Chi-square test (x^2)

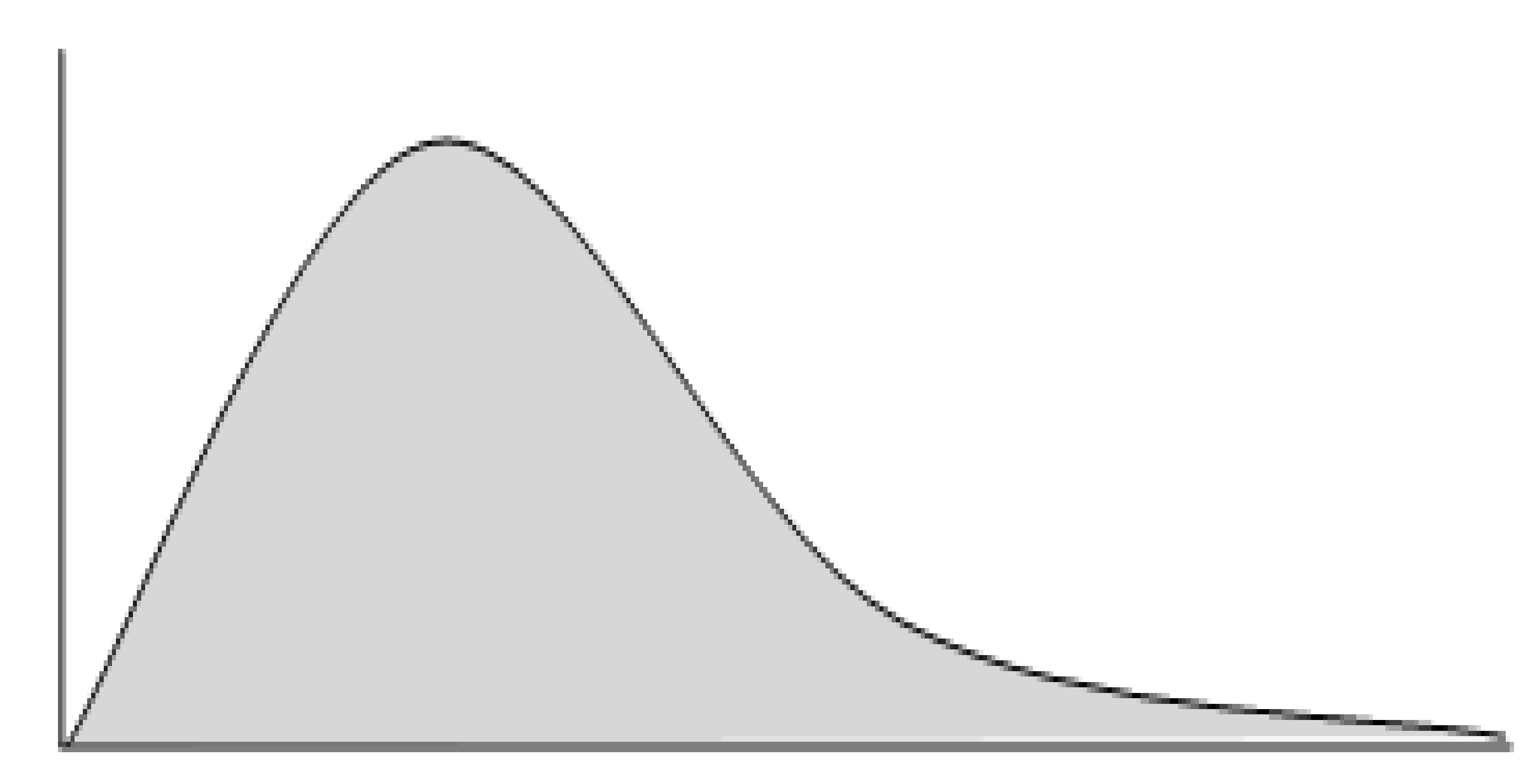
Dr. Nadhim Ghazal

- The chi-square test is the most frequently employed statistical technique for the analysis of count or frequency data. For example we may know for a sample of hospitalized patients how many are male and how many are female. For the same sample we may also know how many have complications and how many did not have complications. We may wish to know, for the population from which the sample was drawn, if the development of complications differs according to gender. The test is designated by the Greek letter chi (x^2) and the distribution is called chi-square distribution which has a shape quite different from the general normal distribution of having a value between 0 and infinity. It can not take on negative values, since it is the sum of values that have been squared.

