Austrian Journal of Statistics
2023, Volume 52, 1–15.
http://www.ajs.or.at/
doi:10.17713/ajs.v52i5.1607

On the P-value for Members of the Cressie-Read Family of Divergence Statistics

Eric J. Beh[1,2] **Ting-Wu Wang**[3]
[1]University of Wollongong, Australia,
[2]Stellenbosch University, South Africa
[3]University of Newcastle, Australia

Abstract

A few years ago a paper appeared in this Journal that proposed a closed-form approximation of the p-value for Pearson's chi-squared statistic. Extensive empirical and simulation studies were performed and it was shown that the approximation provides very accurate p-values when compared with what we considered to be the "true" p-value (obtained using a base R function). It is important to note, however, that Pearson's chi-squared statistic is a special case of the Cressie-Read family of divergence statistics as is, for example, the log-likelihood ratio statistic, the Freeman-Tukey statistic and the Cressie-Read statistic. Therefore, this paper adapts the previously published closed form approximation of the p-value by demonstrating its applicability to any member of the Cressie-Read family of divergence statistics. We also give two further closed form approximations and assess their accuracy by analysing three contingency tables of varying sample size, degrees of freedom and statistical significance of the association.

Keywords: Pearson's chi-squared statistic, Freeman-Tukey statistic, likelihood ratio statistic, Cressie-Read divergence statistic, Hoaglin's approximation, p-value approximation.

1. Introduction

Historically, calculating the quantile of any random variable has been a computationally labourious task. In particular, doing so for a chi-squared random variable was the focus of many during the first half of the 20th century and much of the effort went into producing tables for specific values of a chosen level of significance, α. One may consider, for example, Pearson (1922), Fisher (1928), Thompson (1941), Merrington (1941), Aroian (1943), Goldberg and Levine (1946) and Hald and Sinkbaek (1950) for discussions of such excellent contributions. In addition to tabulating quantiles, attention was also given to deriving simple formulae for producing such values; see, for example, Wilson and Hilferty (1931), Heyworth (1976) and Hoaglin (1977). Additional contributions focused on approximating the p-value for a chi-squared random variable are, to various degrees, computationally intensive, including those of Elderton (1902), Russell and Lal (1969), Khamis and Rudert (1965), Terrell (1984) and Lin (1988); many of these require knowing the quantile of the chi-squared or standard normal distribution.

With the ever increasing use of computer technology over the decades, researchers have had at their finger-tips various packages, such as those in the R programming environment, that will calculate quantiles and p-values without the need to understand exactly how these calculations are performed. As such, with the exception of the contributions mentioned above, very little attention has been given to the matter of providing simple solutions that provide excellent quantile approximations. Much less attention has been given to developing closed-form solutions designed to precisely approximate the p-value of a chi-squared random variable. However, recently Beh (2018) derived a simple closed-form solution that yields very accurate approximations when compared to the p-value calculated using the `pchisq()` function in R which makes use of the algorithm of Ding (1992). The approximation given by Beh (2018) is derived from Hoaglin (1977) who studied the relationship between χ^2_α, the degrees of freedom, v, of the test statistic (X^2) and the level of significance, α.

The approximation to the p-value described by Beh (2018) was for a Pearson chi-squared statistic, X^2, given its degrees of freedom, v. However, Pearson's statistic is only one of many that follow a chi-squared random variable. For example, four additional common measures of association include the log-likelihood ratio statistic (Wilks 1938), the Freeman-Tukey statistic (Freeman and Tukey 1950), the modified chi-squared statistic (Neyman 1940, 1949) and the modified log-likelihood ratio statistic (Kullback 1959). These, and other, statistics are all special cases of the Cressie-Read family of divergence statistics (Cressie and Read 1984) which we denote as $CR(\delta)$; changes in the value of δ lead to the various special cases and we shall be confining δ to lie within the interval $[-2, 1]$. Therefore, rather than considering only Pearson's chi-squared statistic to approximate its p-value, one can instead consider any member of the family of chi-squared statistics that belong to this family. Hence, this paper will outline how the p-value approximation of Beh (2018) can be applied, and extended, to any member of the Cressie-Read family of divergence statistics. To do so, this paper consists of three further sections. In Section 2 we revisit the closed form approximation of the p-value for a Pearson chi-squared statistic described by Beh (2018) (Section 2.1). We then provide an overview of the Cressie-Read family of divergence statistics including five of the most common special cases that come from it (Section 2.2). The p-value approximation of Beh (2018) is then shown to be applicable to any member of this family and two further approximations are described (Section 2.3).

A demonstration of the precision of the three approximations described in Section 2.3 is made in Section 3 for $\delta = -2, -1, -0.5, 0$ and 1. Rather than repeating the simulation study undertaken by Beh (2018), we instead take a more empirical approach by studying their application to three contingency tables. The first table is of size 2×2 and is a revisitation of the hydronephrosis data of Chu, Jacobs, Schwen, and Schneck (2013) where the sample size is quite small ($n = 51$). The second table is an artificial 5×4 contingency table of moderate sample size ($n = 193$) used by Greenacre (1984) to demonstrate the features and application of correspondence analysis. For this contingency table, the p-value of Pearson's chi-squared statistic and other members of the Cressie-Read family of divergence statistics, is larger than the nominal 0.05 commonly used to assess the statistical significance of the association between the variables. The third table we consider comes from Maxwell (1961) and is a 5×4 contingency table of moderate size ($n = 222$) but where the p-value is rather small (< 0.001 for most values of δ). Some final comments will be left for Section 4.

2. The Cressie-Read divergence statistic

2.1. An overview of approximating the p-value

Consider an $I \times J$ two-way contingency table, $\mathbf{N}$, where the (i, j)th cell entry has a frequency of n_{ij} for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. Let the grand total of $\mathbf{N}$ be n and let the matrix of relative frequencies be $\mathbf{P}$ so that its (i, j)th cell entry is $p_{ij} = n_{ij}/n$ where

$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$. Define the ith row marginal proportion by $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$. Similarly, define the jth column marginal proportion as $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$.

