COUNTINUOUS MEDICAL EDUCATION
pISSN09763325 | eISSN22296816
Open Access Article (CC BY-SA)
www.njcmindia.com
DOI: 10.55489/njcm.130920222200

Qualitative - Binary, Nominal and Ordinal Data Analysis in Medical Science

Swati Patel[1]

[1]Surat Municipal Institute of Medical Education & Research, Surat

ABSTRACT

The outcome of any medical research is belonged to the human beings. The correct application of statistical test has its paramount importance. This article provides the details of categorical data analysis test with example and with its interpretation. It is included when to use Chi-square test (2*2 & R*C), Yate`s Correction, Fisher Exact test, Mc Nemartest manually as well as using Open Epi and R codes.

Keywords: Qualitative analysis, nominal data, ordinal data

INTRODUCTION

In study the categorical /Binary data represented by frequency table whereas to know the Association between either binary or categorical data Chi- Square test is used. The chi square analysis determines association between categorical responses between two or more independent groups. Chi square tests can only be used on actual numbers and not on percentages, proportions and means. There are two type of chi- square test (i) Good ness of fit chi-square test (II) Chi-square test of Independence. The Chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not. It is often used to evaluate whether sample data is representative of the full population. Chi-square test of independence is a statistical hypothesis test used to determine whether two Binary / Categorical variables are likely to be related/associated or not. It can be applying when two qualitative variables i.e binary/categorical (nominal or ordinal) are independent.

In medical research when researcher mainly interested to know the association between two binary or dichotomous variable, different statistical test applies for that 2* 2 table to know the association between them according to the frequency of each cell.

This article mainly describes the appropriate usage of chi-square test, Yate`s correction of chi-square test, Fisher Exact test and Mc-Nemar test.

Chi – Square goodness of fit

The chi-square test of goodness-of-fit use when you have one nominal variable within two or more categories (such as <18 ,18-30, 31-45,...) You compare the observed counts of observations in each category with the expected counts, which you calculate using some kind of theoretical expectation (such as a 1:1 Male: Female ratio or a 1:2:1 ratio in a genetic cross). In other words when you draw random samples and to know the proportion of outcome is equal or not to check it chi- square good ness of fit is useful.

If the expected number of observations in any category is too small, the chi-square test may give inaccurate results, and you should use an exact test.

Pearson's Chi- Square test (Chi- Square test of independence)

Chi-square test applied for actual frequencies/numbers as cross tabulated in a contingency table. It is not appropriate to use for percentages/proportion. It can be applied on 2* 2 contin-

How to cite this article: Patel S. Qualitative - Binary, Nominal and Ordinal Data Analysis in Medical Science. Natl J Community Med 2022;13(9):663-668. DOI: 10.55489/njcm.130920222200

Financial Support: None declared | **Conflict of Interest:** None declared | **Date of Submission**: 27-06-2022 **Date of Acceptance**: 28-08-2022 **Date of Publication**: 30-09-2022

Correspondence: Dr. Swati Patel (E-mail: Swati84patel@gmail.com)

gency table or R*C table (R is number of rows and C is Number of Columns).

When a researcher interested to analyze his/her data using a chi-square test of independence, he/she need to make sure that the data should follow the following assumptions. If the data doesn't meet the assumption of chi-square test, you can`t apply the chi-square test for independence.

Assumption 1: Two variables should be measured at binary or **ordinal** or **nominal level**.

Assumption 2: Your two variables should consist of **two or more categorical**, **independent groups**. Example independent variables that meet this criterion include Outcome of Patients (2 groups: Death, Survive), Socioeconomic level (e.g., 3groups: Upper, Middle, lower), Pain Score (4 group: No Pain, Mild, Moderate, Severe), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist).

Assumption: 3 Cells in the contingency table should be mutually exclusive.

It's assumed that individuals can only belong to one cell in the contingency table. That is, cells in the table are mutually exclusive — an individual can not belong to more than one cell.

We can verify that this assumption is met by checking that no individual has been counted in more than one cell.