The characteristics of chi distribution;

1. It always takes a positive value
2. It may be derived from the normal distribution, so no need for the assumption that the sample is taken from normally distributed population
3. It has one accepting region and one rejecting region.
4. The degree of freedom depend on the number of groups of subgroups than on the sample size

Chi-square test is most appropriate for use to test the significance of difference for the qualitative data (categorical variables) such as marital status, sex, and disease type, etc.., comparing different proportion and test the significance of difference between these proportions by employing or using the frequency in the calculation to reach the conclusion.



We have an observed frequency (the number of objects or subjects in our sample that fall into the various categories of the variable of interest) and an expected frequency (the number of objects or subjects in our sample that we would expect to observe if some null hypothesis about the variable is true).

The chi-square test has the following formula;

$$x^2 = \sum \frac{(0 - E)^2}{E}$$

Applications of Chi-square test:

1. Goodness-of-fit
2. The 2 x 2 chi-square test (contingency table, four fold table)
3. The a x b chi-square test (r x c chi-square test)

1- Goodness of fit:

In this application of chi square test we are going to test the goodness of fit of sample distribution (observed frequency "O") of a qualitative variable of one sample with a theoretical (preconceived or hypothesized) distribution (theoretical, expected, estimated frequency "E"), that could be one of the following theoretical distributions; Normal population distribution, Prevalence, Incidence, or Ratio.

Example;

The following data represents the sex distribution of patients with duodenal ulcer (DU). A sample of 120 patients were selected randomly from patients with DU, they were 70 males and 50 females. What is your conclusion if you know that M:F ratio in the population is 1.1:1 (use $\alpha=0.05$).

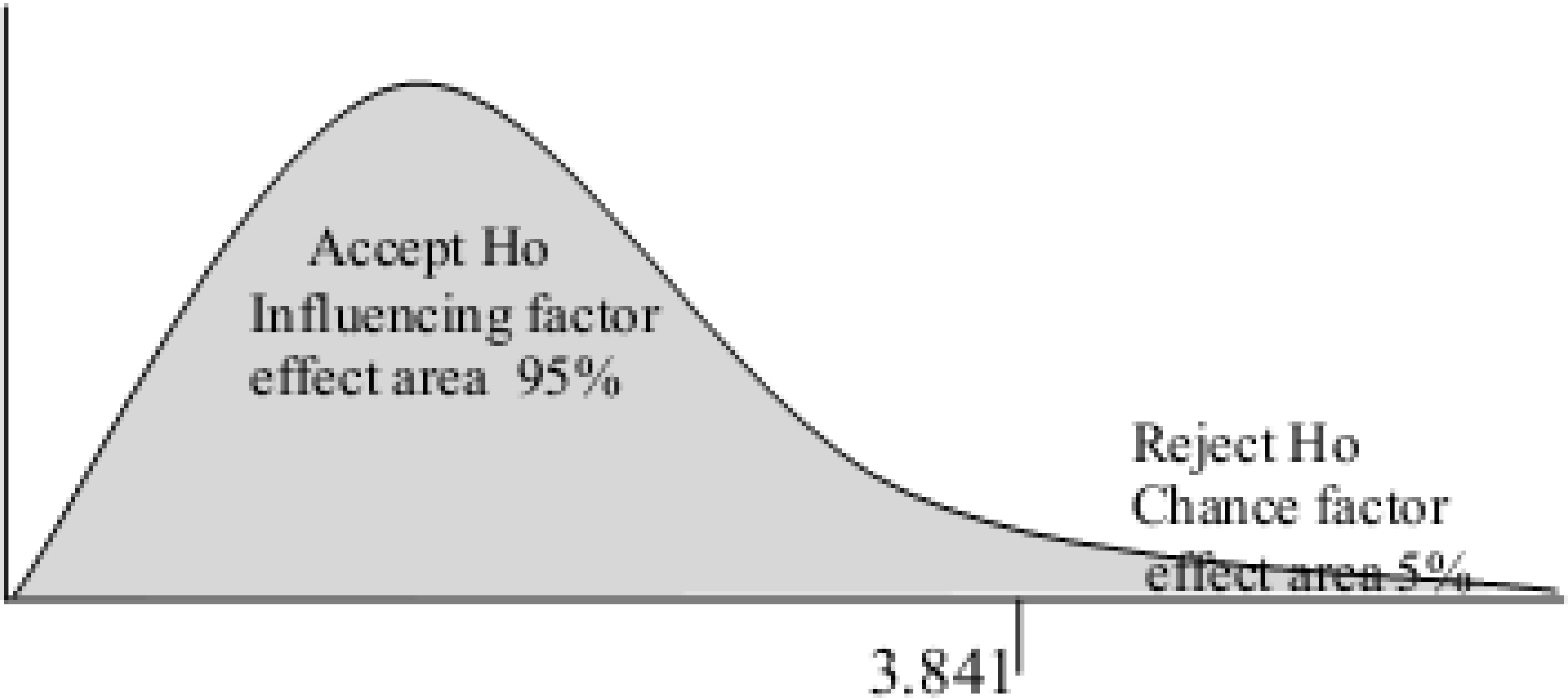
Percentage of males with DU = 70/120 x 100 = 58.3%

