An Insightful Alternative for Calculation of the Pearson Chi-squared Test Statistic

Duane K. Allen
Lean Six-Sigma Black Belt Consultant
302 Massachusetts Avenue, Riverside, CA 92507

Abstract
Pearson's classical formula for the chi-squared test statistic is applicable to the observed (*o*) and expected (*e*) frequencies of categorical data. This poster presentation develops a general Pearson's chi-squared test statistic formula by introducing a parameter *q*. Proper selection of *q* can improve the contrast of the values of the formula's summation terms, which leads to improved visual exploration of the categorical data.

Key Words: Contingency table, goodness-of-fit, testing for independence

Introduction

The Pearson chi-squared test statistic is often used as a tool for the analysis of categorical data. The statistic supports goodness-of-fit tests and testing for independence. Pearson (1900) developed the classical formula

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where

- X^2 is the Pearson chi-squared test statistic,
- *k* is, depending on context, the number of categories, cells, or summation terms,
- *n* is the total number of observations,
- o_i is the observed frequency of the *i*th category, $o_i \geq 0$, and
- e_i is the expected frequency of the *i*th category, $e_i > 0$.

The total number of the expectations equals the total number of observations. That is, $n = \sum o_i = \sum e_i$.

This presentation develops a general formula for the Pearson chi-squared test statistic by introducing a real parameter *q*.

$$X^2 = \sum_{i=1}^{k} \left(\frac{o_i}{e_i} - 1 \right) \left(\frac{o_i}{e_i} - q \right) e_i$$

The parameter *q* scales the values of the individual summation terms of the test statistic while not affecting the value of the test statistic.

Derivation of a General Pearson Chi-squared Test Statistic

Where $n = \sum` o_i = \sum` e_i$, then

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \ldots = \sum_{i=1}^{k} \frac{o_i^2}{e_i} - n$$

$$= \sum_{i=1}^{k} \frac{o_i^2}{e_i} + qn - n - qn = \sum_{i=1}^{k} \frac{o_i^2}{e_i} + (q-1)\sum_{i=1}^{k} o_i - q\sum_{i=1}^{k} e_i$$

$$\vdots$$

$$= \sum_{i=1}^{k} \left(\frac{o_i}{e_i} - 1\right)\left(\frac{o_i}{e_i} + q\right) e_i$$

Cases q = -1, 0, and 1

While any real value may be assigned to the parameter q, this presentation focuses on the values of q being -1, 0, and 1.

$$X^2 = \begin{cases} \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} & q = -1 \\ \sum_{i=1}^{k} o_i \left(\frac{o_i}{e_i} - 1\right) & q = 0 \\ \sum_{i=1}^{k} \left(\frac{o_i^2}{e_i} - e_i\right) & q = 1 \end{cases}$$

For q = -1, the general equation becomes Pearson's classical formula.
For q = 0, the general equation provides insight on how the test statistic X^2 value will change with respect to a changes in o_i.
For q = 1, figure 1 shows that each summation term increases monotonically from $-e_i$ as o_i/e_i increases from zero. Also for q = 1, the summation terms closely approximate the corresponding terms of the G-test, especially when $\frac{1}{2} \le o_i/e_i \le 2$.

