PAPER COM 202.2

ADVANCED BUSINESS STATISTICS

Chapter 5: Non-Parametric Test

Traditional parametric testing methods are based on the assumption that data are generated from a population for which the population distribution (normal / binomial etc.) is known for which the parameters (mean, variance, etc.) are tested and the hypotheses of the problems are formulated as equalities / inequalities related to these unknown parameters. In other words parametric methods are based on the introduction of stringent assumptions, often quite unrealistic, unclear and connected with the availability of inferential method. Hence the critical values or alternatively the *p*-values can be computed according to the distribution of the test statistic under the null hypothesis, which can be derived from the assumptions related to the assumed underlying distribution of data. When the assumed distribution of the population and thereby the sample is not true, other methods, which ignore the true distribution of data, are needed. These methods are called *nonparametric* or *distribution-free method* of hypothesis testing.

Since, when the parametric assumptions hold, the nonparametric procedures are less powerful than the parametric methods and they are the only valid solution when the parametric assumptions do not hold. Nonparametric tests are in general more flexible and often more appropriate than parametric counterparts.

Some of the popular non parametric tests are:

a) Chi square test
b) Sign test
c) Run test
d) Mann-Whitney U test
e) Kolmogorov -Smirnov (K-S) test
f) Kruskal- Wallis H test etc.

Chi square test (χ^2 test): Chi square test is one of the most popular and widely used non parametric tests under test of statistical hypotheses. Chi square test is used for-

a) Test of goodness-of-fit

b) Test of independency

c) Test of homogeneity

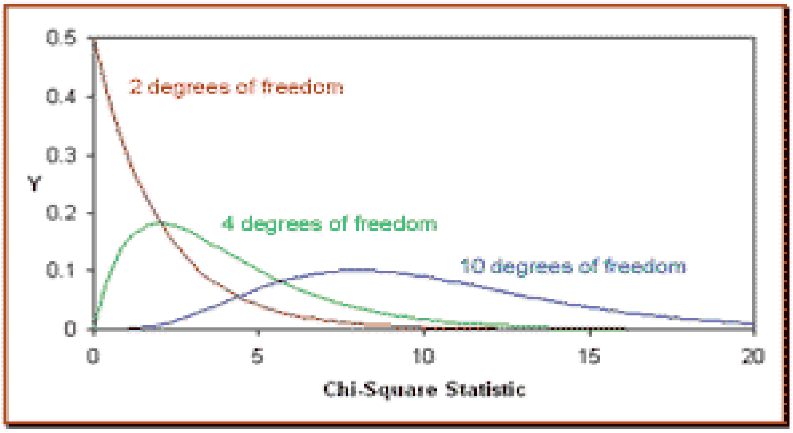
Test of goodness-of-fit:

Computation of chi square:

$$\chi^2 = \sum \frac{(observed\ frequency - expected\ frequency)^2}{expected\ frequency}$$

$$\text{i.e.} = \sum \frac{(fo - fe)^2}{fe}$$

Chi square distribution is a positively skewed distribution and the value of chi square varies from 0 to ∞. The specific distribution of chi square depends of the degrees of freedom.



For goodness-of-fit test the degree of freedom is calculated as k-1, where k = the number of classes of the observed data.

Problem 1:

A die is rolled 120 times in order to determine whether or not it is fair (unbiased). The value **1** appears 20 times, the value **2** appears 14 times, the value **3** appears 18 times, the value **4** appears 17 times, the value **5** appears 22 times, and the value **6** appears 29 times. Do the data suggest that the die is biased?

Solution 1: let us take the hypothesis that the dice is unbiased.

Therefore, our null hypothesis H0: ($P_1 = P_2 = P_3 = \ldots\ldots = P_6$)

And the alternative hypothesis H1: (at least one proportion is not equal)

Let us fix the level of significance of the test $\alpha = 0.05$

We know if we roll an unbiased dice then the expectation of getting any of the six faces is the same. Thus, the probability of getting each face on the upper part is 1/6. Hence, if we roll an unbiased dice for 120 times then the expected frequency of each face is 1/6 * 120 = 20.

