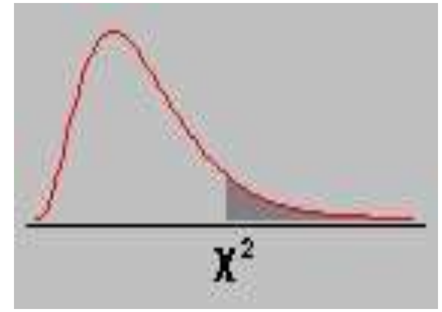
\+ **Age groups** .<20

Sex 20 – 24 **occupation**

25 – 29

The data consist of counting No. in each sample or group

An important thing is the type of the variable concerned.



180
Baghdad

170
Mutah

100
♀

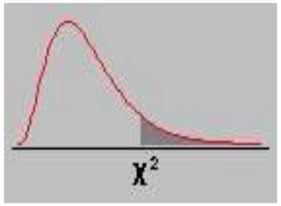
75
♂

Numbers of students who were succeeded

??????

????????

cause could be



	succeeded
Baghdad	180
Mutah	170

????? cause could be

	succeeded
Baghdad	180
UiTM	220
Syria	200
Mutah	170

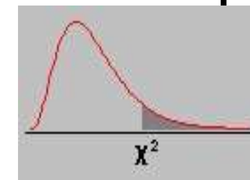
?????

Numbers of students who were succeeded

cause could be

Therefore

	Total	succeeded	%	Not succeeded
Baghdad	240	180	75%	60
Mutah	200	170	85%	30
	440	350		90

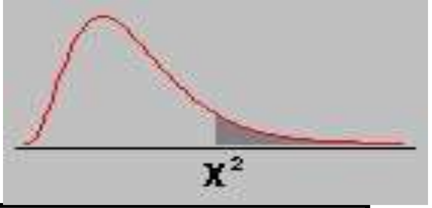


Proportion succeeded
350/440=0.80

Proportion succeeded at Mutah ??

Proportion succeeded at Baghdad ??

cause could be



	Total	succeeded	%	Not succeeded
Baghdad	220	180	82%	40
Mutah	200	170	85%	30
Syria	320	200	62.5%	120
UiTM	380	220	57.9%	160
	1120	770		350

770/1120 = 0.687

350/1120 = 0.3125

770/1120 X 100 = 68.7%

350/1120 X100 = 31.25%

When data measurement is

Qualitative data
counting data
Categorical data
Discrete.

The data consist of proportion of individuals in each group or sample,

- We have absolute numbers
- We have counting numbers
- so

- comparing between
- Rates , proportions of individuals in each group

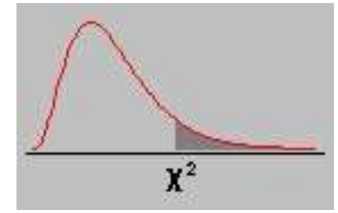
Two groups

More than two groups

statistical inference are made in term of difference in proportions

$$Ho = P_1 = P_2 = P_0$$

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



We classify persons into categories such as

	male	female	total
Present			
Absent			
total			

- male female
- smoker not smoker
- Succeeded and not succeeded…. etc smoker, not smoker and X smoker then

➢count the number of observation fall in each category

The result is frequency data

enumerative data because we
enumerate the No. of person in each category

Categorical data , because we
count the No. of person in each category

When measurement

When measurement is

merely the presence or absence of certain condition,

Absolute No X

✓ Proportion

the population parameter is

P: :the proportion of condition in population

which is estimated by

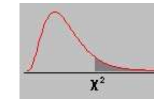
P: the proportion of condition in the sample

So

testing hypothesis about population proportion "P"

based on sample proportion P

is similar to testing hypothesis about μ .



The techniques for testing hypothesis concerning

Qualitative data
counting data
Categorical data
Discrete

is known as
chi square (χ^2) test .

<u>Chi square is</u>
used in testing difference in proportions

$$Ho = P_1 = P_2 = P_0$$

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

while t test and F test are used in testing difference in means .

Also classification could be more than 2 groups, could be three, four, five ………. K groups .

P1 P2 P3 P4 P5 ………… Pk

Tumour stage I II III ……..

