Pearson Chi-Square Goodness-of-Fit (test_pearson_gof)

a.k.a. Relative Risk

Author: P. Stikker
Website: https://peterStatistics.com
YouTube: https://www.youtube.com/stikpet
Version: 0.1 (2023-01-06)

Introduction

The *test_pearson_gof* function (and *test_pearson_gof_arr* in VBA) performs a Pearson chi-square Goodness-of-Fit test. The test could be used to compare the proportions from different categories. The null-hypothesis is roughly that the proportions are all the same. If the p-value is too small (usually below 0.05) the assumption is rejected, indicating that at least two categories will have a different proportion in the population.

This document contains the details on how to use the functions, and formulas used in them.

1 About the Function

1.1 Input parameters:

- **data**
  The data to be used. Note for Python this needs to be a pandas data series.

- *Optional parameters*
  - **expCount** (default is none)
    A table with two columns. One with the categories and another with the expected counts. In Pandas this needs to be a dataframe.

  - **cc** (default is none)
    which (if any) continuity correction to use. Either
    - "none": no correction
    - "yates": Yates
    - "pearson": E.S. Pearson
    - "williams": Williams

  - **out** (default is "pvalue") – only applies to VBA *test_pearson_gof*
    Choice what to show as result. Either:
    - "pvalue": show the p-value (significance)
    - "df": the degrees of freedom
    - "statistic": show the test-statistic used

1.2 Output:

- The **test-statistic (chi-square value)**, **degrees of freedom**, **p-value** and **test** used. Except for the non-array version in VBA (Excel) which will only show the requested Alternative Ratio.

- The array version in VBA (*test_pearson_gof_arr*) requires **two rows** and **four columns**.

1.3 Dependencies

- **Excel**
  None.
  You can run the **test_pearson_gof_addHelp** macro so that the function will be available with some help in the 'User Defined' category in the functions overview.

- **Python**
  The following additional libraries will have to be installed:
  - *pandas*
    the data input needs to be a pandas data series, and the output is also a pandas dataframe.

- **R**
  No other libraries required.

2 Examples

2.1 Excel

	A	B	C	D	E	F	G
1	Marital		Expected counts				
2	MARRIED		MARRIED	5			
3	DIVORCED		DIVORCED	5			
4	MARRIED		NEVER MARRIED	5			
5	SEPARATED		SEPARATED	5			
6	DIVORCED						
7	NEVER MARRIED				E10 =	=ts_pearson_gof($A$2:$A$20;;$D10;E$9)	
8	DIVORCED						
9	DIVORCED			cc	statistic	df	pvalue
10	NEVER MARRIED			none	3,105263	3	0,375679
11	MARRIED			yates	1,842105	3	0,605816
12	MARRIED			pearson	2,941828	3	0,400681
13	MARRIED			williams	2,97479	3	0,395528
14	SEPARATED						
15	DIVORCED			0,375679	=ts_pearson_gof(A2:A20;C2:D5)		
16	NEVER MARRIED						
17	NEVER MARRIED			statistic	df	p-value	test
18	DIVORCED			3,105263	3	0,375679	Pearson chi-square test of goodness-of-fit
19	DIVORCED						
20	MARRIED			D12:G13	=ts_pearson_gof_arr(A2:A20)		

2.2 Python

```
[286]: data = pd.DataFrame(["MARRIED", "DIVORCED", "MARRIED", "SEPARATED", "DIVORCED",
                          "NEVER MARRIED", "DIVORCED", "DIVORCED", "NEVER MARRIED",
                          "MARRIED", "MARRIED", "MARRIED", "SEPARATED", "DIVORCED",
                          "NEVER MARRIED", "NEVER MARRIED", "DIVORCED", "DIVORCED", "MARRIED"],
                         columns=["marital"])
```

```
[287]: ts_pearson_gof(data)
```

[287]:

	statistic	df	p-value	test
0	3.105263	3	0.375679	Pearson chi-square test of goodness-of-fit

```
[288]: eCounts = pd.DataFrame({'category' : ["MARRIED", "DIVORCED", "NEVER MARRIED", "SEPARATED"],
                             'count' : [5,5,5,5]})
       ts_pearson_gof(data, eCounts)
```

[288]:

	statistic	df	p-value	test
0	3.105263	3	0.375679	Pearson chi-square test of goodness-of-fit

```
[289]: ts_pearson_gof(data, cc="pearson")
```

[289]:

	statistic	df	p-value	test
0	2.941828	3	0.400681	Pearson chi-square test of goodness-of-fit, with E. Pearson continuity correction