Calculation of χ^2

Face	Observed frequency (fo)	Expected frequency (fe)	$(fo\text{-}fe)^2$	$(fo\text{-}fe)^2$ / fe
1	20	20	0	0
2	14	20	36	1.8
3	18	20	4	0.2
4	17	20	9	0.45
5	22	20	4	0.2
6	29	20	81	4.05
			$\chi^2 = \sum \frac{(fo-fe)^2}{fe}$	**6.7**

Here, the degrees of freedom = k -1 = 6-1 = 5

The table value of chi square for 5 degrees of freedom at 5% level of significance is 11.07.

Since our computed value of chi square (6.7) is less than the critical value (11.07), there is no reason to reject our null hypothesis at 5% level of significance.

We may therefore conclude that the dice is unbiased.

Problem 2:

A librarian wishes to determine if it is equally likely that a person will take a book out of the library each of the six days of the week the library is open (assume the library is closed on Sundays). She records the number of books signed out of the library during one week and obtains the following frequencies: Monday, 15; Tuesday, 9; Wednesday, 13; Thursday, 12; Friday, 17; and Saturday, 24. Assume no person is permitted to take out more than one book during the week. Do the data indicate there is a difference with respect to the number of books taken out on different days of the week?

Solution 2: let us take the hypothesis that the number of books taken out from the library is evenly distributed throughout the week.

Therefore, our null hypothesis H0: ($P_1 = P_2 = P_3 = \ldots\ldots = P_6$)

And the alternative hypothesis H1: (at least one proportion is not equal)

Let us fix the level of significance of the test $\alpha = 0.05$

Here, the total number of books taken from the library during the week is 90. If the books taken from the library is evenly distributed on six days of the week then the expected number of books issued on any day will be 90/6 = 15.

Calculation of χ^2

Days	Observed frequency (fo)	Expected frequency (fe)	$(fo\text{-}fe)^2$	$(fo\text{-}fe)^2$ / fe
Monday	15	15	0	0.00
Tuesday	9	15	36	2.40
Wednesday	13	15	4	0.27
Thursday	12	15	9	0.60
Friday	17	15	4	0.27
Saturday	24	15	81	5.40
			$\chi^2 = \sum \frac{(fo-fe)^2}{fe}$	**8.93**

Here, the degrees of freedom = k -1 = 6-1 = 5

The table value of chi square for 5 degrees of freedom at 5% level of significance is 11.07.

Since our computed value of chi square (8.93) is less than the critical value (11.07), there is no reason to reject our null hypothesis at 5% level of significance.

We may therefore conclude that the issue of books from the library is evenly distributed in all days throughout the week.

Exercise:

Problem 3: The proportions of blood types O, A, B and AB in the general population of a particular country are known to be in the ratio 49 : 38 : 9 : 4 respectively. A research team, investigating a small isolated community in the country, obtained the following frequencies of blood type.

Blood type	O	A	B	AB
Frequency	87	59	20	4

Test the hypothesis at 5% level that the proportions in this community do not differ significantly from those in the general population.

[Hints: $\chi^2 = \sum \frac{(fo-fe)^2}{fe} = 3.247$; H0 is accepted.]

Problem 4: Five identical coins are tossed for 320 times. The number of heads appeared at each time of tossing are recorded, which are given below:

Number of heads	0	1	2	3	4	5	Total
Frequency	7	42	112	90	57	12	320

Test at 5% level whether the coins are unbiased.

[Hints: $\chi^2 = \sum \frac{(fo-fe)^2}{fe} = 6.0$; H0 is accepted.]

