

Exploring How to Simply Approximate the P-value of a Chi-Squared Statistic

Eric J. Beh

University of Newcastle

Abstract

Calculating the p-value of any test statistic is of paramount importance to all statistically minded researchers across all areas of study. Many, these days, take for granted how the p-value is calculated and yet it is a pivotal quantity in all forms of statistical analysis. For the study of 2×2 tables where dichotomous variables are assessed for association, the chi-squared statistic, and its p-value, are fundamental quantities to all analysts, especially those in the health and allied disciplines. Examining the association between dichotomous variables is easily achieved through a very simple formula for the chi-squared statistic and yet the p-value of this statistic requires far more computational effort.

This paper proposes and explores a very simple approximation of the p-value for a chi-squared statistic given its degrees of freedom. After providing a review of a variety of common ways for determining the quantile of the chi-squared distribution given the level of significance and degrees of freedom, we shall derive an approximation based on the classic quantile formula given in 1977 by D. C. Hoaglin. We examine this approximation using a simple 2×2 contingency table example then show that it is extremely precise for all chi-squared values ranging from 0 to 50.

Keywords: chi-squared statistic, Fisher's approximation, Hoaglin's approximation, p-value approximation, Wilson and Hilferty's approximation.

1. Introduction

For over 100 years Pearson's chi-squared statistic (Pearson 1904) has been one of the most popular test statistics used within, and outside of, the statistics discipline. This is because it can be generally used either as a test statistic when undertaking a test of independence between two or more categorical variables or for goodness-of-fit purposes. Its popularity and utility has therefore seen a great deal of attention (largely outside of the statistics literature) given to its use and misuse for at least 60 years; see, for example, Lewis and Burke (1949), Edwards (1950), Delucchi (1983) and Dijkers (2005). Even the name of the statistic – “chi-square” versus “chi-squared” – has created much discussion (Beh and Lombardo 2014). We shall not be drawn into that issue here but some excellent insights into the issues surrounding the use of the chi-squared statistic in a variety of situations can also be found in Fienberg (1979). Excellent discussions are available on the development of the chi-squared statistic including those given by Lancaster (1969), Cressie and Read (1984) and Fleiss, Levin, and

Paik (2003).

At the core of the chi-squared statistic lays its simplicity and applicability from which the analyst may make conclusions on the statistical significance of the test being undertaken. Central to making inferential decisions using the statistic is the calculation and determination of its p-value. While it is relatively simple to calculate Pearson's chi-squared statistic in many cases, determining its p-value requires more computational effort. For example, when testing the dependence between two variables of a 2×2 contingency table, the chi-squared statistic can be calculated using a scientific desk calculator, although its p-value cannot be obtained so easily.

Therefore, the aim of this paper is to derive a very simple expression for approximating the p-value of a chi-squared statistic given its degrees of freedom. In doing so, the p-value may be approximated without the aid of statistical packages. The resulting formula is based on the classic approximation of the quantile of the chi-squared distribution given by Hoaglin (1977). We shall demonstrate the accuracy and reliability of the simple approximation by comparing the results with those obtained by more traditional, or computational, means.

2. A simple illustrative example

Suppose we have a 2×2 contingency table with the notation defined in Table 1. When testing the statistical significance of the association between the rows and columns of Table 1, Pearson's chi-squared statistic is

$$X^2 = n \frac{(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}.$$

Yates' continuity correction (Yates 1934) is often included in the calculation of the chi-squared statistic such that

$$X_Y^2 = n \frac{(|ad - bc| - n/2)^2}{(a + c)(b + d)(a + b)(c + d)}.$$

The statistical significance of both test statistics can be assessed by considering that they are random variables from chi-squared distributed with 1 degree of freedom, for a given level of significance, α .

Table 1: Notation for a 2×2 contingency table

	Column 1	Column 2	Total
Row 1	a	b	$a + b$
Row 2	c	d	$c + d$
Total	$a + c$	$b + d$	$n = a + b + c + d$

To help illustrate that calculating the chi-squared statistic for a 2×2 table is very straightforward, suppose we consider data collected as part of a study undertaken to examine kidney function after transplant in paediatric patients (Chu, Jacobs, Schwen, and Schneck 2013). The focus of the study was to look at factors of hydronephrosis (swelling of the kidney due to a build-up of urine) after each of 51 kidney transplants had taken place at the Children's Hospital of Pittsburgh between May 1998 and May 2008. Table 2 is a cross-classification of a patient's gender and whether patients experienced hydronephrosis.

The simplicity of calculating the chi-squared statistic of Table 2 is apparent since

$$X^2 = 51 \times \frac{(22 \times 13 - 13 \times 3)^2}{25 \times 26 \times 35 \times 16} = 8.548$$

and can be found without the need of using statistical packages; all that is required is a hand held calculator (not even a scientific calculator). Incorporating Yates' continuity correction is an equally straight forward task:

$$X_Y^2 = 51 \times \frac{(|22 \times 13 - 13 \times 3| - 51/2)^2}{25 \times 26 \times 35 \times 16} = 6.874.$$

The p-value of X^2 and X_Y^2 are 0.0035 and 0.0087, respectively, and both suggest that there is a statistically significant association between a patient's gender and whether they experience hydronephrosis after a kidney transplant. However, these p-values cannot be calculated as simply as their test statistic. Instead, their calculation requires more computational effort.

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

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

This paper shall derive and demonstrate a simple approximation for calculating the p-value from a chi-squared test statistic. Before doing so, an overview will be given of some of the popular ways in which the quantile of the chi-squared distribution can be estimated.