To determine whether there exists a statistically significant association between the row and column variables of $\mathbf{N}$, one may calculate any number of measures. The most popular measure is Pearson's chi-squared statistic. When there are v degrees of freedom the statistic is defined as

$$X^2 = n \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)^2}{p_{i\bullet}p_{\bullet j}} \tag{1}$$

while its p-value can be calculated by

$$P_0\left(\chi^2_{\alpha,v} > X^2\right) = \frac{1}{\Gamma\left(v/2\right)2^{v/2}} \int_{X^2}^{\infty} e^{-u/2} u^{(v/2)-1} du \tag{2}$$

where $\Gamma\left(\bullet\right)$ on the denominator is the gamma function. Computing (2) is computationally difficult because of the intractable nature of the integral. To overcome this issue, Beh (2018) provided a closed-form approximation of (2) that is based on the following relationship. Hoaglin (1977, eqs (4.3) & (4.4)) showed that for a given level of significance, α, the upper-tail quantile of the chi-squared distribution, $\chi^2_{\alpha,v}$ can be approximated by his "Fit E"

$$\chi^2_{\alpha,v} \approx \left\{a + b\sqrt{v} + \left(c + d\sqrt{v}\right)\sqrt{-\log_{10}\alpha}\right\}^2 . \tag{3}$$

where a, b, c and d are the constants

$$a = -1.37266, \quad b = 1.06807, \quad c = 2.13161 \quad \text{and} \quad d = -0.04589\,. \tag{4}$$

Hoaglin (1977) notes that (4) is suitable when $v < 30$ which makes it ideal for studying contingency tables since most studies involve data whose degrees of freedom do not exceed this value. He also demonstrates that (3) – (4) has a relative fit that does not exceed about 0.025%.

Based on (3) – (4), Beh (2018) showed that the p-value of a Pearson's chi-squared statistic, X^2, can be well approximated by

$$P_1\left(\chi^2 > X^2\right) \approx \begin{cases} \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{X^2} - (a + b\sqrt{v})}{(c + d\sqrt{v})}\right)^2} & , \quad X^2 \geq \left(a + b\sqrt{v}\right)^2 \\ 1 & , \quad X^2 < \left(a + b\sqrt{v}\right)^2 \end{cases} \tag{5}$$

where a, b, c and d are defined by (4). When compared with the p-value obtained from the `pchisq()` function in R – which we refer to in this paper as the "true" p-value – P_1 does not always achieve the same level of accuracy as Hoaglin (1977) reported for (3) - (4). However, Beh (2018) did show that when $v = 5, 10$ and 20, for example, the error in approximating the p-value using P_1 is no more than 2% for $X^2 \in [10, 30]$; our three examples in Section 3 have a Pearson chi-squared within this interval or lie just outside of its limits. For $X^2 > 30$, the p-values are all small and any error in (5) has no impact on the conclusions reached.

There are a range of alternatives to (1) that one may consider for assessing the statistical significance of the association between categorical variables. Some of the more popular options are special cases of the Cressie-Read family of divergence statistics (Cressie and Read 1984). We therefore now turn our attention to defining this family and give a few of the popular special cases that may be derived from it.

2.2. The family of divergence statistics

For some value of $\delta \in (-\infty, \infty)$, Cressie and Read (1984) proposed the following

$$CR(\delta) = \frac{2}{\delta(\delta+1)} \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \left\{ \left(\frac{n_{ij}}{np_{i\bullet}p_{\bullet j}} \right)^{\delta} - 1 \right\} \tag{6}$$

which, like Pearson's statistic, is a chi-squared random variable with $(I-1)(J-1)$ degrees of freedom that remain fixed for all δ; see, for example Cressie and Read (1984, p. 443), Drost, Kallenberg, Moore, and Oosterhoff (1989) and Agresti (2013, p. 34). The measure $CR(\delta)$ given by (6) is referred to as the *Cressie-Read family of divergence statistics*. The choice of δ determines the measure of association used to examine the nature of the association between two cross-classified categorical variables. These include Pearson's statistic, $X^2 = CR(\delta = 1)$, and the following

$$\begin{aligned}
G^2 &= CR(\delta = 0) = 2n \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \ln \left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}} \right) , \\
T^2 &= CR\left(\delta = -\frac{1}{2}\right) = 4n \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\sqrt{p_{ij}} - \sqrt{p_{i\bullet}p_{\bullet j}}\right)^2 , \\
N^2 &= CR(\delta = -2) = n \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(p_{ij} - p_{i\bullet}p_{\bullet j})^2}{p_{ij}} , \\
M^2 &= CR(\delta = -1) = 2n \sum_{i=1}^{I} \sum_{j=1}^{J} p_{i\bullet}p_{\bullet j} \ln \left(\frac{p_{i\bullet}p_{\bullet j}}{p_{ij}} \right) .
\end{aligned}$$

These four measures are, respectively, the log-likelihood ratio statistic (Wilks 1938), the Freeman-Tukey statistic (Freeman and Tukey 1950), the modified chi-squared statistic (Neyman 1940, 1949) and the modified log-likelihood ratio statistic (Kullback 1959) for a two-way contingency table. Another commonly used measure of association that is a special case of (6) is when $\delta = 2/3$ yielding the *Cressie-Read statistic*.

Thus, while the approximation to the p-value given by (5) was confined to Pearson's chi-squared statistic, it can be easily applied to G^2, T^2, N^2, M^2 and any other measure of association that is derived from (6). We now turn our attention to demonstrating how this can be done.

2.3. Three approximations of the p-value

Since X^2, G^2, T^2, N^2 and M^2, for example, are all chi-squared random variables then the p-value of any member derived from the Cressie-Read family of divergence statistics, (6), can be approximated by amending (5) so that

$$P_1\left(\chi^2 > CR(\delta) \mid \delta\right) = \begin{cases} \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{CR(\delta)} - (a + b\sqrt{v})}{(c + d\sqrt{v})}\right)^2} & , \quad CR(\delta) \geq (a + b\sqrt{v})^2 \\ 1 & , \quad CR(\delta) < (a + b\sqrt{v})^2 \end{cases} \tag{7}$$

for a given value of δ and where a, b, c and d are defined by (4).

