Testing differences in proportions

Murray J Fisher
RN, ITU Cert., DipAppSc, BHSc, MHPEd, PhD

Senior Lecturer and Director Preregistration Programs
Sydney Nursing School (MO2)
University of Sydney NSW 2006

Phone: +61 2 9351 0587
Fax: +61 2 9351 0654
Email: murray.fisher@sydney.edu.au

Andrea P Marshall

Sesqui Senior Lecturer (Critical Care) and Director Postgraduate Programs
Sydney Nursing School (MO2)
University of Sydney NSW 2006

Marion Mitchell

Senior Research Fellow Critical Care
Griffith University & Princess Alexandra Hospital
Address: Nurse Practice Development Unit
Princess Alexandra Hospital
Ipswich Road
Woolloongabba, Qld 4102

Abstract

This paper is the sixth in a series of statistics articles recently published by *Australian Critical Care*. In this paper we explore the most commonly used statistical tests to compare groups of data at the nominal level of measurement. The chosen statistical tests are the chi-square test, chi-square test for goodness of fit, chi-square test for independence, Fisher's exact test, McNemar's test and the use of confidence intervals for proportions. Examples of how to use and interpret the tests are provided.

Introduction

This article presents the most commonly used statistical tests to compare groups of data at the nominal level of measurement. Nominal (or categorical) level of measurement is the sorting of cases into one of several categories (for example, types of religion), where the measure of dispersion is based on the count or frequency of cases in each category of measurement. Concepts around levels of measurement are explained in the second article of this series *Understanding Descriptive Statistics*[1].

The number of cases in each category for a given sample is known as the frequency distribution. A common way of presenting frequency distributions for nominal data is in a table, sometimes referred to as a contingency table or cross-tabulation. An example is depicted in Table 1, which shows the frequency distribution of the incidence of diarrhoea in an intensive care unit (ICU) population over a 12 month period during which time an intervention was introduced[2].

Specific methods of inferential statistics are required to determine differences between samples in nominal level measurement. In the example depicted in Table 1, we would be determining the difference in frequency of diarrhoea in patients before implementation of a bowel management protocol with those after the protocol was implemented. The tests of significance for nominal data vary depending on the nature of the chosen measurements for the variables. Table 2 presents the most commonly used tests for comparing two groups using nominal measurement level.

Chi-square test

The chi-square test compares the observed frequency distribution (f_o) for each category of the scale with the expected frequency distribution (f_e) of the null hypothesis. When using a chi-square test it is assumed that there has been random sampling; that 80% of the cells have an expected frequency of greater than five; that no cell has an observed frequency of 0; and, that a large sample is used, as small sample sizes lead to a small expected frequency which causes large chi-square values[3]. A limitation of the chi- square test is that it is sensitive to either very small or large samples. Quantifying the minimum sample is difficult as it is dependent on the number of cells in the crosstab. A sample is considered too small when the above assumptions are not met. When these assumptions are not met the chi-square cannot be meaningfully interpreted.[4] The chance of finding a significant difference between samples is greater with larger samples. If you double the sample size, the chi-square statistic will double due to the large sample size rather than a strong pattern of dependence between the variables[5].

When these assumptions are violated the results may lead to erroneous interpretation of the data; that is, the results may be considered statistically significant when in reality they may not be statistically significant. When samples are small and the assumptions for the chi-square are violated, the Fisher's exact test could be used[6].

The formula for calculating the chi-square statistic is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Where χ^2 is equal to the sum of the squared difference between the observed and the expected frequencies divided by the expected frequency for each cell.

The concept of degrees of freedom (*df*) is important and is a mathematical limitation that needs to be factored in when calculating an estimate of one statistic from an estimate of another. The *df* are used in conjunction with the table of critical values for chi- square. The *df* for a chi- square test is calculated with the following equation:

$df = (R-1) \times (C - 1)$

Where:

R equals the number of rows

C equals the number of columns[3]

Chi-square test for goodness of fit

The chi-square test for goodness of fit is used for a single population and is a test used when you have one categorical variable. This test determines how well the frequency distribution from that sample fits the model distribution. Consider the data provided in the contingency table (Table 3) which reports the frequency of patients who developed diarrhoea for three different wards within a hospital.