Assuming each individual in the dataset was only surveyed once, this assumption should be met because it's not possible for an individual to be, say, a Male Republican & Female Democrat simultaneously.

Assumption 4: Expected value of cells should be 5 or greater in at least 80% of cells.

It's assumed that the expected value of cells in the contingency table should be 5 or greater in at least 80% of cells and that no cell should have an expected value less than 1.

Methods:

Researcher can enter the data in Excel. It has also facility to prepare cross table. Using the pivot table in excel cross table can be prepared. [2,3]

Excel → Insert → Pivot table 

1. The above table.1 full filed the all assumptions of Pearson's chi- square test, two categorical/binary independent variables are given, Cells in the contingency table are mutually exclusive
2. Calculate expected value of each cell. (no more than 20% of cells expected value is < 5)
   Expected cell = row total (rij)*Colum total(cij)/N
3. Calculate $\chi^2_{cal} = \frac{\sum(o_{ij}-e_{ij})^2}{e_{ij}}$
   o_{ij} = observed frequency of i^{th} row and j^{th} Column
   e_{ij} = expected frequency of i^{th} row and j^{th} Column

Post hoc test

When chi-squared test is significant, the next step is to perform a post hoc test to find out which cells from the contingency table are different from their expected values. Without any prior knowledge of each cell, we are interested in testing all cells in a contingency table at once. Three test statistics are often calculated for each cell: Raw Residual (RawR), Standardized Residual (StdR), and Adjusted Residual (AdjR). The larger these residuals are, the greater the contribution of these residuals to the overall chi-squared test. (5)

When the chi-square test of a more than 2×2 is significant (and sometimes when it isn't), it is desirable to investigate the data further. MacDonald and Gardner (2000) use simulated data to test several post-hoc tests for a test of independence, and they found that pairwise comparisons with Bonferroni Correction *P* values work well.

Raw Data application

Researcher planned study to assess the prevalence and associated factors of depression in breast cancer patients. Total 152 patients included with Brest cancer, depression measure using PHQ -2 Scales. The researcher hypothesised that

Null Hypothesis: There is no association between depression and stages of cancer.

Alternative Hypothesis: There is an association between depression and stages of cancer.

How to prepare table using raw data

Table 1: Details of Depression and Stages of Cancer in Brest cancer patients

Stages of Cancer	Depression	
	Present	Absent
I	8	18
II	13	30
III	26	41
IV	11	5

Table 1.1: Details of observed and expected value

Stages of Cancer		Depression		Total
		Yes	No	
I	Observed	8(o11)	18(o12)	26
	expected	12(e11)	16(e12)	
II	Observed	13(o21)	30(o22)	43
	expected	16(e21)	27(e22)	
III	Observed	26(o31)	41(o32)	67
	expected	26(e31)	41(e32)	
IV	Observed	11(o41)	5(o42)	16
	expected	6(e41)	10(e42)	
Total		58	94	152

$$\chi^2_{cal} = \frac{(o11-e11)^2}{e11} + \cdots\ldots + \frac{(o42-e42)^2}{e42}$$

$$= \frac{(8-12)^2}{12} + \frac{(18-16)^2}{16} + \cdots\ldots\ldots + \frac{(5-10)^2}{10}$$