Class stage level I II III IV V

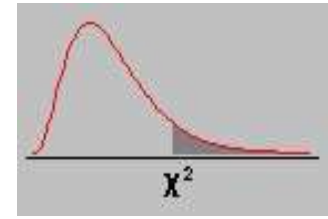
P1 P2 P3 P4 P5 ………… Pk

In this case

$$Ho = P_1 = P_2 = P_3 = P_4 = P_5 = P_0$$

$$H_A = P_1 \neq P_2 \neq P_3 \neq P_4 \neq P_5 \neq P_0$$

	Jordanian	Iraqi	Syrian	Egyptian	total
smoker					
Not smoker					
total					



When measurement is

merely the presence or absence of certain condition,

Absolute No X

✓ Proportion

the population parameter is

P: :the proportion of condition in population

which is estimated by

P: the proportion of condition in the sample

So

Testing hypothesis about population proportion "P"

based on sample proportion P

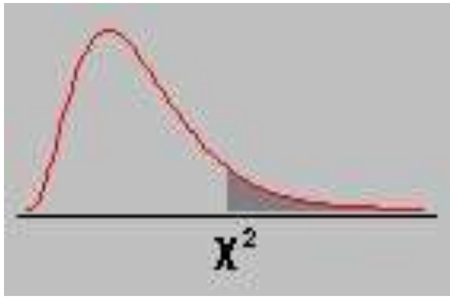
If the true population proportion of condition is Po

and sample size is N, So

Po N = total No. of condition that expected (E) in population .

Chi square test denoted X^2

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



This has two common applications:

first as test

whether two categorical variables are independent or not;

second as a test of

whether two proportions are equal or not

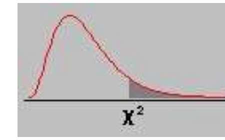
$$Ho = P_1 = P_2 = P_0$$

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

$$Ho = P_1 = P_2 = P_3 = P_4 = P_5 = P_0$$

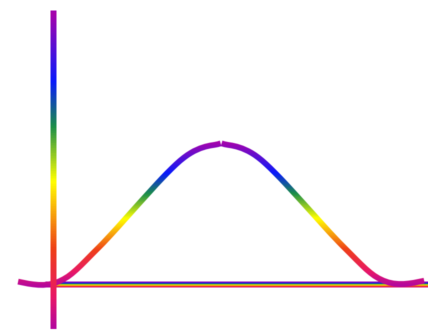
$$H_A = P_1 \neq P_2 \neq P_3 \neq P_4 \neq P_5 \neq P_0$$

contingency table



The chi square test is applied to frequency data in form of a contingency table i.e. a table of cross- tabulations) with the rows represent categories of one variable and the columns categories of a second variable.

	♂	♀	total
succeeded	70	90	160
not succeeded	10	30	40
Total	80	120	200

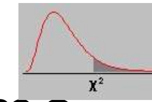


The null hypothesis is that the two variables are unrelated the rows represent categories of one variable and the columns categories of a second variable

Sex	succeeded	not succeeded	Total
♂	70	10	80
♀	90	30	120
Total	160	40	200

The H0; is that the two variables are unrelated

The HA ???????????????



If the variables display are Exposure and outcome.

Then

we usually we arrange the table with

Exposure as the row variable and

Out come as the column variable .

and display % corresponding the exposure variable

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

Example

smoking during pregnancy and relation to small birth weight

smoker or non smoked mother during pregnancy??

small birth weight no small birth weight ???

	♂	♀	total
succeeded	70	90	160
not succeeded	10	30	40
Total	80	120	200

SEX	succeeded	not succeeded	Total
♂	70	10	80
♀	90	30	120
Total	160	40	200

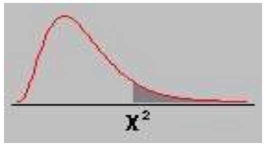
	♂	♀	total
succeeded	70	90	160
not succeeded	10	30	40
Total	80	120	200

????

merely the presence or absence of certain condition,

Absolute No X

✓ Proportion



	♂	♀	total
succeeded	70 87.5%	90 75%	160 80%
not succeeded	10 12.5%	30 25%	40
Total	80	120	200

If the true population proportion of condition is

160/200 =0.8 40/200 = 0.2

Po =0.8 and

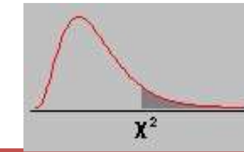
Rate (proportion) of succeeded ♂ (p_1)=70/80= 87.5%

Rate(proportion) of succeeded ♀ (p_2)= =90/120= 75%

$$Ho = P_1 = P_2 = P_0$$

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

????