While the approximation to the upper-tail quantile of the chi-squared distribution proposed by Hoaglin (1977) provides excellent approximations of the p-value, he also proposed his "Fit A",

$$\chi^2_\alpha \approx \left\{1.00991\sqrt{v} + 1.95188\left(-\log_{10}\alpha\right)^{1/2} - 1.14485\right\}^2 . \tag{8}$$

See his equations (4.1) and (4.2) or, equivalently, his (5.1). Hoaglin (1977, eq (5.3)) considered a further simplification of the quantile approximation by defining his "Fit S" so that

$$\chi^2_\alpha \approx \left\{\sqrt{v} + 2\left(-\log_{10}\alpha\right)^{1/2} - 7/6\right\}^2 . \tag{9}$$

Equations (8) and (9) lead to two further approximations of the p-value for any member of (6). Based on (8), one may obtain

$$P_2\left(\chi^2 > CR\left(\delta\right) \mid \delta\right) = \begin{cases} \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{CR(\delta)}-\left(1.00991\sqrt{v}-1.14485\right)}{1.95188}\right)^2} & , \quad X^2 \geq \left(1.00991\sqrt{v}-1.14485\right)^2 \\ 1 & , \quad X^2 < \left(1.00991\sqrt{v}-1.14485\right)^2 \end{cases} \tag{10}$$

while, from (9), a more simplified approximate may be obtained from

$$P_3\left(\chi^2 > CR\left(\delta\right) \mid \delta\right) = \begin{cases} \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{CR(\delta)}-\left(\sqrt{v}-7/6\right)}{2}\right)^2} & , \quad X^2 \geq \left(\sqrt{v}-7/6\right)^2 \\ 1 & , \quad X^2 < \left(\sqrt{v}-7/6\right)^2 \end{cases} . \tag{11}$$

While Beh (2018) provided an extensive simulation study of the accuracy of (5), we shall refrain from repeating the study for (7), (10) and (11). This is because the Cressie-Read family of divergence statistics, (6), can produce any chi-squared statistic, including those ranging from 0 to 100, which Beh (2018) considered in his study of (5). Instead, the next section shall take a more empirical approach and consider the accuracy of the three approximations to the p-value – (7), (10) and (11) – by studying three two-way contingency tables that differ in sample size (n), degrees of freedom (v), and in the statistical significance of the association between their variables. All testing is performed at the 0.05 level of significance.

3. Three examples

3.1. Kidney data

The data set

The first example is a re-examination of the the kidney function data of Chu *et al.* (2013) that Beh (2018) considered in his study of (5). The data comes from a study that looks at the swelling of the kidneys due to urine build up (called hydronephrosis) after 51 kidney transplants were performed at the Children's Hospital of Pittsburgh between May 1998 and May 2008. Table 1 gives the 2×2 contingency table that is formed from the cross-classification of the patient's *Gender* and whether they experienced *Hydronephrosis*.

Table 1: 2×2 table of gender and hydronephrosis status

	Hyrdonephrosis		
Gender	Yes	No	Total
Male	22	13	35
Female	3	13	16
Total	25	26	51

The true p-value

A chi-squared test of independence of Table 1 is performed with 1 degree of freedom and yields a Pearson statistic of 8.5480; it was calculated without incorporating Yates' continuinty correction, although one could, as Beh (2018) did. Using the R function `pchisq()` the true p-value for this statistic is, to 9 decimal places (9dp),

```
> 1 - pchisq(8.5480, 1)
[1] 0.003459021
>
```

so that

$$P_0\left(\chi^2 > 8.5480\right) \equiv P_0\left(\chi^2 > CR\,(1)\right) = 0.003459021\,.$$

Three approximation of the p-value

Since X^2 for Table 1 is greater than $\left(-1.37266 + 1.06807\sqrt{1}\right)^2 = 0.09277507$ we can approximate its p-value using (7) for $\delta = 1$ so that

$$P_1\left(\chi^2 > 8.5480 \,|\, \delta = 1\right) = \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{8.5480} - (-1.37266+1.06807\sqrt{1})}{(2.13161-0.04589\sqrt{1})}\right)^2} = 0.00402$$

and is within 17% of the true p-value. This may appear to be rather imprecise however given the small magnitude of the true p-value, the approximate value is certainly acceptable for inferential purposes.

The p-value can also be approximated using (10). Since Pearson's statistic of the table exceeds $\left(-1.14485 + 1.00991\sqrt{1}\right)^2 = 0.0182088$ then

$$P_2\left(\chi^2 > 8.5480 \,|\, \delta = 1\right) = \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{8.5480} - (-1.14485+1.00991\sqrt{1})}{1.95188}\right)^2} = 0.00350$$

and is within 1.27% of the true value. Using (11), the p-value is approximated to be

$$P_3\left(\chi^2 > 8.5480 \,|\, \delta = 1\right) = \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{8.5480} - (\sqrt{1}-7/6)}{2}\right)^2} = 0.00410$$

and, for practical purposes, provides just as good as approximation as P_1, being within 18.4% of the true value.

Assessing the three approximations

We can repeat the calculations given above by approximating the p-value for the log-likelihood ratio statistic, Freeman-Tukey statistic, modified chi-squared statistic and the modified log-likelihood ratio statistic of Table 1; we note that all calculations are rounded to six decimal places. These statistics are $G^2 = 9.0591$, $T^2 = 9.5472$, $N^2 = 12.4935$ and $M^2 = 10.2401$ respectively. Using the `pchisq()` function in R, the true p-value of these statistics is 0.00261, 0.00225, 0.00041 and 0.00150, respectively and, like the p-value for Pearson's statistic, all show that a statistically significant association exists between the variables of Table 1. These p-values are summarised to 5dp in Table 2; see the column labelled P_0.

