

A green ribbon banner with a central rectangular section and pointed ends on both sides. The text is centered within the central section.

Chi-square test (χ^2)

A red, horizontally-oriented oval with a slight gradient and a drop shadow.

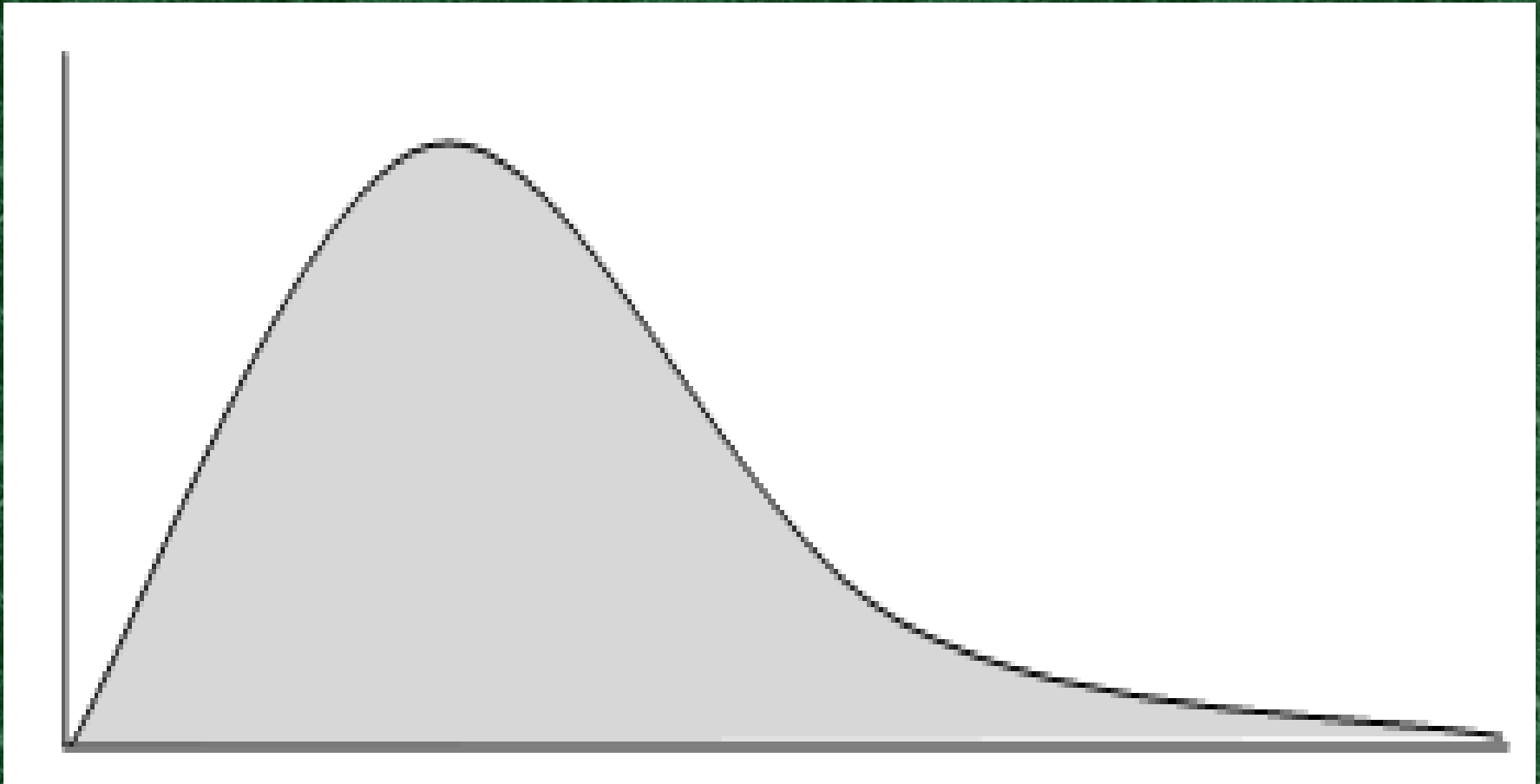
Lecture seven

- 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 (χ^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;

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

Applications of Chi-square test:

1. Goodness-of-fit
2. The 2×2 chi-square test (contingency table, four fold table)
3. The $a \times b$ chi-square test ($r \times c$ 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 $\alpha=0.05$).