$$= 8.104$$

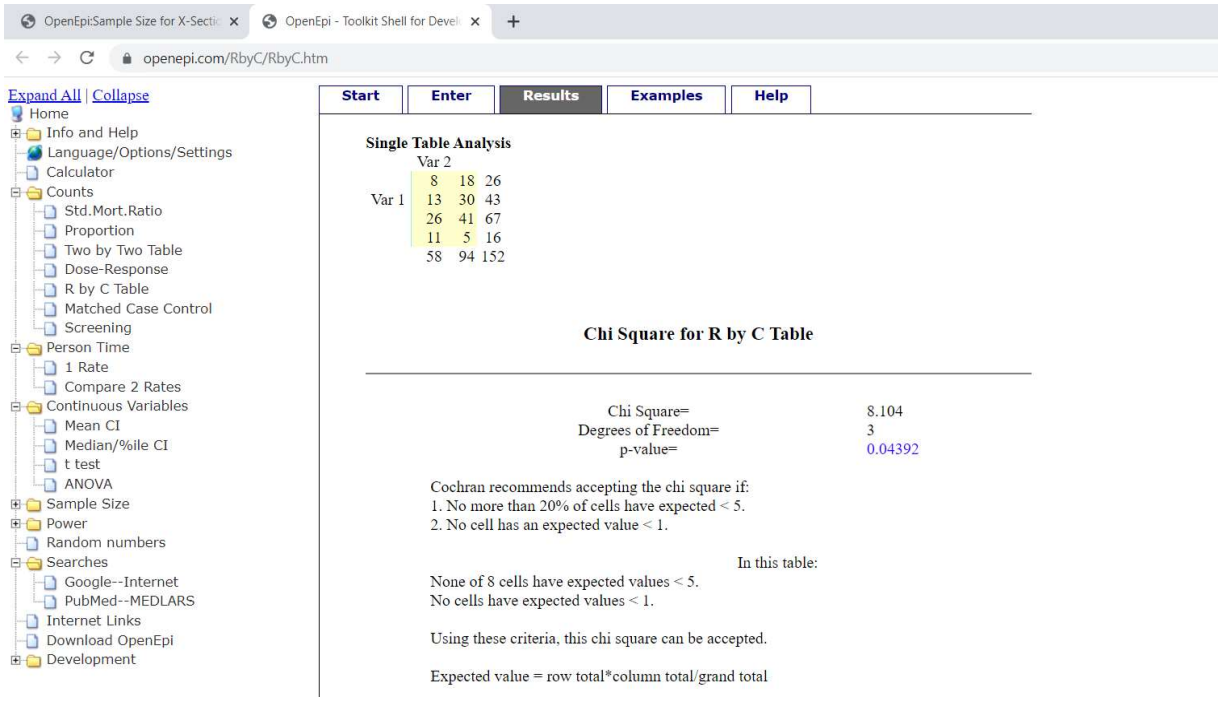
4. Compare χ^2_{cal} value with tabulated value according to degree of freedom

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

5. Conclusion: - $\chi^2_{cal} > \chi^2_{0.05,3}$, reject the null hypothesis. i.e there is association in between stages of cancer and depression among the breast cancer patients.

Manually Calculation part is very laborious for chi-square test, especially for R*C table. Now days many software's/ online statistical calculators are available using this anyone can do calculation very easily but important thing is that researcher need to know the assumption of it, once you prepare the cross table (R*C), to calculate chi square value you can use OPEN EPI & EPI TOOLS software, which are free of cost.

Figure 1: chi-Square test using Open EPI



1. https://www.openepi.com/SampleSize/SSCohort.htm use this website.
2. Click on R*C, you will get new window.
3. Click on enter in new window, it will ask your number of rows and column of cross table for chi- square calculation.
4. Click on calculate.
5. In Result you will get three values i.c chi square =8.104, df =3 and P-value = 0.04392.
6. This software will also mention the assumption of chi- square test (i.e % of expected cell`s value < 5)

Interpretation

Here 4×2 table has a chi-square value is 8.104 with 3 degrees of freedom, *P* value 0.04392, which only indicating that there is an association between Depression and different stages of cancer among the Brest Cancer patients. But to know which stages of cancer is significantly associated with the depression to know that pairwise comparison researcher needs to apply Post hoc test.

Post hoc chi-square: -

There are six possible pairwise comparisons, so you can do a 2×2 chi-square test for each one. Post hoc Chi-square computed p-values for each cell in this contingency table of this example. All Pairs P value is > 0.008 (Adjusted Bonferroni level of significance 0.05/6 =0.008). No pairwise association has been observed between stages of cancer and presence of depression using R software.

2* 2 contingency table data analysis

When in Study researcher has 2*2 contingency table is easy to calculate chi-square than more than 2*2 table. According to the objective of study researcher can apply different test for 2*2 table.

Chi-square test –

For 2 * 2 cross table assumptions are same as more than 2*2 chi-square test. When 2*2 cross table is given to know the association between two binary variables, it will be applying when all cell`s frequency is >5.

