Early Human Development xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Early Human Development

journal homepage: www.elsevier.com/locate/earlhumdev

WASP (Write a Scientific Paper) using Excel – 10: Contingency tables

ARTICLE INFO

Keywords:
Software
Computers
Statistics
Biostatistics

ABSTRACT

Contingency tables may be required to perform chi-test analyses. This provides pointers as to how to do this in Microsoft Excel and explains how to set up methods to calculate confidence intervals for proportions, including proportions with zero numerators.

1. Introduction

A contingency table (a cross tabulation or crosstab) is a matrix, a table that depicts the frequency distribution of variables, typically as counts, i.e. the distribution of one variable in rows and another in columns. These are useful to study the association/s between two variables. This sort of data does not make (and indeed does not need to make) any assumptions about the data being normally distributed. The commonest form is a 2 by 2 matrix (Table 1).

The dependent variable (the outcome) is conventionally placed at the top and the independent variable (predictor/exposure/test) is placed at the side. Thus, the outcome of interest is typically the top left cell (a).

Most outcomes are usually negative events or failures such as death, disablement, disease or discomfort. Examples of predictors or exposures are smoking, alcohol, sex, blood group, hypertension, diabetes, or treatments.

Some outcomes are not dichotomous (binary) but can be transformed into a binary format using cut-offs. A typical example is body mass index (BMI), which is a continuous variable. Sets of BMI values can be analysed in contingency tables by using cut-off thresholds, such as World Health Organisation values for obesity. Values are thus categorised as obese/not obese and can be input into a 2 by 2 matrix. Tables such as these may be extended to 2 by n.

It is important to note that the comparison for two sets of predictor/exposure variables are "with" or "without". The denominator (the total) is never invoked (and is unnecessary) as this is not a rate calculation.

2. Chi function

Excel's inherent chi testing functions are rudimentary. The CHITEST function requires not only observed but also expected values. For example, Table 2 depicts a small database that comprises monthly live births, by sex, for the state of California for the year 2013.

A pivot table easily extracts male and female totals as has been shown in a previous paper in this series [1]. The ratio of male to total births is 0.5124. Were the expected ratio 0.5 (equal number of male and female births), the expected values are calculated on the right hand side along with the formulas depicting the calculations. The observed ratio departs significantly from the expected ratio ($p < < 0.0001$) as shown, along with the Excel chi function.

However, it is known that this ratio approximates 0.515. The expected values calculation is also shown on the right as well as the chi function, which is still statistically significant, but less so ($p = 0.0003$).

3. Excel limitations

The following are not natively available in Excel.

In 2 × 2 tables, a "Yates continuity correction" is conventionally applied so as to prevent overestimation of statistical significance for small datasets.

There are additionally three common variants of the chi test:

- A Mantel-Haenszel test for the analysis of stratified or matched categorical data.

https://doi.org/10.1016/j.earlhumdev.2018.02.020

0378-3782/ © 2018 Elsevier B.V. All rights reserved.

Early Human Development xxx (xxxx) xxx–xxx

Table 1
Template for a 2 by 2 table.

		Outcome	
		Present	Absent
predictor/	Present	a	b
exposure	Absent	c	d

Table 2
Monthly live births, by sex, for California, 2013 and chi calculations.

	A	B	C	D	E	F	G	H
1	**Month**	**Sex**	**Births**					
2	January	Female	20,243		**Row Labels**	**Sum of Births**		
3	January	Male	21,008		Male	253,495		
4	February	Female	17,939		Female	241,210		
5	February	Male	18,772		**Grand Total**	**494,705**		
6	March	Female	19,889					
7	March	Male	20,815					
8	April	Female	18,960			Observed	Expected	
9	April	Male	19,834		Male	253,495	247,352.5	=F11/2
10	May	Female	19,707		Female	241,210	247,352.5	=F11/2
11	May	Male	20,809			494,705	494,705	
12	June	Female	18,902		Male/Total	0.5124	0.5000	
13	June	Male	20,104		p	2.5854E-68		
14	July	Female	20,761			=CHITEST(F9:F10,G9:G10)		
15	July	Male	21,931					
16	August	Female	21,818					
17	August	Male	23,070			Observed	Expected	
18	September	Female	21,537		Male	253,495	254,773.075	=F20*0.515
19	September	Male	22,267		Female	241,210	239,931.925	=F20*0.485
20	October	Female	21,399			494,705	494,705	
21	October	Male	22,254		Male/Total	0.5124	0.5150	
22	November	Female	19,600		p	0.000277041		
23	November	Male	21,073			=CHITEST(F18:F19,G18:G19)		
24	December	Female	20,455					
25	December	Male	21,558					

- A chi test for trend to ascertain whether a ratio is changing, e.g. with time and
- McNemar's test for paired observation.

Furthermore, it is considered inappropriate to use the chi test for 2 × 2 tables if the totals in the table are < 20 or if the totals are in the range 20–40 and the smallest expected value is < 5. In this case, Fisher's exact test is considered the better choice.

4. Excel add-ins

Data for such tests can usually be extracted using pivot tables from a database, as shown above. It is laborious to attempt to carry out these analyses in Excel and for this reason, this paper will not demonstrate further how to construct more functional chi tests in Excel. However, several add-ins are available which can do all of these tests. A popular (and free) Excel add-in called Bio-Med-Stat performs all of these functions [2].

5. Confidence intervals for proportions

Confidence intervals are derived from the standard error which is in turn calculated from the standard deviation. However, contingency tables assume non-normality and these calculations are therefore not an inherent part of the calculation of statistical significance.

However, the confidence interval of a proportion is still important as it gives researchers the 95% limits of probability of where the true population proportion lies.

There are two common ways to calculate this. A good approximation is the following which first calculates the standard error (p is the proportion):

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

From this point, the confidence interval is conventionally calculated thus, where z is conventionally given as 1.96 (95%)

$$CI = p \pm (z \times SE)$$

However, the so-called binomial methods should be used for the calculation of confidence intervals for greater precision when the proportion is ≤0.3 or ≥0.7 (i.e. $0.3 \le p \le 0.7$) [3]. These are calculated thus (where q = 1-p):

Download English Version:

https://daneshyari.com/en/article/8777642

Download Persian Version:

https://daneshyari.com/article/8777642

Daneshyari.com