The chi-square test for goodness of fit determines difference by comparing the observed frequency distribution with the frequency distribution of the null hypothesis. The null hypothesis is the expected frequency distribution of all wards is the same. That is, approximately 33.3 patients in each ward would be expected to have developed diarrhoea.

$$\chi^2 = \frac{(-3.3)^2}{33.3} + \frac{(-8.3)^2}{33.3} + \frac{(6.7)^2}{33.3}$$

$$= 0.33 + 2.09 + 1.36$$

$$= 3.78$$

The critical value for χ^2 needs to be determined. First determine the *df* (see textbox) and determine the level of significance (often set at 0.05) (please refer to the fourth article of this series *Statistical and clinical significance, and how to use confidence intervals to help interpret both*[8]). Referring to a table listing the critical values of chi-square (available in most statistics texts) and using the calculated degrees of freedom (*df* = 1) and level of significance of 0.05, the critical value for χ^2 is 3.84. The computed chi-square value of 3.78 is lower than the critical value of 3.84, therefore the null hypothesis is not accepted and we conclude that there is not a statistical difference in the distribution of the frequency of patients who developed diarrhoea for the three different wards.

Chi-square test for independence

The chi-square test for independence is also used for a single population but where there are two categorical variables. The test examines if there is a relationship between the two variables for the one sample.[3]Consider the observed frequency distribution on the difference in the incidence of diarrhoea before and after the implementation of a bowel management protocol (Table 1).

The contingency table (Table 1) demonstrates that 36.41% of the pre-intervention sample and 22.74% of the post-intervention sample experienced diarrhoea. In order to determine whether there is a statistical difference between the pre-intervention and post-intervention groups, the chi-squared test of independence is used as these are two independent samples. The chi-square statistic compares the observed frequency distribution (f_o), for example the frequencies that are depicted in Table 1, with the expected frequency distribution of the null hypothesis (f_e). The null hypothesis expresses the expected frequency for each category if there is no statistical difference between categories (see previous publication in this series[8] for further information on hypothesis testing).

In this case the null hypothesis is that there is no statistical difference between the number of patients with diarrhoea in the pre-intervention sample as compared to those in the post-intervention sample. To calculate the frequency distribution of the null hypothesis (f_e) the following formula is used:

$$f_e = \frac{f_c f_r}{n}$$

Where, f_c is the frequency total for the column, f_r is the frequency total for the row and n is the total sample size. To calculate the expected frequency for each cell you simply substitute the observed frequency with the calculated expected frequency using the formula.

Refer to Table 4 and the calculation below to determine the expected number of patients with diarrhoea in the pre-intervention sample for the null hypothesis. In this case the f_e would be:

$$f_e = \frac{379(201)}{656}$$

$$= 116.13$$

The expected frequency distribution for the null hypothesis in this example would be calculated as depicted in Table 5.

At a glance it would appear that in this example there is a difference between frequency observed (f_o) and the expected frequency (f_e). Table 6 presents the difference between the observed and expected frequency for each cell.

To calculate whether there is a statistical difference the chi-square formula is used.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Where χ^2 is equal to the sum of the squared difference between the observed and the expected frequency divided by the expected frequency for each cell. In this case the chi-square statistic is equal to:

$$\chi^2 = \frac{(21.87)^2}{116.13} + \frac{(-21.87)^2}{84.87} + \frac{(-21.87)^2}{262.87} + \frac{(21.87)^2}{192.13}$$

$$= 4.12 + 5.63 + 1.82 + 2.49$$

$$= 14.06$$

The critical value for χ^2 needs to be determined; first calculate the *df* (see textbox) and determine the level of significance. Referring to a table listing the critical values of chi-square and using the calculated *df* (1) and level of significance of 0.05, the critical value for

χ^2 is 3.84. The chi-square value of 14.06 calculated above exceeds that of the critical value, therefore the null hypothesis is rejected and we conclude that there is a statistical difference between the number of patients with diarrhoea in the pre-intervention sample as compared to those in the post-intervention sample.

The Statistical Package for the Social Sciences (SPSS version 18) was used to examine this sample of patients with or without diarrhea. The reported SPSS output confirms that there was a statistical difference in the incidence of diarrhoea between the pre-intervention and post-intervention samples χ^2 (1, n=656) =14.06, p<0.0001. In the original study Ferrie and East[2] identified a statistical difference in the incidence of diarrhoea between the two samples (p<0.0001), however this claim could have been strengthened by reporting the χ^2 statistic.