Case study

During the Pandemic of COVID19 a team of doctors recorded that the severity has been observed among the patients who admitted in COVID ward with hypertension. Using this data, they assumed that severity of COVID 19 is related or associated with hyper-

tension. Total 276 patients recorded for this study, divided them two parts hypertensive and non-hypertensive according to severity and non-severity.

Table 1.2 Details of COVID 19 admitted patient's severity wise

	Hypertensive	Non-Hypertensive	Total
Sever	52	89	141
Non-Sever	29	106	135
Total	81	195	276

Null Hypothesis: Severity of disease is not associated with hypertension.

Alternative hypothesis: Severity of disease is associated with hypertension

Calculation of Chi square:

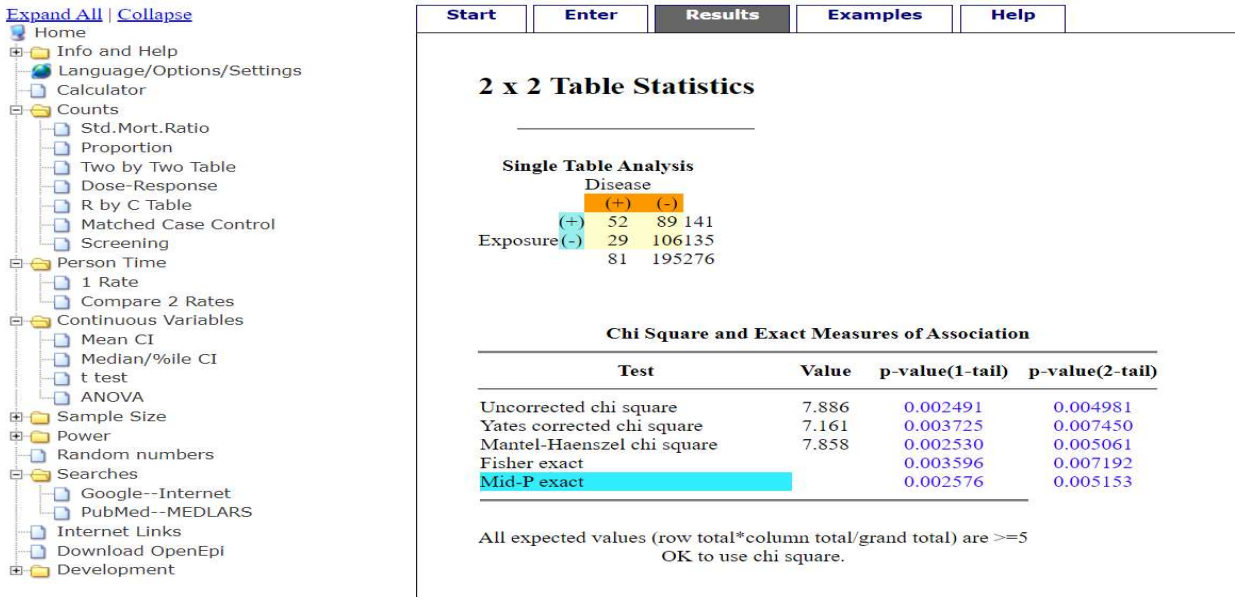
$$\chi^2_{cal} = \frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$= \frac{((52*106)-(89*29))^2}{(52+89)(29+106)(52+29)(89+106)} = 7.886$$

Interpretation: Here in the given 2*2 cross table $\chi^2_{cal} > \chi^2_{0.05,1}$ (3.84 will get from statistical tables) , accept the alternative hypothesis. i.e., Severity of disease is associated with hypertension.

Using Open EPI software, the calculation will become easy.

Figure1.2 Calculation using OPEN EPI



Interpretation – The given 2×2 table has a chi-square value is 7.886 with 2 degrees of freedom, *P* value(2-tail) 0.004981, which only indicating that severity of disease is associated with hypertension and 2.136(OR) times more chance of severity among hypertensive patients of COVID 19 as compared to non-hypertensive COVID 19 Patients.