Percentage of females with DU = 50/120 x 100 = 41.7%

- **Data:** Data represent a sample of DU patients selected randomly, 58.3% males and 41.7% females with 1.1:1 M:F ratio in the population.
- **Assumption:** We assume that the sample of 120 DU patients was selected randomly from a population of DU patients.
- **HO:** There is no significant difference between the proportions of males and females with DU from population proportions. OR There is no significant association between sex and DU.

- **H_A:** There is significant difference between the proportions of males and females with DU from population proportions. OR There is significant association between sex and DU.
- **Level of significance;** (α = 0.05); 5% Chance factor effect area 95% Influencing factor effect area (association between DU and sex) d.f.=K-1; (K=Number of subgroups).

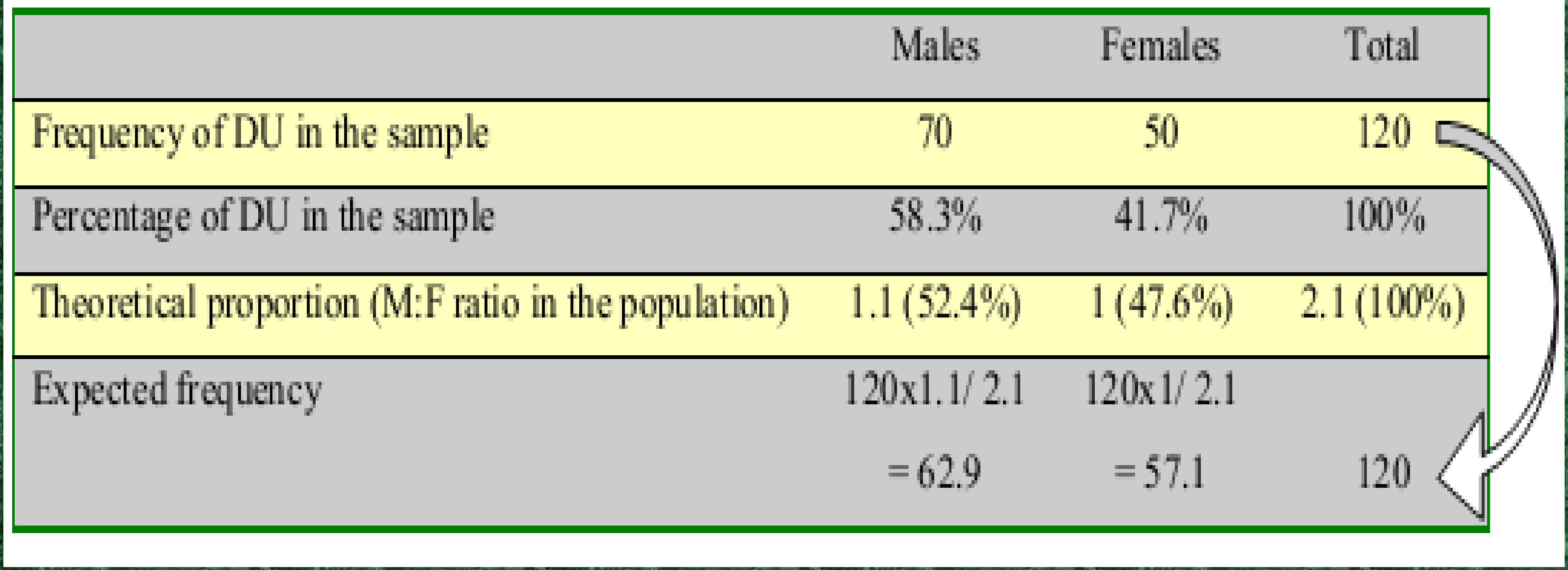
Tabulated χ^2 for d.f {(K-1)=2-1=1}, for α 0.05 equal to 3.841