Fisher's exact test

The Fisher's exact test is used in cases where there are cells with an expected frequency (f_e) less than 5 and/or with small sample sizes, as the Fisher's exact test has no sample size restriction[6]. The method of calculation of the Fisher's exact test is different to the chi-square statistic and is calculated by determining the probability of getting the observed frequency distribution by establishing and comparing to all other possible distributions where the column and row totals remain the same as the observed distribution. In this case the null hypothesis indicates that all the cells would be close to equal. The calculation of the Fisher's exact test is complex and is not available in all statistical packages but can be performed using the Statistical Package for Social Sciences.

McNemar's test

The McNemar test compares dependent (paired or matched) samples in terms of a dichotomous variable.[4] It is the best test for comparing dichotomous variables with two dependent sample studies as opposed to the chi-square test which examines nominal level variables with two samples that are independent of each other.[4] A dichotomous variable has only two possible outcomes, for example yes or no and it results in a binomial distribution.[9] The McNemar test may be used for pretest-posttest design or in time series data where the same sample is tested at least in two points in time. The main assumption of the McNemar text is that the data comes from two samples that are matched. This can either be as a paired sample or a before/after sample. The McNemar test is a non parametric test and thus assumes that the data are not normally distributed.[4]

Consider the following fictitious two by two contingency table (Table 7) which shows the incidence of diarrhoea at two time periods in a sample (n=200).The McNemar test is similar to the chi-squared test in that it examines the difference between expected and observed cell frequencies. The following formula is used to calculate the McNemar test. There is one *df* which is derived from the following equation: (rows – 1)x (columns – 1) = 1.

$$\chi^2_M = \frac{(N_a - N_d - 1)^2}{N_a + N_d}$$

Where:

N_a equals the frequency of observed responses – see Cell marked A in Table 7

N_d equals the frequency of observed responses – see Cell D marked in Table 7

In the above example the χ^2_M is equal to:

$$\chi^2_M = \frac{(40 - 33 - 1)^2}{40 + 33}$$

$$= \frac{36}{73}$$

$$= 0.04$$

With one degree of freedom and level of significance of .05, based on the chi-square distribution the critical value for χ^2_M is 3.84. The McNemar chi-square value is less than that of the critical value, therefore the Null Hypothesis (H_o) is retained and we conclude that there is not a statistical difference in the incidence of diarrhoea between the two time periods.

Confidence intervals for differences in proportion

Confidence intervals (CI) are now being reported along with p values in clinical studies and their use has been described in an earlier paper in this series[5]. Calculation of CI is based on the assumption that the variable is normally distributed in the population and is dependent on the level of measurement and therefore the statistical test used.

The formula to construct the CI for a proportion will be available in most statistics textbooks[10]. There are also computer programmes that perform these tasks for researchers.

The statistical programme will calculate the CI and the researcher selects the level of confidence (for example, 95%). Below is an example of how CIs for nominal data may be used in determining clinical significance. In this sample the proportional difference of those with diarrhoea before the intervention compared to those with diarrhoea after the intervention is 13% (36% - 23% - Table 1). We will now calculate the CI around this sample result.

The equation for an approximate 95% confidence interval for the difference between two population proportions ($p_1 - p_2$) based on two independent samples of size n1 and n2 with sample proportions $\hat{p}_1$ and $\hat{p}_2$ is given by the following equation:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example

Using the data provided in Table 1, we will calculate the 95% CI using the equation above where $\hat{p}_1$= 36%; $\hat{p}_2$ = 23%; n_1= 379 and n_2 = 277. The figure of 1.96 indicates we are computing a CI of 95%.

$$\text{CI} = (0.36 - 0.23) \pm 1.96 \sqrt{\frac{0.36(1-0.36)}{379} + \frac{0.23(1-0.23)}{277}}$$

$$= 0.13 \pm 1.96 \sqrt{0.000607316 + 0.00063335}$$

$$= 0.13 \pm 1.96 \text{ x } 0.035223089$$

$$= 0.13 \pm 0.069037$$

upper limit = 0.199 and lower limit = 0.060

These results indicate that the lower limit of a 95% CI is 6% and the upper limit is 20% with the sample proportion difference at 13%. Note that CIs may not be symmetrical around the sample proportion, it just happens to be in this instance. With a 0.05 level of significance, there is a significant result with $p<0.0001$ (as reported earlier in the paper) and the CI provides additional information as it gives a range of where the population proportion is likely to lie. Patients with the intervention are somewhere between 6% and 20% more likely to experience no diarrhea than those without the intervention. The clinical significance and research conclusions should be drawn from the individual context for the study[3].