Yate Correction –

Yate`s correction is the method of modification of chi-square for only for 2'*2 contingency table, to reduce the error in approximation, Frank Yates, an English statistician, suggested a correction for continuity which adjusts the formula for Pearson's chi-square test by subtracting 0.5 from the difference between each observed value and its expected value in a 2 × 2 contingency table (Yates, 1934). This reduces the chi-square value obtained and thus increases its p-value. (6)

The effect of Yates' correction is to prevent overestimation of statistical significance for small data. This formula is chiefly used when at least one cell of the table has an expected count smaller than 5. Unfortunately, Yates' correction may tend to overcorrect. This can result in an overly conservative result that fails to reject the null hypothesis when it should. So it is suggested that Yates' correction is unnecessary even with quite low sample sizes (Sokal and Rohlf, 1981), such as total sample sizes less than or equal to 20.

Fisher Exact test

Fisher's exact test assesses the null hypothesis of independence applying hypergeometric distribution of the numbers in the cells of the table instead of approximation. Fisher's exact test is practically useful only in analysis of small samples but actually it is valid for all sample sizes. While the chi-squared test relies on an approximation, Fisher's exact test is one of exact tests. Especially when more than 20% of cells have expected frequencies < 5, we need to use Fisher's exact test because applying approximation method is inadequate. Manually calculation of Fisher Exact test is very difficult but today many packages

provide the calculation of Fisher's exact test for 2 × 2 contingency tables but not for bigger contingency tables with more rows or columns. For example, the SPSS statistical package and OPEN EPI automatically provides an analytical result of Fisher's exact test as well as chi-squared test only for 2 × 2 contingency tables. But When Fisher's exact test of bigger contingency tables, we can use web pages providing such analyses. SAS and R can be also use for to calculate Fisher Exact test for more than 2 rows and 2 columns.

A researcher studied to know the association between age group and severity of CIVID19 hospitalised cases. This study included total 99 patients of COVID 19.

Table 1.3: Details of Age group and Severity of COVID 19 Case

Age	Mild	Moderate	Severe
<18 yrs	6	4	0
18-30 yrs	12	7	3
31-50yrs	12	6	5
>50 yrs	4	23	17

There are 4 rows and 3 columns in the above table, given data has 25% of 12 cells have expected values < 5 which doesn`t meet the Cochran's criteria, so to know the proportions are independent or not Instead of Chi- Square test Fisher Exact test should be use.

Interpretation: Here P< 0.05, so the proportion of age group and severity of COVID19 are Associate (Not Independent). To know the age wise severity comparison post hoc pairwise comparison.

For 4*3 contingency table if P< 0.05 (by Fisher Exact test), you could do a 2×3 Fisher's exact test for each of these pairwise comparisons, but there are 6 possible pairs, so you need to do correct multiple comparison. One way to do this is with a modification of the Bonferroni-corrected pairwise technique suggested by MacDonald and Gardner (2000), replacing Fisher's exact test for the chi-square test they used. You do a Fisher's exact test on each of the 6 possible pairwise comparison (i.e <18 yrs & 18-30 Yrs, <1 8yrs &31-50 Yrs ,..... etc) then apply the Bonferroni Correction multiple tests. With 6 pairwise comparisons, the *P* value must be less than 0.05/6, or 0.008, to be significant at the *P*<0.05 level.

Age group	P- Value adjusted Fisher
<18 Yrs & 18-30 Yrs	0.8480
<18 Yrs & 31-50 Yrs	0.6510
<18 Yrs & >50 Yrs	0.0017*
18-30 Yrs & 31-50 Yrs	0.8480
18-30 Yrs & > 50 Yrs	0.0017*
31-50 Yrs & >50 Yrs	0.017

Interpretation: - In this pairwise comparison, two pairs of age group <18 yrs. & >50 yrs & 18-30 yrs & > 50 yrs are significantly associated with severity of COVID19.