The approximation of these p-values using (7), (10) and (11) are also summarised in Table 2; see the columns labelled P_1, P_2 and P_3, respectively. The precision of these approximations, when compared with their true p-value (P_0), is also given but to 3dp and is calculated by

$$\text{Precision} = 100\left|1 - \frac{P_n}{P_0}\right|$$

for $n = 1, 2$ and 3. These results show that (7) is within 0.17% of the true p-value for the modified chi-squared statistic, N^2 (when $\delta = -2$), but the approximation worsens as $\delta \to 1$. Although, irrespective of the choice of δ, (7) provides acceptable approximations of the p-value since their magnitude remains small enough so that the conclusion reached on the statistical significancy of the association between the variables does not change. A similar behaviour can also be observed for the accuracy of (11); that is, P_3. Contrary to P_1 and P_3, the approximation given by (10) improves from 28.47% for N^2 to within 1.3% of the true p-value when $\delta = 1$.

Table 2: Comparison of p-value and its approximations for the five special cases of the family of divergence statistics for Table 1

δ	Measure	Statistic	P_0	P_1	%	P_2	%	P_3	%
-2	N^2	12.49352	0.00041	0.00041	0.17	0.00029	28.47	0.00038	7.94
-1	M^2	10.24012	0.00137	0.00150	9.29	0.00120	12.36	0.00147	6.74
-1/2	T^2	9.54172	0.00200	0.00225	12.14	0.00186	6.93	0.00223	11.48
0	G^2	9.05911	0.00261	0.00298	14.15	0.00254	2.97	0.00300	14.86
1	X^2	8.54796	0.00346	0.00402	16.24	0.00350	1.27	0.00410	18.43

A visual inspection of the precision of (7), (10) and (11) when compared with P_0 can be seen by observing Figures 1 and 2 where the findings of Table 2 described above are reflected in these figures. Figure 1 also shows that P_1 and P_3 have quite similar levels of accuracy to P_0, especially for $\delta \in [-1\,, 1]$ while P_2 gives the best of the three approximations for δ lying between about -0.25 to 1.

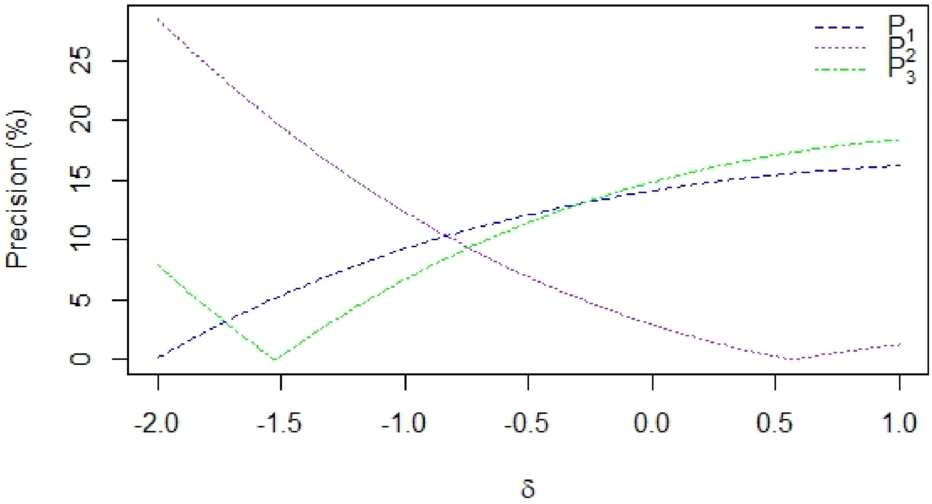

Figure 1: Precision of P_1, P_2 and P_3 for the divergence statistic for Table 1; $\delta \in [-2,\, 1]$

3.2. Smoking data

The data set

Our second example is an artificial contingency table of size 5 × 4 first used by Greenacre (1984, Table 3.1) to describe the key algebraic and visual features of correspondence analysis; see Table 3. It is a classification of 193 fictitious staff according to how often they smoke cigarettes (*Smoking*) and their position (*Position*) within the fictitious company. For the *Smoking* variable, those who do not smoke are classified by "None", "Light" smokers are those who smoke between 1 and 10 cigarettes a day, "Medium" smokers are those who smoke between 11 and 20 cigarettes a day while a "Heavy" smoker is classed as one who smokes more than 20 cigarettes a day; see Greenacre (1984, p. 56). The *Position* variable consists of the categories "Senior Manager", "Junior Manager", "Senior Employee", "Junior Employee" and "Secretaries".

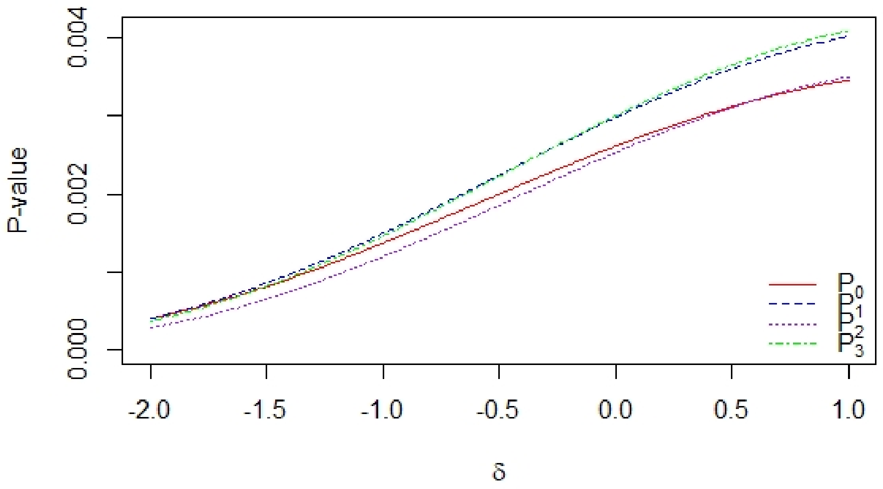

Figure 2: Comparison of P_0, P_1, P_2 and P_3 for the divergence statistic for Table 1; $\delta \in [-2, 1]$

Table 3: Ficticious contingency table of 193 workers according to their smoking status and position within the company

	Smoking Status				
Position	None	Light	Medium	Heavy	Total
Senior Managers	4	2	3	2	35
Junior Managers	4	3	7	4	18
Senior Employees	25	10	12	4	51
Junior Employees	18	24	33	13	88
Secretaries	10	6	7	2	25
Total	61	45	62	25	193

The true p-value

A chi-squared test of independence of Table 3 gives a Pearson statistic of 16.4416. With 12 degrees of freedom, this statistic has a p-value of 0.17184 (to 5dp);

```
> 1 - pchisq(16.4416, 12)
[1] 0.1718366
>
```

so that one may conclude that there is no evidence of a statistically significant association between *Smoking* and *Position*.

