-7-

The χ^2 distribution

7.1 Introduction

This is a continuous distribution that is widely used to ascertain whether different observations are independent or related, through what is known as **independence test**. This test is also used to check whether a random variable follows a certain distribution in what is known as **goodness of fit test.**

The parameter included in this distribution is K = number of degrees of freedom, to be defined later.

The density function of this distribution is:

$$f(x) = \frac{\left(\frac{1}{2}\right)^{\frac{K}{2}} . x^{\frac{K}{2}-1} . e^{-\frac{x}{2}}}{\Gamma\left(\frac{K}{2}\right)} \qquad (7.1)$$

This function, when plotted against x shows different curves corresponding to different values of the parameter K.

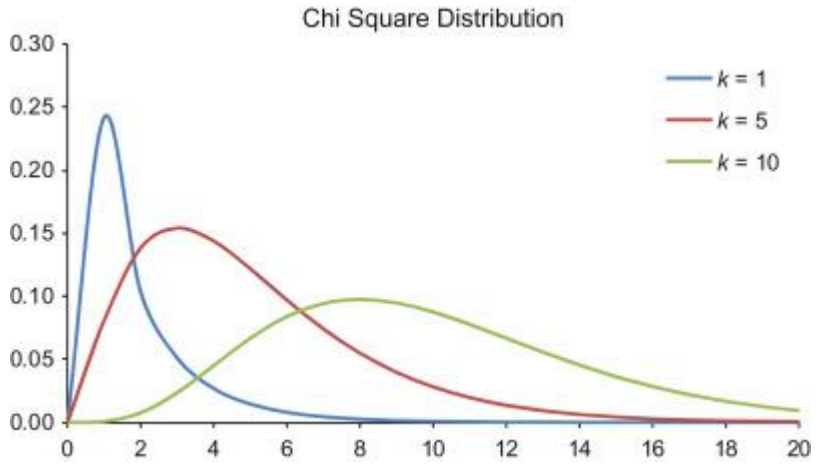

Fig (7.1): χ^2 density function for different values of K

As seen from Figure (7.1), as K increases, the distribution approaches a normal curve.

7.2 Independence tests

Consider the following simple situation: A dice is thrown 60 times. The following table shows the outcomes. The hypothesis to be tested is: the dice is uniform (that is the probability of getting any of its 6 numbers is alike). We compare in the table the observed frequency of occurrence with the hypothetical frequency.

If f_o is the value of observed frequency and f_h, the value of hypothetical frequency, then the value of χ^2 is obtained by the formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_o - f_h)^2}{f_h} \qquad (7.2)$$

In the EXCEL function CHIINV, there are two entries: *K*, the number of degrees of freedom and the level of significance $\alpha = 1 - L$.

In the present case, $K = 6 - 1 = 5$. This is since if we know the number of outcomes for any five numbers, we can get the sixth since the total number is known to be 60.

Number	1	2	3	4	5	6
Frequency (Obs.)	13	6	8	12	11	10
Frequency (Hyp.)	10	10	10	10	10	10

$$\chi^2 = \frac{(13-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(10-10)^2}{10}$$
$$= 3.4$$

For $K = 5$, the value of 3.4 is less than the critical value of 11.05 at a 0.05 significance level. **This means that there is a 95% probability that differences between observed values and predicted ones are due to chance**. We thus accept the hypothesis that the coin is uniform.

In the previous example, the hypothetical value of frequencies was easy to calculate as well as the number of degrees of freedom. When the hypothetical values are not known, the problem gets more complicated as follows:

It is claimed that the sales of a certain brand of detergent are independent of the geographic sector where it is sold. To test this hypothesis, we choose samples from three districts A, B and C, and obtain the number of customers using this brand.

The following table summarizes the results obtained. Such tables are called **Contingency tables**

Table (7.1): Observed frequencies of users in different districts

	Number of users	**Number of non-users**	**TOTAL**
A	125	75	200
B	143	107	250
C	89	61	150
Total	**357**	**243**	**600**

To test the validity of the hypothesis, we calculate the "theoretical frequency" of occurrence. This is done by dividing the total number of users and non – users by the total number of customers then multiplying this figure by the total number of customers in each district.