	♂	♀	total
succeeded	70 (87.5%)	90 (75%)	160 80%
not succeeded	10 (12.5%)	30 (25%)	40
Total	80	120	200

If the true population proportion of condition is

160/200 =0.8 40/200 = 0.2

Po =0.8 and

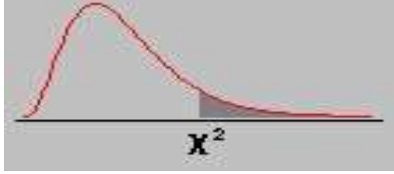
sample size is N, (200) So

Po N =Total No. of condition that expected (E)

in Each population .

♂ 80X 0.8= 80X 0.2 =

♀ 120X 0.8= 120X 0.2=



expected (E)

♂	80X.8=	80X.2 =
♀	120X.8=	120X.2=

	♂ O	♂ E	♀ O	♀ E	total
succeeded	70	64	90	96	160
not succeeded	10	16	30	24	40
Total	80		120		200

$$\sum O - E = Zero$$

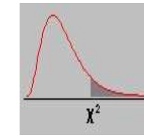
$$\sum \frac{O - E}{E} = Zero$$

the actual observed No. of subject with condition (O) and the expected No. of condition (E)

- Looking for the difference between the observed and expected frequencies

$$\sum O - E = Zero$$

$$\sum \frac{O - E}{E} = Zero$$



So if the actual No. of subject with condition observed No. (O) is close to the expected No. (E) then

the Ho will be not rejected ( ).

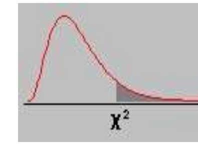
This mean that P=Po .

Usually summation $\sum O - E = Zero$ $\sum \frac{O - E}{E} = Zero$ So

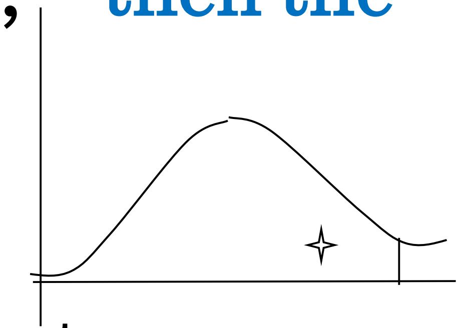
To overcome this result, we have to square O-E make it as $(O-E)^2$ then divided by E $\frac{(O - E)^2}{E}$ for each cell

Then we have to do the summation $\chi^2 = \sum \frac{(O - E)^2}{E}$

Therefore, χ^2 is always UPPER ONE SIDED TEST



❖ When O and E are close together, then the computed χ^2 is small and Ho is not Rejected .



❖ When O and E values are far apart
Then O-E is great, $(O-E)^2$ be more great
This will lead to Reject Ho .

In Enumerate (Discrete) value variable,
we classified individuals into :

Those having the condition P1

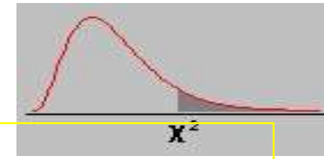
Those having no condition P2

	male	female	total
Present			
Absent			
total			

sign. Difference in proportion

$$Ho = P_1 = P_2 = P_0$$

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



Chi square (χ^2)

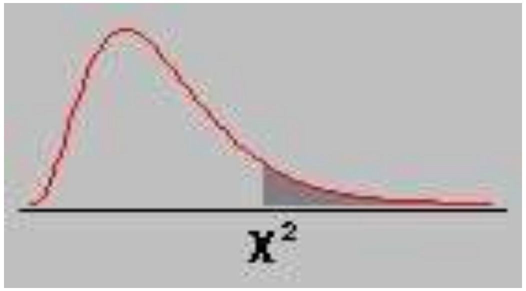
It is the sum of the squared difference between the observed frequency and expected frequency, divided by the expected frequency .