Mc Nemarc test

It is not showing the association between two variables. It is non-parametric (distribution-free) test assesses if a statistically significant change in proportions have occurred on a dichotomous attribute at two time points on the same population. It is applied using only a 2×2 contingency table with the dichotomous variable at time 1 and time 2. In medical research, if a researcher wants to determine whether or not a particular therapy has an effect on a disease (e.g., yes vs. no), then a count of the individuals is recorded (as + and – sign, or 0 and 1) in a table before and after being given the Theraphy. Then, McNemar's test is applied to make statistical decisions as to whether or not a drug has an effect on the disease.

Example

A researcher planed study to know the effect of two different treatments on for breast cancer after mastectomy. The two group of treatment group compared on the basis of other prognostics factors. Total 621 patients included in this study. Patients are assigned to pairs matched on age (within 5 yrs.) and clinical condition

Outcome of Treatment B	Outcome of Treatment A	
	Survive for 5 yrs	Die within 5 yrs
Survive for 5 yrs	510 (a)	16(b)
Die within 5 yrs	5 (c)	90(d)

Ho: No Difference in proportion of survival/(5 yrs) by treatment A and B.

$$\chi^2_{cal} \ or \ Q = \frac{((b-c)-1)^2}{(b+c)} = \frac{(16-5)-1)^2}{16+5} = 4.76$$

In order to test if the treatment is helpful, we use only the number discordant pairs of twins, b and c, since the other pairs of twins tell us nothing about whether the treatment is helpful or not.

$\chi^2_{cal} \ or \ Q = \frac{(b-c)^2}{(b+c)}$ or

which for large samples is distributed like a chi-squared distribution with 1 degree of freedom. A closer approximation to the chi-squared distributing uses a continuity correction.

$$\chi^2_{cal} \ or \ Q = \frac{((b-c)-1)^2}{(b+c)}$$

CONCLUSION

Calculated value (4.76) > $\chi^2_{0.05,1}$ (3.84) ,reject the null hypothesis. i.e treatment give different result from subjects of matched pairs, then the treatment

A subjects of pair is significantly more likely to survive for 5 yrs than the treatment subjects.'

REFERENCES

1. Kim H. Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative dentistry & endodontics*, *42*(2), 152–155. https://doi.org/10.5395/rde.2017.42.2.152
2. Shan G, Gerstenberger S (2017) Fisher's exact approach for post hoc analysis of a chi-squared test. PLoS ONE 12(12): e0188709. https://doi.org/10.1371/journal.pone.0188709
3. Yates F. Contingency tables involving small numbers and the χ2 test. *Supplement to the Journal of the Royal Statistical Society*. 1934;1(2):217–235. [Google Scholar
4. https://support.microsoft.com/en-us/office/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576
5. https://www.excel-easy.com/data-analysis/pivot-tables.html
6. https://cran.r-project.org/web/packages/exact2x2/vignettes/exactMcNemar.pdf
7. http://www.biostathandbook.com/fishers.html

Annexure

Name of Test	R code
Post hoc Chi-square test	R1 = c(8,18) R2 = c(13,30) R3 = c(26,41) R4 = c(11,5) rows = 4 x = matrix(c(R1,R2,R3,R4), nrow=rows, byrow=TRUE) rownames(x) = c("One", "Two", "Three", "Four") colnames(Matriz) = c("Yes", "No") x library(rcompanion) pairwiseNominalIndependence(Matriz, fisher = FALSE, gtest = FALSE, chisq = TRUE, method = "fdr")

Fisher Exact test	x <- matrix(c(6,4,0,12,7,3,12,6,5,4,23,17), byrow = TRUE, nrow = 4, ncol = 3);
	fisher.test(x);
	
	Outcome:
	p-value = 8.491e-05 or 0.00008491
	
	library(rcompanion)
	pairwiseNominalIndependence(x,
	fisher = TRUE,
	gtest = FALSE,
	chisq = FALSE,
	method = "fdr")

McNemar's Chi-squared test	> x <- matrix(c(510,5,16,90) , nrow =2); > x [,1] [,2] [1,] 510 16 [2,] 5 90 > mcnemar.test(x) data: x McNemar's chi-squared = 4.7619, df = 1, p-value = 0.029