- Apply the proper test of significance

$$x^2 = \sum \frac{(0 - E)^2}{E}$$

$$= \frac{(01 - E1)^2}{E1} + \frac{(02 - E2)^2}{E2}$$



	Males	Females	Total
Frequency of DU in the sample	70	50	120
Percentage of DU in the sample	58.3%	41.7%	100%
Theoretical proportion (M:F ratio in the population)	1.1 (52.4%)	1 (47.6%)	2.1 (100%)
Expected frequency	120x1.1/ 2.1 = 62.9	120x1/ 2.1 = 57.1	120

$$= \frac{(01 - E1)^2}{E1} + \frac{(02 - E2)^2}{E2}$$

$$= \frac{(70 - 62.9)^2}{62.9} + \frac{(50 - 57.1)^2}{57.1}$$

= 0.8015 + 0.8829 = 1.6844 (calculated value) (Calculated chi)

1.6844 < 3.841

Since Calculated x^2 < Tabulated x^2

So P>0.05

Then accept Ho.... (Ho Not rejected)

- There is no significant difference between the proportions of males and females with DU from population proportions.
- There is no significant association between sex and DU
- Sex distribution in patients with DU follows the sex distribution in the population

2- **2x2 chi square test:**

Sometimes when we consider two groups for comparison and the variable of interest have a criteria of two in the classification, so the data here will be cross classified in a manner resulted in a contingency table consisting of two rows and two columns, such a table is referred to as 2x2 or four-fold or contingency table. For such table the degree of freedom will be determined by applying the rule of (r-1)(c-1) which will result in (2- 1)(2-1)= 1x1=1 degree of freedom.

In this application of chi square test we are going to compare the of sample distribution (observed frequency "O") of a qualitative variable of two samples with a theoretical (preconceived or hypothesized) distribution (theoretical, expected, estimated frequency "E"), that can be estimated using the following rule

$$E = \frac{(Tr\ X\ Tc)}{Trc}$$

Total of row (Tr); Total of column (Tc); Grand total (Trc) for each cell.

Example;

A researcher interested in studying the association between cancer of bladder (ca bladder) and smoking. He took the records of 20 patients with ca bladder and compared them with 200 healthy control subjects selected at random from the population. He found that half of patients with ca bladder were smokers and only 20 of the healthy controls were smoker. What is the conclusion he reached from such data (use $\alpha=0.05$).