The ratio of users is 357/600 = 0.595 and that of non – users: 243/600 = 0.405

The hypothetical number of users in zone A should be: $200 \times 0.595 = 119$, in zone B: $250 \times 0.595 \approx 149$ and in zone C: $150 \times 0.595 = 89$

Similarly, the corresponding figures for non – users are: A: $200 \times 0.405 = 81$, B: $250 \times 0.405 = 101$ and C: $150 \times 0.405 = 61$

The following contingency table shows the actual frequency and the hypothetical frequency of users.

Table 7.2 Observed and hypothetical frequencies of users

	Observed Nº of users	**Hyp. Nº of users**	**Observed Nº of non-users**	**Hyp. Nº of non-users**	**TOTAL**
A	125	119	75	81	200
B	143	148.75	107	101.25	250
C	89	89.25	61	60.75	150
Total	**357**		**243**		**600**

In such contingency tables, K, the number of degrees of freedom is obtained by the following equation:

$$K = (C - 1) \times (R - 1) \qquad (7.3)$$

Where, *C* is the number of columns in Table 7.1 (disregarding totals) and *R* the number of rows.

In the current case, $K = (2 - 1) \times (3 - 1) = 2$

And from equation (7.2), calculation of χ^2 on EXCEL, yields $\chi^2 = 1.2975$

From the χ^2 function for $K = 2$, at a 0.05 level of significance, the value of 1.2975 is less than the critical value of 5.99

This means that the differences between the observed and the hypothetical values are most probably due to chance. The hypothesis is thus accepted.

Example 7.1

A large company owns three factories A, B and C. Each factory produces four brands of chemicals (C_1, C_2, C_3 and C_4). The following table shows the distribution of their products (in tons per year). Check the null hypothesis: The number of brands produced does not depend on a specific factory.

	C_1	**C_2**	**C_3**	**C_4**
A	16	18	15	10
B	50	8	8	10
C	30	20	10	5

Solution:

The following table summarizes the values of observed and hypothetical frequencies.

Note for example that the total of C_1 brand produced by the three factories is 96, while the total number of products is 200. So, the C_1 brand figures have to be multiplied by 96/200 = 0.48 etc…

	C_1		C_2		C_3		C_4		**Total**
	Obs.	**Hyp.**	**Obs.**	**Hyp.**	**Obs.**	**Hyp.**	**Obs.**	**Hyp.**	
A	16	28.32	18	13.57	15	9.74	10	7.38	**59**
B	50	36.48	8	17.48	8	12.54	10	9.35	**76**
C	30	31.20	20	14.95	10	10.73	5	8.13	**65**
Total	**96**		**46**		**33**		**25**		**200**

From equation (7.2), calculation of χ^2 on EXCEL, yields $\chi^2 = 25.41$

The number of degrees of freedom is $(4 - 1)\times(3 - 1) = 6$

From the χ^2 function for $K = 6$, the value of 25.41 is greater than 12.59, the value for $\alpha = 0.05$. **The hypothesis cannot therefore be accepted**.

7.3 Goodness of fit tests

The χ^2 distribution can also be used to test how good a fit of data as compared to some theoretical model. The following example explains the method used.

Example 7.2

A common empirical rule, known as Trouton's rule, states that the entropy of vaporization of a solid is 88 J/mole.K. The following table gives experimental values for the latent heat of vaporization and the boiling point of some solids. From these data test the suitability of this rule at a significance level = 0.05

Solid	**Cd**	**Na**	**Mg**	**Zn**	**KCl**	**NaCl**
ΔH_{vap}**.kJ.mol^{-1}**	100	99.2	127.5	112.5	159	166
T **K**	1038	1155	1378	1180	1680	1738

Solution:

First, we recall that: $\Delta S_{vap} = \frac{\Delta H_{vap}}{T}$

The hypothesis under test is therefore: ΔS_{vap} = 88 J.mole^{-1}.K^{-1}

So, we build a table where the observed and hypothetical values of ΔS_{vap} are represented. The number of degrees of freedom is $(2 - 1)\times(6 - 1) = 5$

Solid	$\Delta S_{vap\ obs.}$	$\Delta S_{vap\ hyp.}$
Cd	96.339	88
Na	85.887	88
Mg	92.525	88
Zn	95.339	88
KCl	94.643	88
NaCl	95.512	88

We get: $\chi^2 = 2.828$

From the χ^2 function, for $K = 5$, the value of 2.828 is less than 11.07, the value corresponding to $\alpha = 0.05$. **The hypothesis is thus accepted.**

Example 7.3

A company produces computer chips of uniform power. To check for that uniformity, eight specimens were tested for power. Check at a significance level = 0.05 whether the power can effectively be considered uniform.