2.3 R

```
> data <- c("MARRIED", "DIVORCED", "MARRIED", "SEPARATED", "DIVORCED",
+           "NEVER MARRIED", "DIVORCED", "DIVORCED", "NEVER MARRIED",
+           "MARRIED", "MARRIED", "MARRIED", "SEPARATED", "DIVORCED",
+           "NEVER MARRIED", "NEVER MARRIED", "DIVORCED", "DIVORCED", "MARRIED")
> eCounts = data.frame(c("MARRIED", "DIVORCED", "NEVER MARRIED", "SEPARATED"), c(5,5,5,5))
> ts_pearson_gof(data)
    chiVal df      pVal                                   testUsed
1 3.105263  3 0.3756787 Pearson chi-square test of goodness-of-fit
> ts_pearson_gof(data, cc="yates")
    chiVal df      pVal                                                                      testUsed
1 1.842105  3 0.6058155 Pearson chi-square test of goodness-of-fit , with Yates continuity correction
> ts_pearson_gof(data, eCounts)
    chiVal df      pVal                                   testUsed
1 3.105263  3 0.3756787 Pearson chi-square test of goodness-of-fit
> 
```

3 Details of Calculations

3.1 The Original Test

The Pearson chi-square test uses:

$$\chi^2_{P.GoF} = \sum_{i=1}^{k} \frac{(F_i - E_i)^2}{E_i}$$

$$df = k - 1$$

$$sig. = 1 - \chi^2(\chi^2_{P.GoF}, df)$$

Note that if the expectation about the population, is that all categories have the same frequency, then:

$$E_i = \frac{n}{k}$$

$$n = \sum_{i=1}^{k} F_i$$

Symbols used:

- k the number of categories
- F_i the (absolute) frequency of category i
- E_i the expected frequency of category i
- n the sample size, i.e. the sum of all frequencies
- $\chi^2(\dots)$ the chi-square cumulative density function

3.2 Yates Continuity Correction

This correction is usually only recommended if the degrees of freedom is two. For a goodness-of-fit test this means only if you have two categories.

$$\chi^2_{P-Y.GoF} = \sum_{i=1}^{k} \frac{(|F_i - E_i| - 0.5)^2}{E_i}$$

3.3 E.S. Pearson correction

$$\chi^2_{P-EP.GoF} = \frac{n-1}{n} \times \chi^2_{P.GoF}$$

3.4 Williams correction

$$\chi^2_{P-W.GoF} = \frac{\chi^2_{P.GoF}}{q}$$

With:

$$q = 1 + \frac{k^2 - 1}{6 \times n \times df}$$

If $df = k - 1$ (which usually is the case with a GoF test, except if you have an intrinsic null hypothesis), the formula can be simplified to:

$$q = 1 + \frac{k + 1}{6 \times n}$$

4 Sources

Pearson described this test in an article in *Philosophical Magazine Series 5* (K. Pearson, 1900).

Yates describes this for a 2x2 table:

> tribution. This is equivalent to computing the values of χ^2 for deviations half a unit less than the true deviations, 8 successes, for example, being reckoned as $7\frac{1}{2}$, 2 as $2\frac{1}{2}$. This correction may be styled the *correction for continuity*, and the resultant value of χ denoted by χ'.

(Yates, 1934, p. 222)

The Pearson correction is found as:

> **and $m+n = N$.* It is seen that the ratio d/s_d is identical with the ratio u of equation (22), except for a factor $\sqrt{[(N-1)/N]}$ which is unimportant in large samples. Thus the classical test is practically identical with that suggested in paras. 40–42 above, though the two tests are differently derived.**

(E. S. Pearson, 1947, p. 157)

The Williams correction is from Williams (1976)

> $q = 1 + \frac{1}{6\nu n}$(sum of reciprocals of expected cell frequencies
>
> − sums of expectations of marginal frequencies in the numerators of the maximum likelihood estimators
>
> + sums of expectations of marginal frequencies in the denominators of the maximum likelihood estimators).
>
> In general q is a function of the expected frequencies. To determine a numerical value for q these expected frequencies must in practice be replaced by their maximum likelihood estimators.
>
> A much easier alternative is to use the minimum value q_{min} of q given by
>
> $$q_{min} = 1 + \phi(a^2, b^2, \ldots)/(6\nu n),$$
>
> where $\phi(a, b, \ldots)$ is the deviance degrees of freedom ν expressed as a function of the factor levels $a, b, \ldots$. The difference between q and q_{min} will often be small, and the use of $q = q_{min}$

(Williams, 1976, p. 36)

The formula used is adopted from McDonald (2014).

References

McDonald, J. H. (2014, December). *Small numbers in chi-square and G–tests*. Handbook of Biological Statistics. http://www.biostathandbook.com/small.html

Pearson, E. S. (1947). The choice of statistical tests illustrated on the Interpretation of data classed in a 2 × 2 table. *Biometrika*, *34*(1/2), 139–167. https://doi.org/10.2307/2332518

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Williams, D. A. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika*, *63*(1), 33–37. https://doi.org/10.2307/2335081

Yates, F. (1934). Contingency tables involving small numbers and the chi square test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217–235. https://doi.org/10.2307/2983604