Percentage of smoking in ca bladder = 10/20 x 100 = 50%

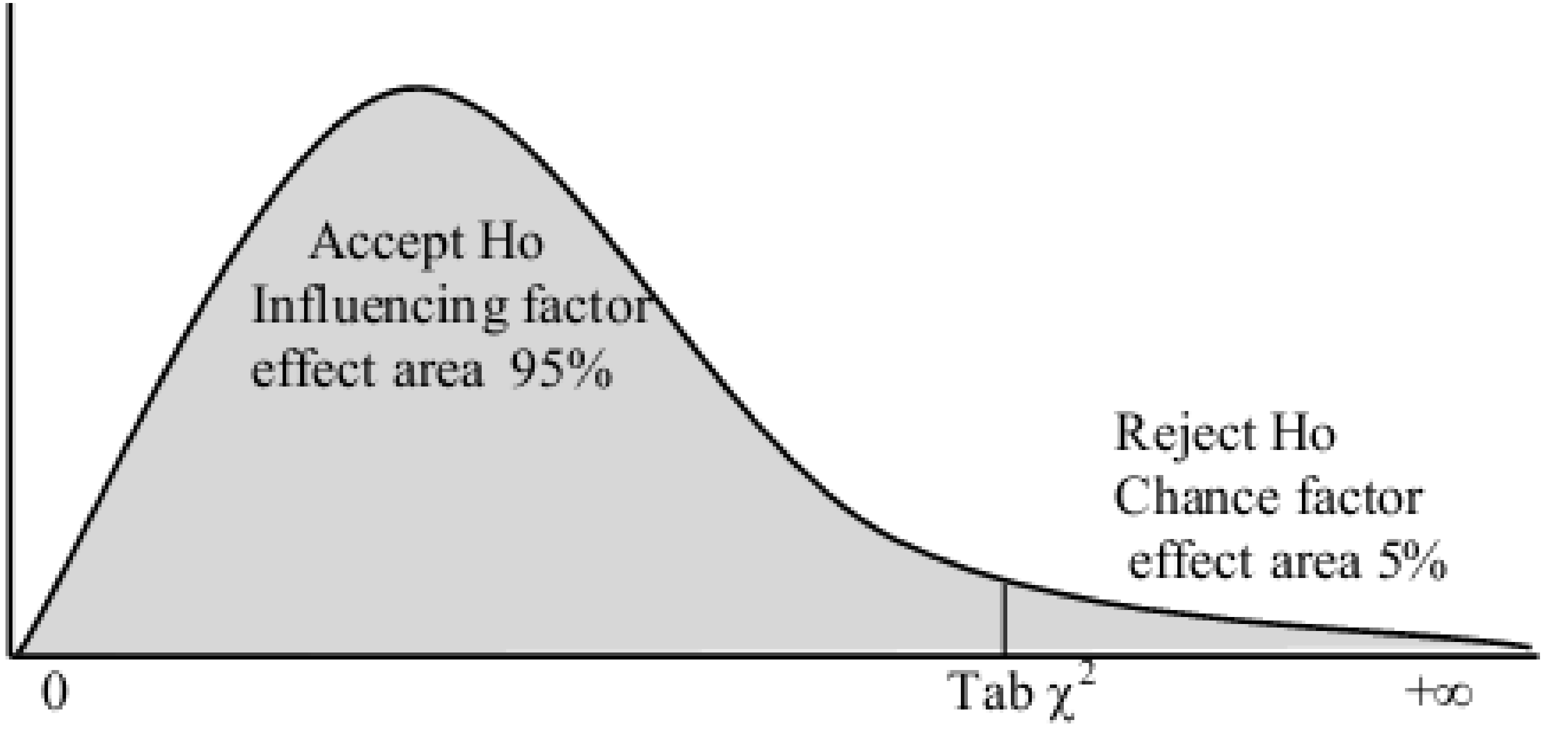
Percentage of smoking in healthy subjects = 20/200 x 100 = 10%

- **Data:** Data represent two samples, the first one consist of 20 patients with ca bladder, half of them smokers (50%), and the other sample of 200 healthy subjects, 10% of them were smoker.
- **Assumption:** We assume that the two independent groups were selected randomly from two independent populations.
- **H_O:** There is no significant difference between the proportions of smokers and non-smoker with or without ca bladder. OR There is no significant association between smoking and ca bladder.

- **H_A :** There is significant difference between the proportions of smokers and non-smoker with or without ca bladder. OR There is significant association between smoking and ca bladder.
- **Level of significance**; (α = 0.05); 5% chance factor effect area 95% influencing factor effect area (association between ca baldder and smoking)