105 115 107 102 103 95 110 108

Solution:

The mean value of the sample is first determined:

$\bar{x} = 105.63$

The null hypothesis to be tested is H_0: $\mu = 105.63$

The following table shows the χ^2 calculations:

P	105	115	107	102	103	95	116	108
P_{th}	105.63	105.63	105.63	105.63	105.63	105.63	105.63	105.63
$\frac{(P - P_{th})^2}{P_{th}}$	0.0038	0.8312	0.0178	0.1247	0.0655	1.0697	0.1808	0.0532

The number of degrees of freedom = 8 – 1 – 1 = 6 since one restriction was added by assuming the constant value as the mean value of sample. $\chi^2 = 2.347$

Critical value from CHIINV = 12.59

Since $2.347 < 12.59$, the hypothesis can be accepted and the distribution cann be assumed to be uniform.

Example 7.4

The following data belong to the population distribution of student marks in an exam. Check, at a 0.05 level of significance, whether these data can be fitted by a normal distribution model.

Class	0 to < 5	5 to < 10	10 to < 15	15 to < 20	20 to< 25	≤ 30
Frequency	2	15	8	13	16	3

Solution:

Calculation of mean value and standard deviation: $\mu = 15.57$, $\sigma = 6.8026$

The number of degrees of freedom must be lessened by 2 since we have added a new constraint regarding the mean and standard deviation of the normal assumption = 6 – 1 – 2 = 3

Normal simulation	$x < 5$	$x < 10$	$x < 15$	$x < 20$	$x < 25$	$x < 30$
Probability	0.06011	0.20645	0.4666	0.7425	0.91716	0.9835
Calc. cum *f*	3.42649	11.7676	26.5968	42.3253	52.2783	56.0338
Obs. cum *f*	2	17	25	38	54	57
$(f_{calc} - f_{obs})^2/f_{calc}$	0.5939	2.3266	0.09587	0.4420	0.05670	0.01666

Calculated value of χ^2:
0.5939 + 3.3266 + 0.09587 + 0.442 + 0.0567 + 0.01666 = **3.532**

Critical value from CHIINV: **7.81**

Since 3.532 < 7.81, **the hypothesis is accepted.**

7.4 The use of χ^2 values to evaluate the confidence interval for variance

χ^2 values can be used to construct a confidence interval for standard deviations or variances and alternatively test hypotheses about their values assuming normally distributed population.

For a significance level α, we first determine two values for χ^2: $\chi^2_{\frac{\alpha}{2}}$ and $\chi^2_{1-\frac{\alpha}{2}}$.

For a sample size n, let the variance of sample be s^2, then the confidence interval for variance of population σ^2 will be:

$$\frac{(n-1).s^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1).s^2}{\chi^2_{1-\frac{\alpha}{2}}} \qquad (7.4)$$

Example 7.4

A 10 – sized sample of crude oil was analyzed for sulfur content. The results are given in the following table.

Specimen Nº	1	2	3	4	5	6	7	8	9	10
% sulfur	1.56	1.8	1.22	1.78	2.11	1.95	1.45	2.05	2.0	1.85

Use this information to construct a confidence interval for the standard deviation of the whole production at a 90% confidence level.

Solution:

The standard deviation of sample is calculated as: $s = 0.286$

The values of $\chi^2_{0.05}$ and $\chi^2_{0.95}$ are obtained from the CHINV function as:

$\chi^2_{0.05}$ = 16.918 and $\chi^2_{0.95}$ = 3.325

Substituting in equation (7.4):

$$\frac{(10-1)\times 0.286^2}{16.918} < \sigma^2 < \frac{(10-1).\times 0.286^2}{3.325} \quad \text{or } 0.0435 < \sigma^2 < 0.2214$$

Hence: **$0.2086 < \sigma < 0.4705$**

7.5 The use of χ^2 values to test hypotheses about the variance

χ^2 values can also be used to test hypotheses about the variance or alternatively the standard deviation of a population along which the variable of interest is normally distributed.