Three approximations of the p-value

The true p-value of Table 3 can be approximated using (7) since Pearson's statistic for the contingency table is greater than $\left(-1.37266 + 1.06807\sqrt{12}\right)^2 = 5.41606$. Therefore, calculating this approximation gives

$$P_1\left(\chi^2 > 16.4416 \,|\, \delta = 1\right) \approx \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{16.4416}-(-1.37266+1.06807\sqrt{12})}{(2.13161-0.04589\sqrt{12})}\right)^2} = 0.17184$$

and, to 5dp is identical to P_0 and within 0.5% of the true p-value beyond 5dp. The value of P_2 may also be approximated using (10) since $\left(-1.14485 + 1.00991\sqrt{12}\right)^2 = 5.539343$. Thus

$$P_2\left(\chi^2 > 16.4416 \,|\, \delta = 1\right) = \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{16.4416}-(-1.14485+1.00991\sqrt{12})}{1.95188}\right)^2} = 0.17391$$

and is within about 1.2% of the true value. Using (11), the p-value is approximated to be

$$P_3\left(\chi^2 > 16.4416 \,|\, \delta = 1\right) = \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{16.4416}-(\sqrt{12}-7/6)}{2}\right)^2} = 0.16900$$

and is within 1.7% of the true value. For these three approximations of the true p-value, the levels of accuracy are excellent given that the true p-value is quite large. Note that, none of the approximations contradict the nature of the association that the true p-value gives.

Assessing the three approximations

Suppose we supplement our findings of the p-value of X^2 and now turn our attention to the four additional special cases of the Cressie-Read family of divergence statistics that we described above. For Table 3, the log-likelihood ratio statistic, Freeman-Tukey statistic, modified chi-squared statistic and modified log-likelihood ratio statistic and are $N^2 = 17.4930$, $M^2 = 16.6921$, $T^2 = 16.4647$ and $G^2 = 16.3476$, respectively. The true p-value of each of these statistics exceeds 0.13 thus, like the p-value of X^2, confirms that there is no evidence of a statistically significant association between the two categorical variables of Table 3. The true p-value for each of these statistics is summarised in the fourth column of Table 4 labelled P_0.

Table 4: Comparison of P_0, P_1, P_2 and P_3 for the five special cases of the family of divergence statistics for Table 3

δ	Measure	Statistic	P_0	P_1	%	P_2	%	P_3	%
-2	N^2	17.49299	0.13198	0.13047	1.14	0.13245	0.36	0.12932	2.01
-1	M^2	16.69207	0.16155	0.16050	0.65	0.16316	0.99	0.15872	1.76
-1/2	T^2	16.46471	0.17087	0.17002	0.50	0.17289	1.19	0.16803	1.66
0	G^2	16.34759	0.17583	0.17511	0.41	0.17810	1.29	0.17301	1.61
1	X^2	16.44164	0.17184	0.17101	0.48	0.17391	1.21	0.16900	1.65

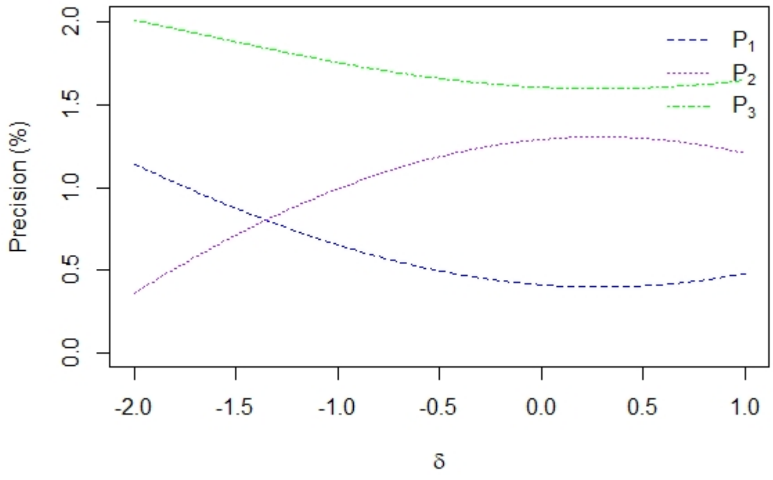

Figure 3: Precision of P_1, P_2 and P_3 for the divergence statistic for Table 3; $\delta \in [-2, 1]$

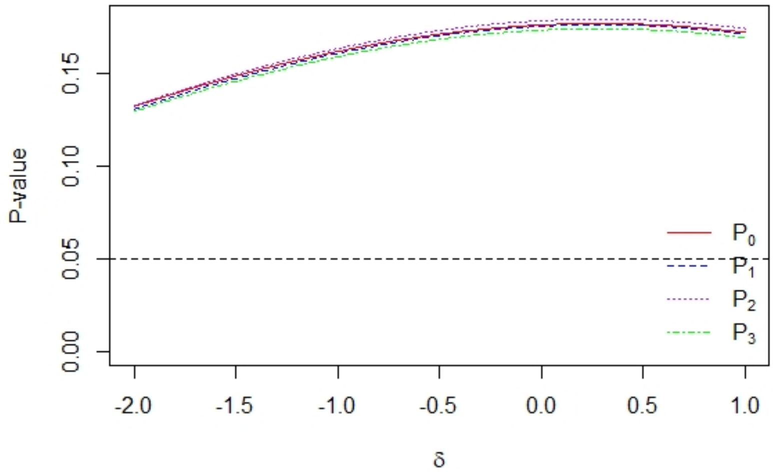

Figure 4: Comparison of P_0, P_1, P_2 and P_3 for the divergence statistic for Table 3; $\delta \in [-2,\, 1]$