d.f.=(r-1)(c-1)= (2-1)(2-1)=1x1=1

Tabulated χ^2 for d.f =1, for α 0.05 equal to 3.841

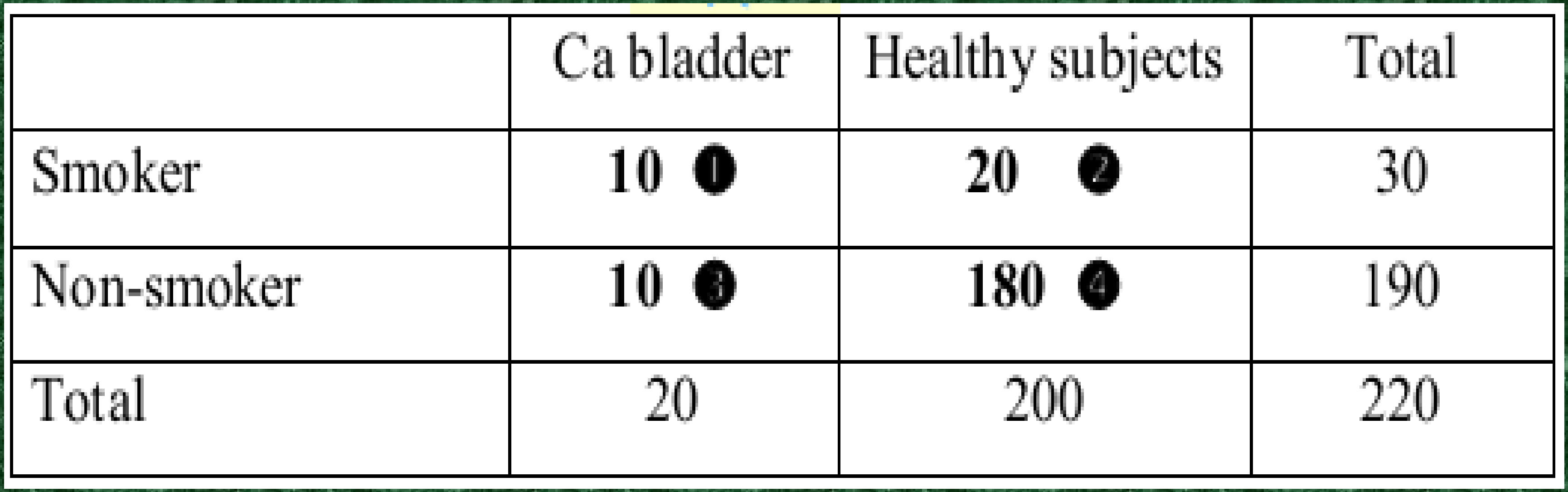


3.841

6.635 (0.01)

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

$$= \frac{(O1 - E1)^2}{E1} + \frac{(O2 - E2)^2}{E2} + \frac{(O3 - E3)^2}{E3} + \frac{(O4 - E4)^2}{E4}$$



	Ca bladder	Healthy subjects	Total
Smoker	**10** ❶	**20** ❷	30
Non-smoker	**10** ❸	**180** ❹	190
Total	20	200	220

$$\mathbf{E = \frac{Tr \times Tc}{Trc}}$$

$$E1 = \frac{Tr \times Tc}{Trc} = \frac{30 \times 20}{220} = 2.7$$

$$E2 = \frac{Tr \times Tc}{Trc} = \frac{30 \times 200}{220} = 27.3$$

$$E3 = \frac{Tr \times Tc}{Trc} = \frac{190 \times 20}{220} = 17.3$$

$$E4 = \frac{Tr \times Tc}{Trc} = \frac{190 \times 200}{220} = 172.7$$

$$= \frac{(10-2.7)^2}{2.7} + \frac{(20-27.3)^2}{27.3} + \frac{(10-17.3)^2}{17.3} + \frac{(180-172.7)^2}{172.7}$$

= **19.36** + 1.93 + 3.06 + 0.30 = **24.65** (calculated value) (Calculated chi)