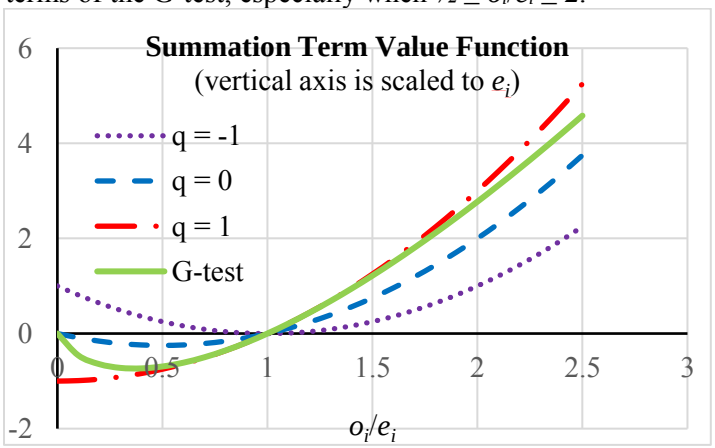


Figure 1. Summation term value functions for several parameter q values

Example with q = -1

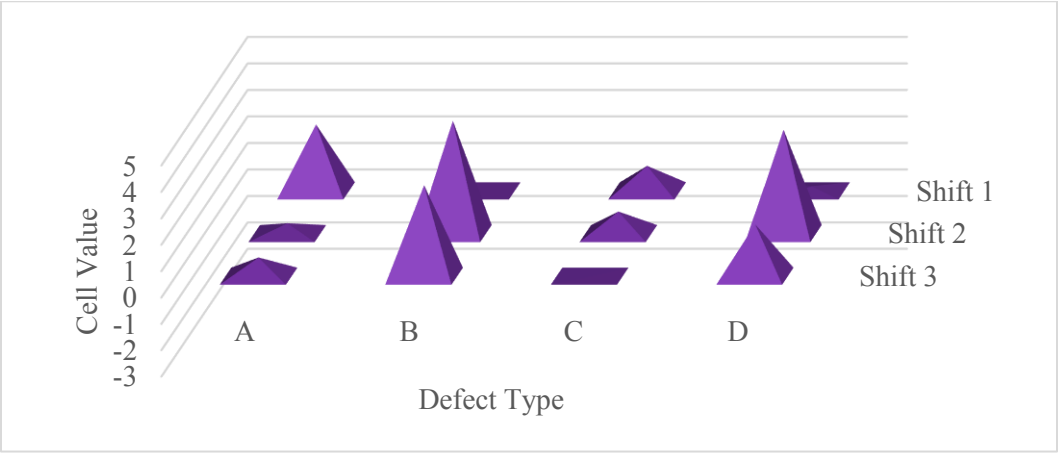
Using the data of the example contained in section 7.4.5 of the *NIST/SEMATECH e-Handbook of Statistical Methods*, the summation terms of the general Pearson chi-squared test statistic with q = -1 can be compared with the summation terms with q = 1.

The example has frequency observations for the four defect types that occur during the three production shifts. In the following table, the expected frequencies are in the parentheses that are beside the observed frequencies for each of the 12 categories of production shift (1, 2, and 3) and defect type (A, B, C, and D).

Data	Defect Type				
Shift	A	B	C	D	Total
1	15 (22.51)	21 (20.99)	45 (38.94)	13 (11.56)	94
2	26 (22.99)	31 (21.44)	34 (39.77)	5 (11.81)	96
3	33 (28.50)	17 (26.57)	49 (49.29)	20 (14.63)	119
Total	74	69	128	38	309

Applying the classical Pearson chi-squared formula, that is, the general formula with q = -1 yields the following contingency table and summation term plot.

Data	Defect Type					
Shift	A	B	C	D	x	p
1	2.5063	0.0000	0.9436	0.1794	3.6293	
2	0.3940	4.2662	0.8363	3.9234	9.4199	
3	0.7111	3.4486	0.0018	1.9673	6.1288	
x^2	3.6114	7.7147	1.7817	6.0702	19.1780	0.0038

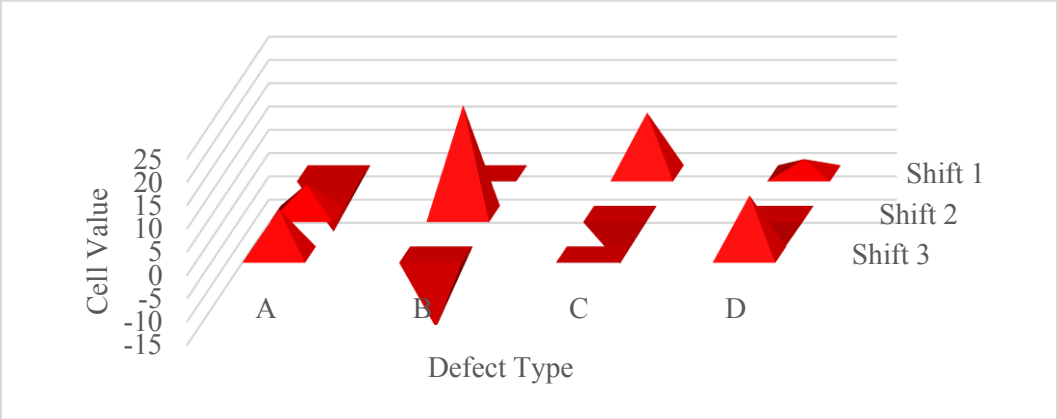


Pearson's classical formula (q = -1) identifies that the data of categories A1, B2, B3, D2, and D3 provide 84% of the test statistic X^2 value of 19.178. With the six degrees of freedom, the p-value of the test statistic is 0.00387.

Example with q = 1

Calculation of the Pearson chi-squared test statistic by the general formula with q = 1 also yields a test statistic X^2 value of 19.178 and a p-value of 0.00387. The 12 summation term values differ from those made using Pearson's classical formula. Negative and greater positive values occur in the following table and summation term plot.

Data	Defect Type					
Shift	A	B	C	D	x^2	p
1	-12.5164	0.0194	13.0666	3.0597	3.6293	
2	6.4134	23.3924	-10.6977	-9.6882	9.4199	
3	9.7143	-15.6970	-0.5872	12.6987	6.1288	
x^2	3.6114	7.7147	1.7817	6.0702	19.1780	0.0038



The alternative formula yields a graphic that identifies that efforts to reduce defects should focus on categories A3, B2, C1, and D3, which have observed defect frequencies higher than expected. The negative summation terms of A1 and B3 may indicate best practices or conditions that minimize defects.

Acknowledgements

The author thanks Dr. Analisa Flores and Ms. Jill E. Smith of University of California Riverside and Prof. Bayo Lawal of Kwara State University for comments and suggestions.

References

1. *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 10/30/2013, Section 7.4.5
2. Pearson (1900), Pearson, Karl, *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, (5) 50, 157-175