Table 4 shows that (7), (10) and (11) all provide excellent approximations of the true p-value with the greatest error of approximation being 2.01%; for P_3 of the modified chi-squared statistic (when $\delta = -2$). Figure 3 and Figure 4 provide a visual description of the three approximations when compared with P_0 for $\delta \in [-2,\, 1]$ and confirm the high precision of P_1, P_2 and P_3 for all δ that lie within this interval. In fact, Figure 3 shows that, irrespective of the choice of $\delta \in [-2,\, 1]$, P_2 lies within 1.5% of the true p-value while P_1 lies within about 1.0%. Figure 4 shows that any error in the approximation of P_0, using (7), (10) and (11), is practically negligible and that all approximations, like P_0, exceed the 0.05 level of significance.

3.3. Dream data

The data set

The last contingency table that we shall consider is of size 5×4 and comes from Maxwell (1961, pp. 70 - 72). A random sample of 222 boys was asked to rate how disturbed they were by their dreams on a four point scale from 1 (least disturbing) to 4 (most disturbing). The age of the boys, in years, was recorded and the contingency table formed from the cross-classification of the *Age* of the boys and the *Rating* of their dreams giving Table 5.

The true p-value

A chi-squared test of independence for Table 5 gives a Pearson statistic of 32.1255. With $(5-1)\,(4-1) = 12$ degrees of freedom, the true p-value is, to 5dp, 0.00132;

```
> 1 - pchisq(32.1255, 12)
[1] 0.001323393
>
```

Therefore, there is enough evidence to conclude that there is a statistically significant association between *Age* and *Rating*.

Three approximations of the p-value

Using (7) for $\delta = 1$, we can directly approximate the p-value for our test statistic of 32.1255. Since the test statistic is nearly six times greater than $\left(-1.37266 + 1.06807\sqrt{12}\right)^2 = 5.41606$,

Table 5: Contingency table of 222 boys according to their age (in years) and dream rating

	Rating				
Age	4	3	2	1	Total
5 - 7	7	3	4	7	21
8 - 9	13	11	15	10	49
10 - 11	7	11	9	23	50
12 - 13	10	12	8	28	58
14 - 15	3	4	5	32	44
Total	40	41	41	100	222

we approximate its p-value to be

$$P_1\left(\chi^2 > 32.1255 \,|\, \delta = 1\right) = \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{32.1255}-(-1.37266+1.06807\sqrt{12})}{(2.13161-0.04589\sqrt{12})}\right)^2} = 0.001355$$

and is within 2.4% of the true p-value. By using (10) we get

$$P_2\left(\chi^2 > 32.1255 \,|\, \delta = 1\right) = \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{32.1255}-(-1.14485+1.00991\sqrt{12})}{1.95188}\right)^2} = 0.00131$$

which is a valid approximation since $32.1255 > \left(-1.14485 + 1.00991\sqrt{12}\right)^2 = 5.539343$. This approximation of the p-value is accurate to the fifth decimal place and lies within 1.15% of the true p-value, showing that P_2 is a better approximation than P_1. However, P_3 proves to be a relatively bad approximation of the p-value of X^2. Using (10)

$$P_3\left(\chi^2 > 32.1255 \,|\, \delta = 1\right) = \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{32.1255}-(\sqrt{12}-7/6)}{2}\right)^2} = 0.00145$$

which is only accurate to the third decimal place and incurs a level of inaccuracy of about 9%.

Table 6: Comparison of p-value and its approximations for the five special cases of the family of divergence statistics for Table 5

δ	Measure	Statistic	P_0	P_1	%	P_2	%	P_3	%
-2	N^2	42.24191	0.00003	0.00003	10.92	0.00003	1.59	0.00004	27.10
-1	M^2	36.03000	0.00032	0.00034	5.46	0.00032	0.12	0.00037	15.45
-1/2	T^2	34.17281	0.00063	0.00066	3.96	0.00063	0.63	0.00071	12.36
0	G^2	32.95764	0.00098	0.00101	3.02	0.00097	0.94	0.00109	10.46
1	X^2	32.12549	0.00132	0.00136	2.41	0.00131	1.15	0.00145	9.20

Assessing the three approximations

Table 6 provides a summary of the true p-value, P_0, and its approximations – P_1, P_2 and P_3 – for Table 5 for $\delta = -2$, -1, $-1/2$, 0 and 1. It shows that P_1 performed relatively well and, for $\delta = -1$, $-1/2$, 0 and 1, is accurate to within about 5% of the true value. The approximation is relatively poor when determining the p-value for the modified chi-squared statistic ($\delta = -2$) and has an error of about 11% when compared with the true value. Although, the more simple approximation of P_2 proves to produce much better approximations of the p-value for our five δ values since P_2 is within about 1.6% of P_0. In fact, we see that the approximate p-value for

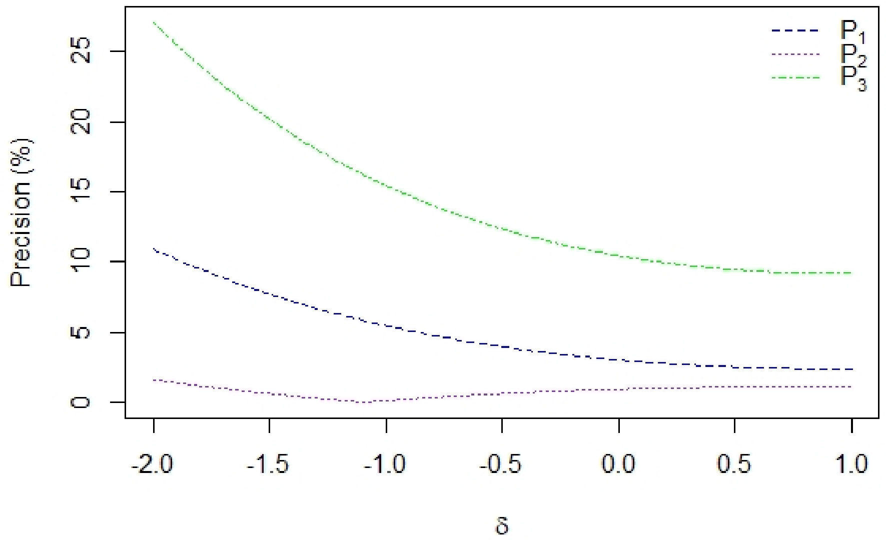

Figure 5: Precision of P_1, P_2 and P_3 for the divergence statistic for Table 5; $\delta \in [-2, 1]$