24.65 > 3.841

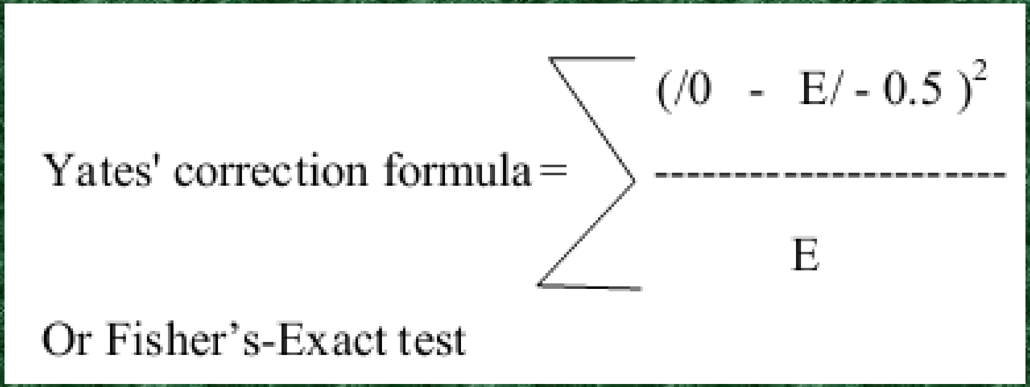
- Since Calculated x^2> Tabulated x^2
- So P< 0.05
- Then reject Ho and accept H_A
- There is significant difference between the proportions of smokers and non-smoker with or without ca bladder.
- There is significant association between smoking and ca bladder
- Smokers with high proportion to develop ca bladder 50% versus 10%, means five times more to develop ca bladder than healthy subjects.

- **3- a x b chi square test:**
- The same application as 2x chi square test, but there are either more groups than two or the qualitative variable of interest is classified to more than two categories so we will have 2x3 or 3x2 or 3x3 etc… so we have r=2 or more and c=2 or more. The difference will be seen in the degree of freedom it will be more than 1.

- **Important notes in chi-square test**;

1- When 2x2 chi-square test have a zero cell (one of the four cells is zero) we can not apply chi-square test because we have what is called a complete dependence criteria. But for a x b chi-square test and one of the cells is zero when can not apply the test unless we do proper categorization to get rid of the zero cell.

2- When we apply 2x2 chi-square test and one of the expected cells was <5 or when we apply a x b chi-square test and one of the expected cells was <2, or when the grand total is <40 we have to apply Yates' correction formula;



$$\text{Yates' correction formula} = \sum \frac{(/0\ -\ E/ - 0.5\ )^2}{E}$$

Or Fisher’s-Exact test

- 3- In case of 2x2 or a x b or even goodness of fit result in significant difference in proportion and significant association, the significancy came from that cell that have a "small x^2 " value of more than 3.841