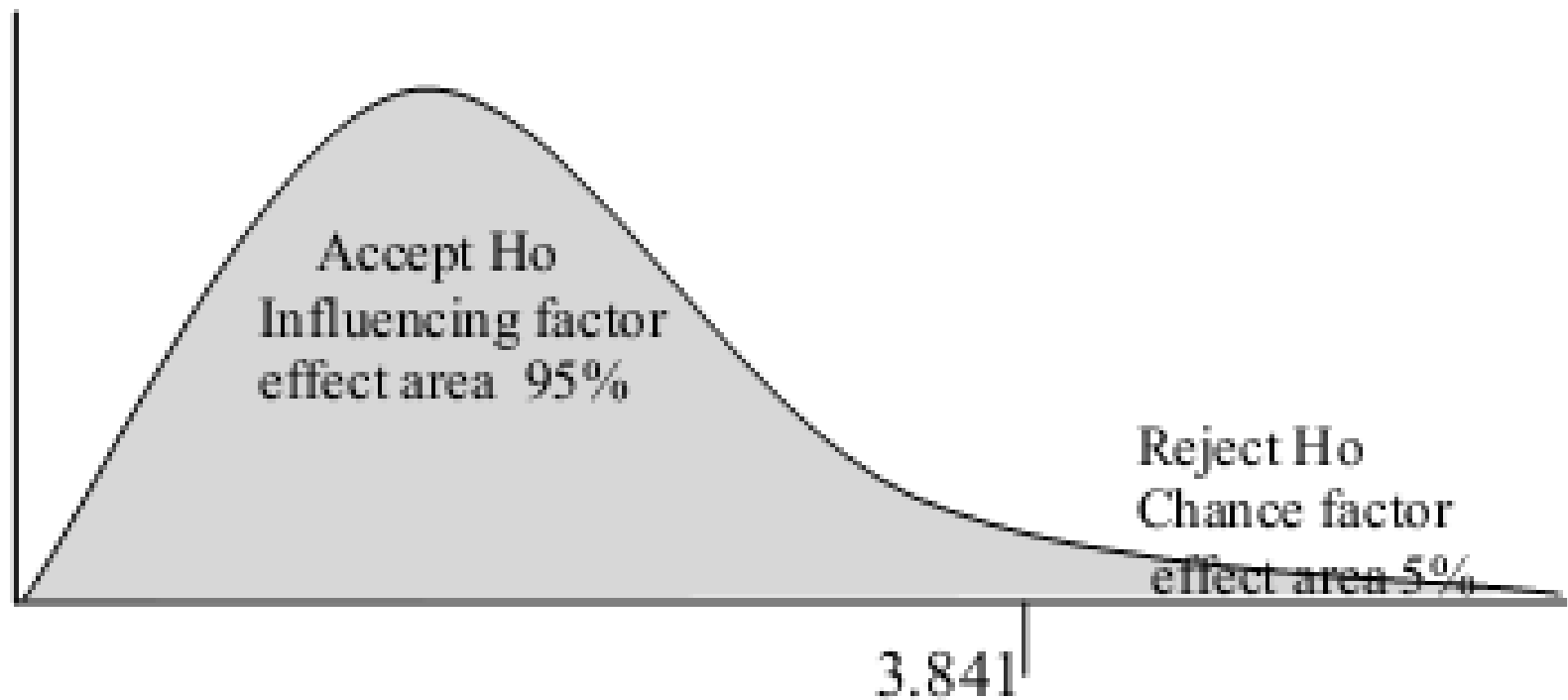
Percentage of males with DU = $70/120 \times 100 = 58.3\%$

Percentage of females with DU = $50/120 \times 100 = 41.7\%$

- **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; ($\alpha = 0.05$); 5%**
Chance factor effect area 95%
Influencing factor effect area
(association between DU and sex)
d.f.=K-1; (K=Number of subgroups).

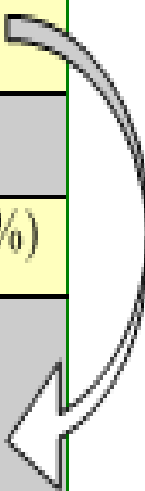
Tabulated χ^2 for d.f $\{(K-1)=2-1=1\}$, for α 0.05 equal to 3.841



- Apply the proper test of significance

$$\begin{aligned} \chi^2 &= \sum \frac{(O - E)^2}{E} \\ &= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \end{aligned}$$

	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	$120 \times 1.1 / 2.1$ = 62.9	$120 \times 1 / 2.1$ = 57.1	120



$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

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

$$= 0.8015 + 0.8829 = 1.6844 \text{ (calculated value)}$$

(Calculated chi)

$$1.6844 < 3.841$$

Since Calculated $\chi^2 <$ Tabulated χ^2

So $P > 0.05$

Then accept H_0 (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 $(2-1)(2-1)=1 \times 1=1$ 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

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

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 = $10/20 \times 100 = 50\%$

Percentage of smoking in healthy subjects = $20/200 \times 100 = 10\%$



END