M^2 using (10) is within 0.12% of the true value. However, it is P_3 that produces the poorest approximations of the true p-value. For our five values of δ, (10) produces approximations of the true p-value that exceed about 10%. The worst approximation is, as we saw for P_1, when $\delta = -2$ and is in error in the order of 27%. Figure 5 clearly shows that, for $\delta \in [-2, 1]$, P_2 provides the best approximation of the true p-value, lying within 2% for all values of δ. This figure also clearly shows P_3 to be the poorest of the three approximation methods, exceeding about 10% for all values of δ.

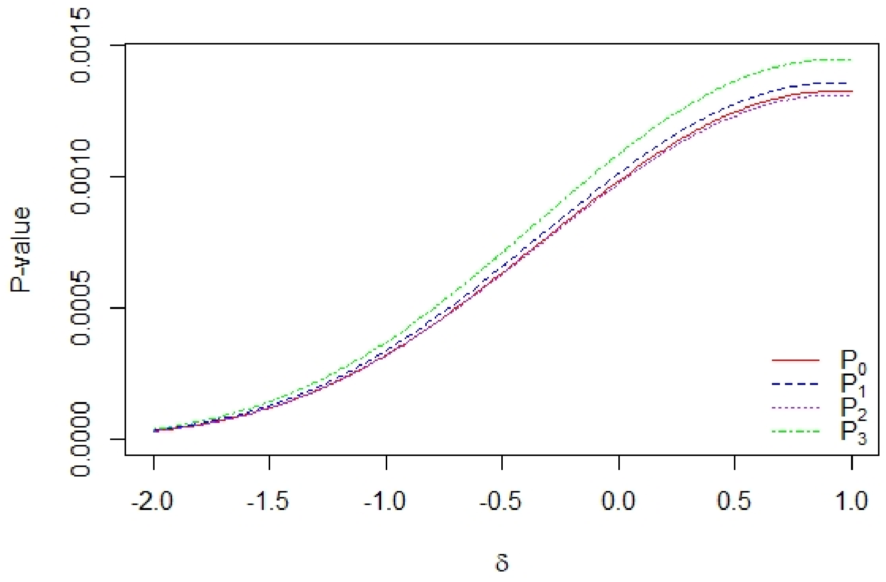

Figure 6: Comparison of P_0, P_1, P_2 and P_3 for the divergence statistic for Table 5; $\delta \in [-2, 1]$

While P_3 for Table 5 appears to produce the worst p-value approximation of any approximation across the three examples we have considered, any discrepancy with the true p-value is not practically relevant. This is because the true p-value for all five special cases of divergence statistic, (6), are all very small (being less than 0.0015). Figure 6 shows how close P_1, P_2 and

P_3 are to P_0; even for values of δ close to -2 where the greatest discrepancies lie, there is no practical difference between the true p-value and the three approximations.

4. Discussion

While calculations of the p-value for a chi-squared random variable have been developed (and largely rely on knowing the quantile of the chi-squared or standard normal distribution), a common difficulty in obtaining the p-value relies on computing reliable approximations of complex formulae like (2). However, the approximation described by Beh (2018), based on the quantile approximation of Hoaglin (1977), has helped to address this difficulty. This paper expands upon the approximation given by Beh (2018) and provides an approximation of the p-value for any chi-squared statistic that belongs to the Cressie-Read family of divergence statistics. This paper also provides two additional approximations of the p-value that prove relatively simple to calculate while providing excellent levels of accuracy for most practical cases.

From the empirical study of the three contingency tables considered, we have shown that all three approximations – P_1 given (7), P_2 given by (10) and P_3 given by (11) – give excellent levels of accuracy when compared with the true p-value. In many cases, the accuracy of these approximations lie within 2% of the true p-value. For small dimensional contingency tables with a relatively small sample size, such as Table 1, P_2 proved accurate for δ ranging between -0.25 to 1 thereby showing that (10) approximates well the p-value for the Freeman-Tukey, log-likelihood ratio and Pearson statistics in such cases. For Table 1, P_1 and P_3 gave better approximations for the remaining chi-squared statistics. For the two larger contingency tables – see Table's 3 and 5 – approximations using P_2 were within 1.6% of the true p-value for all values of $\delta \in [-2, 1]$. While we did not consider the case when $\delta = 2/3$ which lies within this interval, the p-value approximations for the chi-squared statistic $CR\,(2/3)$ – better known as the Cressie-Read statistic (Cressie and Read 1984) – are comparable to those of the Freeman-Tukey statistic.

One observation that can be made from the three approximations studied in this paper is that their performance greatly improves as the chi-squared statistic increases, and this will often be the case for large sample sizes (n); note from (1) that multiplying the sample size by some constant $C > 1$ will increase the value of the chi-squared statistic by a factor of C (this also applies to all members of the Cressie-Read family of divergence statistics). Such a performance of the approximation for large chi-squared values is reminiscent of the finding made by Beh (2018) that a more accurate approximation of the p-value can be gained for larger values of the chi-squared statistic, regardless of the degrees of freedom. Even for cases where an approximation appears relatively inaccurate (being more than, say, 10% different from its true value), the p-value was very small (less than 0.005) and so did not impact on the statistical significance of chi-squared statistic. Therefore, in most practical situations the simplicity of P_3 provides easy to calculate approximations of the p-value, although more accurate values can be obtained using P_1 or P_2. Of course, as Beh (2018) pointed out, if far more precise calculations of the p-value are required then the researcher can still continue to use the range of statistical packages that are available for analysing data.