Conclusion

This paper has provided an introduction to the statistical tests commonly used to test differences in proportions for nominal level data. Chi-square tests are commonly used in health care research and where sample sizes are small, the Fisher's Exact Test may be used. If the data from dependent, paired samples are binomial, then the McNemar's test may be more appropriate. As with many other statistical tests, assessment of the critical values, p values and CI may assist in the reader determining clinical and statistical significance of the results.

Table 1: Incidence of diarrhoea in intensive care following a bowel management protocol

	Pre-intervention **n (%)**	**Post-intervention** n (%)	**Total**
Patients with diarrhoea	138 (36)	63 (23)	201
Patients without diarrhoea	241 (77)	214 (64)	455
Total	379	277	656

Table 2 The tests of significance for nominal data

Sample types	**Test of Significance**
One-sample case	Chi-square goodness of fit
Two or more independent samples	Chi-square test for independence
Two dependent (paired) samples	McNemar test for binomial distributions
Small samples	Fisher’s exact test

Table 3 Frequency of diarrhea in patients admitted to three wards

Ward A	Ward B	Ward C	Total
30	25	40	95

Table 4 Calculating the expected number of patients with diarrhoea in the pre-intervention sample for the null hypothesis

	Pre-intervention	**Post-intervention**	**Total**
Patients with diarrhoea	? (fe)	63	201 (fr)
Patients without diarrhoea	241	214	455
Total	379 (fc)	277	656 (n)

Table 5 Expected frequency distribution for patients with and without diarrhoea at pre-intervention and post-intervention time periods

	Pre-intervention	Post-intervention	Total
Patients with diarrhoea	116.13	84.87	201
Patients without diarrhoea	262.87	192.13	455
Total	379	277	656

Table 6 Difference between frequency observed and expected frequency

	Pre-intervention	Post-intervention
Patients with diarrhoea	21.87	-21.87
Patients without diarrhoea	-21.87	21.87

Table 7 Contingency table of diarrhoea at two time periods (with cells named)

		Time 1		
		No	**Yes**	**Total**
Time 2	**Yes**	Cell A = 40	Cell B = 67	**107** **93**
	No	Cell C= 60	Cell D = 33	
	Total	**100**	**100**	**200**

Text box 1

In statistics the degree of freedom is the number of values in the calculation of a statistic that are free to vary. To calculate the degree of freedom for a chi-square test you count the number of rows and subtract 1 and multiply with the number of columns with 1 subtracted. So for Table 2,
DF = (number of rows – 1) x (number of columns – 1) or (2-1) x (2-1) = 1.

References

1. Fisher M, Marshall AP. Understanding descriptive statistics. *Aust CritCare* 2009; **22**: 93-97
2. Ferrie S, East V. Managing diarrhoea in intensive care. *Aust Crit Care* 2007; **20**: 7-13
3. Corty EW. *Using and interpreting statistics: A Practical text for the Health, Behavioral, and Social Sciences*. St. Louis: Mosby Elsevier; 2007
4. Argyrous G. *Statistics for Social Research*. Melbourne: Macmillan Education Australia Pty. Ltd; 1996
5. Smithson, M.J. Statistics with Confidence: An Introduction for Psychologists Sage, Canberra. 2000
6. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall; 1996
7. Fethney J. Statistical and clinical significance, and how to use confidence intervals to help interpret both. *Aust Crit Care* 2010; 23: 93-97
8. Periera SMC, Leslie G. Hypothesis testing. *Aust Crit Care* 2009; 22:187-191
9. Polit DF, Beck CT. *Nursing Research: Generating and Assessing Evidence for Nursing Practice,* 8th ed. St. Louis: Mosby Elsevier; 2008
10. Carlin JB, Doyle LW. Statistics for Clinicians 6: Comparison of means and proportions using confidence intervals. *J Paediatr Child Health* 2001; 37: 583-586.