Chi-square Distribution {χ^2 -Distribution}

- Chi-square: is defined as the sum 0f squares of independent normally distributed variables with zero mean and unit variance.

$$\chi^2 = \sum_{i=1}^{v} Z_i^2 = \sum_{i=1}^{v} \left(\frac{x_i - \mu}{\sigma_i} \right)^2 = Z_1^2 + Z_2^2 + \ldots + Z_v^2$$

- The mean and variance of chi-square distribution are
- .

$$m = v \qquad \text{and} \qquad s^2 = 2v \;, \quad v = df$$

- Goodness-Of-Fit Test:
- This test refers to the comparison between some observed sample distribution and theoretical frequency distribution. This test is based on the quantity

$$c^2 = \sum \frac{(o_i - e_i)^2}{e_i}, \quad \text{where } o_i = observed \text{ and } e_i = \exp ected\ values$$

- Example: The following table shows the age distribution of cases of a certain diseases reported during a year in state. Test the hypothesis that these data comes from a normal distribution. Let $\alpha = 0.05$.
- From the following table compute $\bar{x} = 35.233$ and $s = 12.963$

Age	Observed $f_i = O_i$	$Z = \frac{x_i - \bar{x}}{s}$	Expected Relative Frequency	Expected Frequency (e_i)	$\{C^2\}$ $\frac{(o_i - e_i)^2}{e_i}$
<5	0		0.0099	0.7425	0.7425
5 - 14	5	-2.33	0.0495	3.7125	0.4465
15 - 24	10	-1.56	0.1554	11.6550	0.2350
25 - 34	20	-0.79	0.2772	20.7900	0.0300
35 - 44	22	-0.02	0.2814	21.1050	0.0379
45 - 54	13	+0.75	0.1623	12.1725	0.0562
55 - 64	5	+1.52	0.0643	4.8225	0.0065
	75		1.0000	75.0000	1.5546

$$Z = \frac{x_i - \bar{x}}{s} = \frac{5 - 35.233}{12.963} = -2.33$$

- H_0: Data were drawn from normally distributed population.
- H_1: Data were not drawn from normally distributed population.

$$c^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}, \qquad \{\text{Befor Combination}\}$$

$$c^2 = \frac{(0 - 0.7425)^2}{0.7425} + \frac{(5 - 3.7125)^2}{3.7125} + \ldots + \frac{(5 - 4.8225)^2}{4.225} = 1.5546$$

- Note: Classes with expectations bellow 1 should combine, and k is the number of classes after such combinations have been made.

$$c^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}, \qquad \{\text{After Combination}\}$$

$$c^2 = \frac{(5 - 4.455)^2}{4.455} + \frac{(10 - 11.6550)^2}{11.6550} + \ldots + \frac{(5 - 4.8225)^2}{4.8225} = 0.4323$$

$$4.455 = 0.7425 + 3.7125$$

- *The number of degree of freedom (k-1-r), in a c^2 goodness-of-fit test is equal to the number of cells minus the number of quantities obtained from the observed data that are used in the calculations of expected frequencies.*
- *df = k-1-r* = 6-1-2= 3 (after combination)

From precentiles of the chi-sequare distribution in Table

$$c^2_{(1-a),v} = c^2_{(0.95),3} = 7.815$$

- *Statistical decision*: Accept H_0, since $c^2 < 7.815$
- *Clinical decision*: Conclude that, on the bases of these data, the sample came from normally distributed population. (P>0.05).

- Example: Calculation of χ^2 goodness-of-fit of data consisting of 100 hair colors to a hypothesized color ratio of 3:1. Let α =0.05.

	Hair color		
	Black	Brown	Total
$f_i = O_i$	84	16	100
$F_i = e_i$	75	25	100

e_1= 100*3/4=75 , e_2 = 100*1/4 =25

- H0: The sample data came from a population having a 3:1 ratio of black to brown hair.
- H1: The sample data came from a population not having a 3:1 ratio of black to brown hair.

$$c^2 = \frac{(84-75)^2}{75} + \frac{(16-25)^2}{25} = 4.320$$

From precentiles of the chi-sequare distribution table,

$$c^2_{(1-a),v} = c^2_{(0.95),1} = 3.841$$

- *Statistical decision*: Reject H0, since *Clinical decision*: Conclude that, on the bases of these data, the sample not having a ratio 3:1 of black to brown hair. (P<0.05).

Test of Independence:

- Use of the chi-square distribution to test the null hypothesis that the two criteria of classification, when applied to the same set of entities are independent. *We say that the two criteria of classification are independent if the distribution of one criterion is the same no matter what the distribution of the other criterion.* This test is based on the quantity

$$c^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

where $o_{ij} = observed$ and $e_{ij} = \exp ected\ values$

- This test is used when the observed number of entities, falling into each cell was determined after the sample was drawn as a result, the row and column totals are chance quantities not under control of the investigator.

- NOTE: {*Single sample drawn from single population*}.

Example: 500 elementary school children were cross classified by socioeconomic group and the presence and absence of a certain speech defect. The results were as follows:

Socioeconomic Group

Speech	Upper		Upper Middle		Lower Middel		Lower		Total
Defect	*o*	*e*	*o*	*e*	*o*	*e*	*o*	*e*	
Present	8	(9.1)	24	(26.39)	32	(30.94)	27	(24.57)	91
Absent	42	(40.9)	121	(118.61)	138	(139.06)	108	(110.43)	409
Total	50		145		170		153		500

Are these data compatible with the hypothesis that the speech defect is unrelated to socioeconomic status? Let $\alpha = 0.05$.

e_{11} =91*50 / 500 = 9.1 , *e*23 =409*170 / 500 =139.6

H_0: Speech Defect and socioeconomic status are independent.

H_1: Speech Defect and socioeconomic status are dependent.

$$c^2 = \frac{(8-9.1)^2}{9.1} + \frac{(24-26.39)^2}{26.39} + \ldots + \frac{(108-110.43)^2}{110.43} = 0.765$$

df = (r-1) (c-1) = (2-1) (4-1) = 3

From precentiles of the chi-sequare distribution table,

$$c^2_{(1-a),v} = c^2_{(0.95),3} = 7.815$$

Statistical decision: Accept H_0, since $\chi^2 < 7.815$

Clinical decision: Conclude that, on the bases of these data, The speech effect and socioeconomic status are independent. (P>0.05).