Test of Independency:

Chi square test is also applied to test whether several attributes are independent. Independency test of chi square is done with the help of contingency table. Calculation process of chi square remains the same as of test for goodness of fit i.e. $\chi^2 = \sum \frac{(fo-fe)^2}{fe}$

Here, degree of freedom is calculated as df = (r-1)* (c-1), that is degrees of freedom equals to (number of rows -1) × (number of columns -1)

Contingency table (3 × 3)

	Column 1	Column 2	Column 3	**Marginal total**
Row 1	√	√	0	**RT 1**
Row 2	√	√	0	**RT 2**
Row 3	0	0	0	**RT 3**
Marginal total	**CT 1**	**CT 2**	**CT 3**	**Grand total**

√ → Values can be assigned freely

0 → Values cannot be assigned freely

Expected frequency for any cell --

= (corresponding row total × corresponding column total) / Grand total

Fro first cell it will be – (**RT 1** × **CT 1) / GT**

Problem 5: People believe that the herbs help to prevent flu. To test the belief a sample of 320 people is considered during the flu season. 120 of them had been given Herb 1, 140 had been given Herb 2 and for the rest Placebo (sugar pill) had been used. The results are presented in the following table.

	Herb 1	Herb 2	Placebo (Sugar pill)
Sick	20	30	30
Not sick	100	110	90

Test at 5% level whether the herbs are effective to control flu.

Solution 5: let us take the hypothesis that the Herbs are not effective to control flu that is the flu is independent from taking herbs.

Here the level of significance α = 0.05.

Calculation of expected frequency

	Herb 1	Herb 2	Placebo (Sugar pill)	**Marginal total**
Sick	= (80*120) / 380 = 25.25	= (80*140) / 380 = 29.5	25.25	**80**
Not sick	94.75	110.5	94.75	**300**
Marginal total	**120**	**140**	**120**	**N = 380**

Calculation of χ^2

Observed frequency (fo)	Expected frequency (fe)	$(fo-fe)^2$	$(fo-fe)^2 / fe$
20	25.25	27.563	1.092
100	94.75	27.563	0.291
30	29.5	0.250	0.008
110	110.5	0.250	0.002
30	25.25	22.563	0.894
90	94.75	22.563	0.238
		$\chi^2 = \sum \frac{(fo-fe)^2}{fe}$	**2.525**

Here the degrees of freedom = (r-1) * (c –1) = (2-1) * (3-1) = 2

The table value of chi square for 2 degrees of freedom and at 5% level of significance is 5.99.

Since our computed value of chi square (2.525) is less than the critical value, there is no reason to reject our null hypothesis at 5% level of significance.

We may therefore conclude that the attack of flu is independent from taking herbs that the herbs has no significant effect in controlling flu.

Problem 6: A public opinion poll surveyed a simple random sample of 1000 American voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table below.

	Voting Preferences			Row total
	Rep	**Dem**	**Ind**	
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

Do the men's voting preferences differ significantly from the women's preferences? Use a 0.05 level of significance.

Solution 6:

Let us take the hypothesis that voting preference is independent from gender.

H_o: Gender and voting preferences are independent.

H_a: Gender and voting preferences are not independent.

Here the level of significance of the test is 5%.

Calculation of expected frequency

	Rep	**Dem**	**Ind**	**Marginal total**
Male	180	180	40 (balance)	**400**
Female	270 (balance)	270 (balance)	60 (balance)	**600**
Marginal total	**450**	**450**	**100**	**1000**

Calculation of χ^2

Observed frequency (fo)	Expected frequency (fe)	$(fo-fe)^2$	$(fo-fe)^2$ / fe
200	180	400.00	2.22
250	270	400.00	1.48
150	180	900.00	5.00
300	270	900.00	3.33
50	40	100.00	2.50
50	60	100.00	1.67
		$\chi^2 = \sum \frac{(fo-fe)^2}{fe}$	**16.20**

Here the degrees of freedom = (r-1) * (c –1) = (2-1) * (3-1) = 2

The table value of chi square for 2 degrees of freedom and at 5% level of significance is 5.99.

Since our computed value of chi square (16.2) is more than the critical value, there is no reason to accept our null hypothesis at 5% level of significance.

We may therefore conclude that the preference of voting is not independent from gender.

Exercise:

Problem 7: A manger of a big factory wants to see if geographical region is associated with ownership of a computer of its employees. The manager surveys 100 employees and the data breaks down as follows:

Location	**Computer**	**No computer**	**Marginal total**
North East	12	14	26
South West	21	18	39
Mid West	17	18	35
Marginal total	50	50	100

Apply chi square test to examine whether ownership of computer is independent from geographical location of the employees. Use $\alpha = 0.05$