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

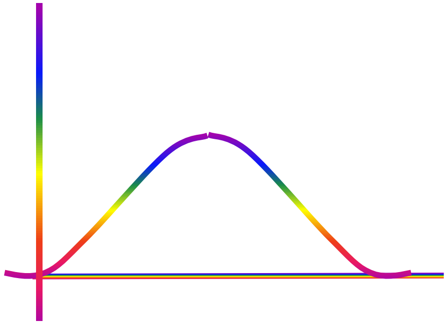
sign. Difference in proportion

Comparing calculated χ2 with tabulated χ^2 in relation to critical region

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

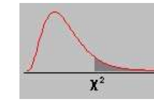


Therefore, χ2 is always UPPER ONE SIDED TEST



Comparing calculated χ2 with tabulated χ²
in relation to critical region

sign. Difference in proportion



Chi square is

used in testing difference in proportions

while t test and F test are used in testing difference in means .

$$Ho = P_1 = P_2 = P_0$$

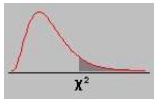
$$H_A = P_1 \neq P_2 \neq P_0$$

Chi square (χ^2)

It is the sum of the squared difference between the observed frequency and expected frequency, divided by the expected frequency .

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

Comparing calculated χ^2 with tabulated χ^2
in relation to critical region

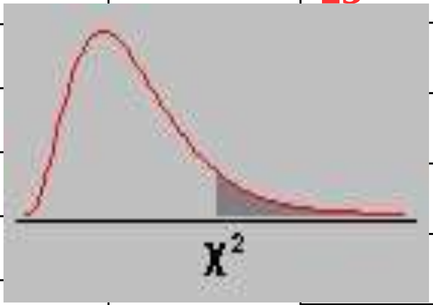


If the variables display are Exposure and outcome.
Then
we usually we arrange the table with
exposure as the row variable and
out come as the column variable .
and display % corresponding the exposure variable

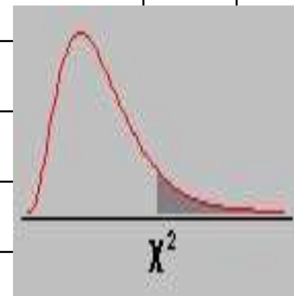
Exposure	Out come +ve	Out come -ve	total
yes			
no			
Total			

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
	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

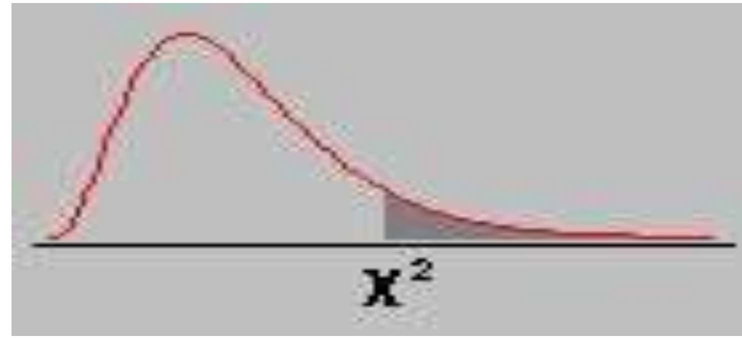


41	56.94	64.95	74.75
42	58.12	66.21	76.09
43	59.30	67.46	77.42
44	60.48	68.71	78.75
45	61.66	69.96	80.08
46	62.83	71.20	81.40
47	64.00	72.44	82.72
48	65.17	73.68	84.03
49	66.34	74.92	85.35
50	67.51	76.15	86.66
51	68.67	77.39	87.97
52	69.83	78.62	89.27
53	70.99	79.84	90.57
54	72.15	81.07	91.88
55	73.31	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	99.62



61	80.23	89.59	100.88
62	81.38	90.80	102.15
[illegible]	82.53	92.01	103.46
[illegible]	83.68	93.22	104.72
[illegible]	84.82	94.42	105.97
[illegible]	85.97	95.63	107.26
[illegible]	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

Thank You

Application of χ2.

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

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