For example, let a population have an unknown variance σ^2. A sample with n values is selected, and its standard deviation (s) determined. To test the null hypothesis: H_0: $\sigma^2 = k^2$, the following statistic is used, where $d.f. = n - 1$:

$$\chi^2 = \frac{(d.f.)s^2}{k^2} \qquad (7.5)$$

The acceptance region will range from $\chi^2_{1-\frac{\alpha}{2}}$ to $\chi^2_{\frac{\alpha}{2}}$ for a significance level α

If, however, the tested hypothesis is **one – tailed**, the acceptance region will be from $\chi^2_{1-\alpha}$ to χ^2_{α}

Example 7.5

The following data show the scores of a sample of 9 students (out of 20) sampled from a large population. It will be assumed that the scores are normally distributed along this population. Test the hypothesis that the standard deviation of the population is 4 at a 5% significance level.

16 12 13 8 17 3 15 18 11

Solution:

The value of s is obtained = 4.77

For α = 0.05: $\frac{\alpha}{2}$ = 0.025 and $1-\frac{\alpha}{2}$ = 0.975

From the CHINV function, the corresponding values of χ^2, for a number of degrees of freedom = 8, are: $\chi^2_{0.025}$ = 17.5 and $\chi^2_{0.975}$ = 2.17.

From equation (7.5), the value of the test statistic for the variance is:

$$\chi^2 = \frac{8\times 4.77^2}{4^2} = 11.37$$

Since this value lies within the interval (2.18; 17.5), the hypothesis is accepted at a 0.05 significance level.

7.6 Exercise problems

In all forthcoming problems take the level of significance as 0.05

(1) A dice was thrown 60 times. The following table was obtained.

1	**2**	**3**	**4**	**5**	**6**
10	14	9	11	9	7

Check the following hypothesis: The coin is uniform.

(2) The following table shows the pattern of sales of a brand of lube oils on different oil stations in four different governorates A, B, C and D. Test the hypothesis that the proportion of sales does not depend of the selling zone.

Zone	**Users**	**Non – users**
A	22	15
B	37	19
C	15	9
D	62	39

(3) The following represents the results obtained on monitoring the BOD of wastewater as function of the distance from the source of pollution:

Distance km	1	2	3	4	5	10	20	50	100
BOD mg.L^{-1}	61	53	48	44	41.5	38	36.5	36	35.5

Conduct a goodness of fit test to evaluate the possibility of using the following empirical equation to predict the BOD level as function of distance (D):

$$BOD = 33 + \frac{30}{D}$$

(4) The following equation represents an empirical relation between the 28-days compressive strength of cement mortar cubes (σ MPa) and the water to cement ratio used (x) for a fixed sand to cement ratio. Check the validity of that relation using the Chi squared test.

$\sigma = 8.5x^{-0.94}$

$\boldsymbol{x}$	0.32	0.305	0.28	0.265	0.245	0.23	0.21
$\boldsymbol{\sigma}$	24	28	31	32	33.5	37	42

(5) Wall tiles can contain 5 types of defects. The following table shows the number of defects found in a sample of 100 tiles.

Nº of defects	**0**	**1**	**2**	**3**	**4**	**5**
Nº of samples	15	26	25	15	12	7

Calculate the mean value and show that this distribution approximately conforms to a Poisson distribution.

(6) Ten specimens were tested from a wastewater stream for suspended solids content (ppm). At 0.05 significance level would you consider the distribution to be uniform?

450 485 460 425 460 470 430 425 450 475

(7) Construct a confidence interval for the standard deviation of the specific gravity of 8 crude specimens:

Specimen Nº	1	2	3	4	5	6	7	8
Sp. Gr.	0.82	0.76	0.77	0.80	0.75	0.76	0.82	0.84

(8) The following table shows the tensile strength (in MPa) of a 11 – sized samples of PE fibers drawn from a large population. Test the hypothesis that the standard deviation of population = 120

410 620 590 840 490 600 820 380 760 540 720

(9) The producer of a lube oil brand claims that its production is extremely uniform in properties. To this aim a client chooses 12 specimens in twelve days and tests them for viscosity at some fixed temperature. Verify the producer claim that the standard deviation does not exceed 2 cP.

32 21 27 24 30 26 28 22 25 27 32 30