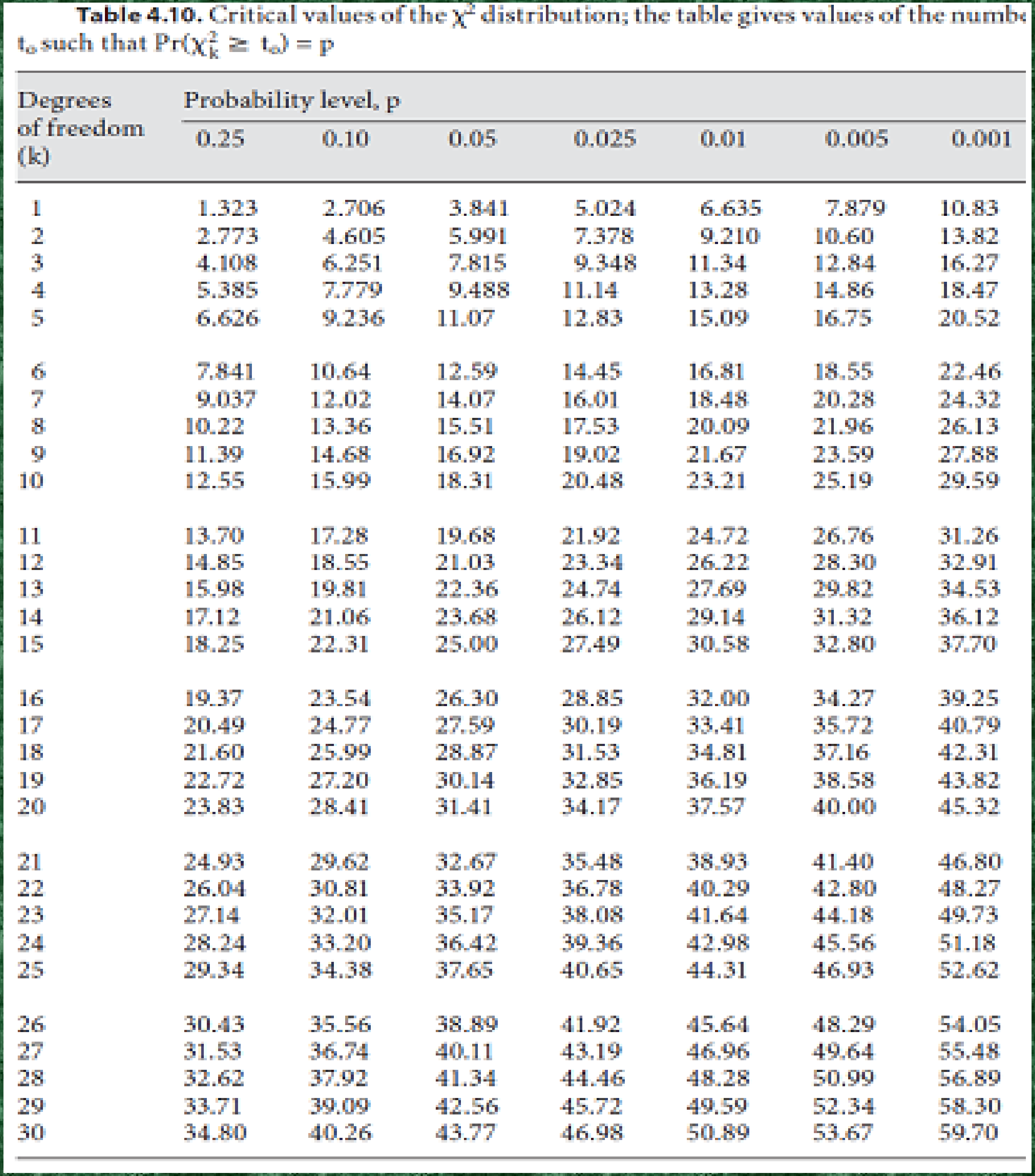
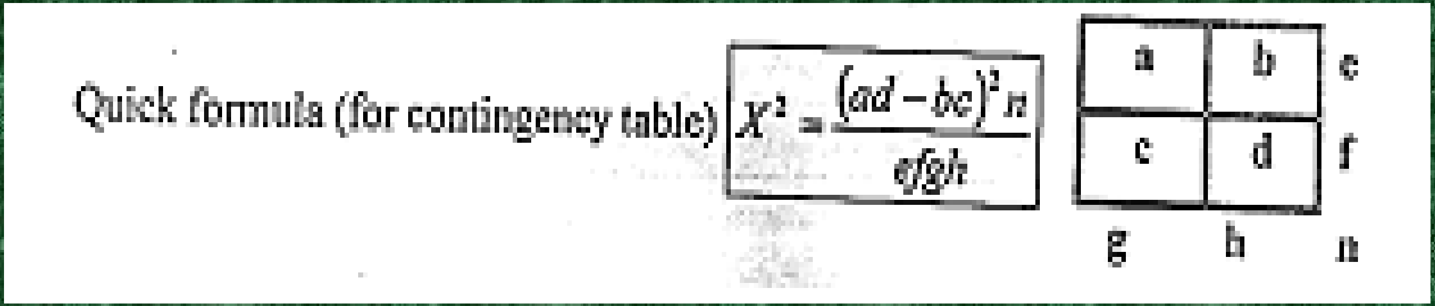


Table 4.10. Critical values of the χ^2 distribution; the table gives values of the numbe t_o such that $\Pr(\chi^2_k \geq t_o) = p$

Degrees of freedom (k)	Probability level, p						
	0.25	0.10	0.05	0.025	0.01	0.005	0.001
1	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	2.773	4.605	5.991	7.378	9.210	10.60	13.82
3	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	6.626	9.236	11.07	12.83	15.09	16.75	20.52
6	7.841	10.64	12.59	14.45	16.81	18.55	22.46
7	9.037	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.13
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70



Quick formula (for contingency table) $\boxed{X^2 = \frac{(ad-bc)^2 n}{efgh}}$

a	b	e
c	d	f
g	h	n

END