References

Agresti A (2013). *Categorical Data Analysis (3rd edn).* Wiley. URL https://www.wiley.com/en-au/Categorical+Data+Analysis%2C+3rd+Edition-p-9780470463635.

Aroian LA (1943). "A New Approximation to the Levels of Significance of the Chi-square Distribution." *Annals of Mathematical Statistics*, **14**, 93 – 95. doi:10.1214/aoms/1177731497.

Beh EJ (2018). "Exploring How to Simply Approximate the P-value of a Chi-Squared Statistic." *Austrian Journal of Statistics*, **47**, 63 – 75. doi:10.17713/ajs.v47i3.757.

Chu L, Jacobs BL, Schwen Z, Schneck FX (2013). "Hydronephrosis in Pediatric Kidney Transplant: Clinical Relevance to Graft Outcome." *Journal of Pediatric Urology*, **9**, 217 – 222. doi:10.1016/j.jpurol.2012.02.012.

Cressie N, Read TRC (1984). "Multinomial Goodness-of-Fit Tests." *Journal of the Royal Statistical Society, Series B*, **46**, 440 – 464. doi:10.1111/j.2517-6161.1984.tb01318.x.

Ding CG (1992). "Algorithm AS275: Computing the Non-central Chi-squared Distribution Function." *Applied Statistics*, **41**, 478 – 482. doi:10.2307/2347584.

Drost FC, Kallenberg WCM, Moore DS, Oosterhoff J (1989). "Power Approximations to the Multinomial Tests of Fit." *Journal of the American Statistical Association*, **84**, 130 – 141. doi:10.1080/01621459.1989.10478748.

Elderton WP (1902). "Tables for Testing the Goodness of Fit of Theory to Observation." *Biometrika*, **1**, 155 – 163. doi:10.2307/2331485.

Fisher RA (1928). *Statistical Methods for Research Workers (2nd ed.).* Oliver and Boyd.

Freeman MF, Tukey JW (1950). "Transformations Related to the Angular and Square Root." *The Annals of Mathematical Statistics*, **21**, 607 – 611. URL http://www.jstor.org/stable/2236611.

Goldberg H, Levine H (1946). "Approximate Formulas for the Percentage Points and Normalization of t and X^2." *Annals of Mathematical Statistics*, **17**, 216 – 225. doi:10.1214/aoms/1177730982.

Greenacre MJ (1984). *Theory and Applications of Correspondence Analysis.* Academic Press, London. URL http://www.carme-n.org/?sec=books5.

Hald A, Sinkbaek SA (1950). "A Table of Percentage Points of the χ^2 Distribution." *Scandinavian Actuarial Journal*, **33**, 168 – 175. doi:10.1080/03461238.1950.10432038.

Heyworth MR (1976). "Approximation to Chi-square." *The American Statistician*, **30**, 204. doi:10.1080/00031305.1976.10479181.

Hoaglin DC (1977). "Direct Approximations for Chi-squared Percentage Points." *Journal of the American Statistical Association*, **72**, 508 – 515. doi:10.1080/01621459.1977.10480604.

Khamis S, Rudert W (1965). *Tables of the Incomplete Gamma Function Ratio: Chi-square Integral, Poisson Distribution.* Justus von Leibig, Darmstad.

Kullback S (1959). *Information Theory and Statistics.* Wiley. doi:10.1137/1002033.

Lin JT (1988). "Approximating the Cumulative Chi-square Distribution and Its Inverse." *The Statistician*, **37**, 3 – 5. doi:10.2307/2348373.

Maxwell AE (1961). *Analysing Qualitative Data.* Methuen, London.

Merrington M (1941). "Numerical Approximations to the Percentage Points of the χ^2 Distribution." *Biometrika*, **32**, 200 – 202. doi:10.1093/biomet/32.2.200.

Neyman J (1940). "Contribution to the Theory of Certain Test Criteria." *Bulletin de L'Institut International de Statistique*, **24**, 44 – 86. doi:10.1525/9780520327016-010.

Neyman J (1949). "Contributions to the Theory of the χ^2 Test." *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pp. 239 – 273.

Pearson K (1904). "On the Theory of Contingency and Its Relation to Association and Normal Correlation." *Drapers Memoirs, Biometric Series*, **1**, 35.

Pearson K (1922). *Tables of the Incomplete Gamma-Function*. Cambridge University Press.

Russell W, Lal M (1969). "Tables of Chi-square Probability Function." Department of Mathematics, St. Johns: Memorial University of Newfoundland. (Reviewed in *Mathematics of Computation*, 23, 211 – 212.).

Terrell GR (1984). "Chi-squared Left-tail Probabilities." *Journal of Statistical Computation and Simulation*, **28**, 264 – 266. doi:10.1080/00949658708811034.

Thompson CM (1941). "Table of Percentage Points of the χ^2 Distribution." *Biometrika*, **32**, 187 – 191. doi:10.1093/biomet/32.2.187.

Wilks SS (1938). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *Annals of Mathematical Statistics*, **9**, 60 – 62. doi:10.1214/aoms/1177732360.

Wilson EB, Hilferty MM (1931). "The Distribution of Chi-square." *Proceedings of the National Academy of Sciences of the United States of America*, **17**, 684 – 688. doi:10.1073/pnas.17.12.684.

Affiliation:

Eric J. Beh
National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong
Wollongong, NSW, 2522, Australia
E-mail: ericb@uow.edu.au
and
Centre for Multi-Dimensional Data Visualisation (MuViSU),
Stellenbosch University, Matieland, 7602, South Africa

Ting-Wu Wang
School of Information and Physical Sciences,
University of Newcastle,
Callaghan, NSW, 2308, Australia

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 52 *Submitted:* 2022-09-07
2023 *Accepted:* 2023-02-24