3. Overview of Pearson's chi-squared statistic and its p-value

Suppose we consider an iid random sample, X_1, X_2, \dots, X_n from the chi-squared distribution with probability density function

$$f(x|v) = \frac{1}{\Gamma(v/2) 2^{v/2}} e^{-x/2} x^{(v/2)-1}, \quad x \geq 0$$

where v is the degrees of freedom of the distribution and $\Gamma(\bullet)$ is the gamma function. Tables of quantiles from the chi-squared distribution (historically also referred to as "percentage points") given specific levels of significance, α , and degrees of freedom, v , are well documented. For example, there are many contributions that provide tables of quantiles for various levels of significance and degrees of freedom at various levels of accuracy. Some of the earlier contributions include those from [Pearson \(1922\)](#), [Fisher \(1928\)](#), [Thompson \(1941\)](#), [Merrington \(1941\)](#), [Aroian \(1943\)](#), [Goldberg and Levine \(1946\)](#), [Hald and Sinkbaek \(1950\)](#), [Vanderbeck and Cooke \(1961\)](#), [Harter \(1964\)](#) and [Krauth and Steinebach \(1976\)](#). While these contributions give tables of quantiles, others provide simple formulae for calculating these values for any level of significance; see, for example, [Wilson and Hilferty \(1931\)](#) and [Heyworth \(1976\)](#). A far more superior approximation was proposed by [Hoaglin \(1977\)](#) which we shall discuss further in Section 4.

The issue of approximating the p-value of a chi-squared statistic given its degrees of freedom, is less well documented. From a mathematical point of view, suppose we consider

$$P(\chi_{\alpha, v}^2 > X^2) = \int_{X^2}^{\infty} \frac{1}{\Gamma(v/2) 2^{v/2}} e^{-u/2} x^{(u/2)-1} du$$

which is the p-value of the test statistic X^2 . Evaluating this integral requires computational procedures. However, tables of p-values have been documented in the past. For example, [Elderton \(1902\)](#) provides the first set of tables based on the approximations

$$P(\chi^2 > X^2) \approx \sqrt{\frac{2}{\pi}} \int_X^{\infty} e^{-X^2/2} dX + \sqrt{\frac{2}{\pi}} e^{-X^2/2} \left(\frac{X}{1} + \frac{X^3}{1 \cdot 3} + \dots, \frac{X^{v-2}}{1 \cdot 3 \cdot \dots \cdot (v-2)} \right) \quad (1)$$

for an odd number of degrees of freedom and

$$P(\chi^2 > X^2) \approx e^{-X^2/2} \left(\frac{X}{1} + \frac{X^3}{1 \cdot 3} + \dots + \frac{X^{v-2}}{1 \cdot 3 \cdot \dots \cdot (v-2)} \right)$$

for an even number of degrees of freedom. [Elderton \(1902\)](#) also provides an approximation to the integral term in (1) and so gives approximations of the p-value for integer valued chi-squared values ranging from 1 to 30, then 40, 50, 60 and 70 and where $v = 2$ to 29. Similarly, [Russell and Lal \(1969\)](#) provide tables of p-values for $v = 1$ (1) 50 and $X^2 = 0.001$ (0.001) 0.01 (0.01) 0.1 (0.1) 107. Here, for example, the notation 0.001 (0.001) 0.01 (0.01) 0.1 (0.1) 107 has been adopted to mean every value between 0.001 and 0.01 (inclusive) in increments of 0.01 then from 0.1 to 107 (inclusive) in increments of 0.1. See also a review of these tables in volume 23 of *Mathematics of Computation* (pages 211 – 212). More tables can be found by considering [Khamis and Rudert \(1965\)](#). Alternative simple procedures for approximating the p-value of a chi-squared statistic can also be found in [Terrell \(1984\)](#) and [Lin \(1988\)](#). The approach that [Lin \(1988\)](#) takes is based on the approximation of the chi-squared quantile proposed by [Fisher \(1928\)](#). While these approximations have been shown to be very effective, they don't lend themselves to such simplicity where anything but statistical, or mathematical, computer packages can be used to approximate the p-value. It should also be noted that this issue has also been explored for other distributions. For example, [Koehler \(1983\)](#) provides an approximation for a random variable from the t-distribution which was subsequently improved upon by [Lin \(1990\)](#).

4. Classical methods for calculating a chi-squared percentile

Given a level of significance, α , and the degrees of freedom v , classic approximations of the quantile of the chi-squared distribution can be made in a number of simple ways. Here we shall very briefly review four of them. They are the standardisation of the chi-squared random variable and the approximations of [Fisher \(1928\)](#), [Wilson and Hilferty \(1931\)](#) and [Hoaglin \(1977\)](#).

4.1. Central limit theorem

It is well known that when n is large, the mean and variance of the chi-squared distribution is v and $2v$, respectively; see, for example, [Haldane \(1940\)](#). Therefore, under the central limit theorem, the chi-squared statistic may be normalised such that

$$z_\alpha = \frac{\chi_{\alpha,v}^2 - v}{\sqrt{2v}}.$$

So, the quantile of the chi-squared distribution can be approximated by

$$\chi_{\alpha,v}^2 \approx v + z_\alpha \sqrt{2v}.$$

Here z_α is the 100(1 - α)'th percentile of the standard normal distribution. This approximation works well when v is large, however, for many practical situations, this may not be the case and so the approximation is not recommended.

4.2. Fisher's approximation

Another approximation, proposed by [Fisher \(1928\)](#), deals with first considering that

$$z_\alpha = \sqrt{2\chi_{\alpha,v}^2} - \sqrt{2v-1}$$

is a standard normally distributed random variable. Such an approximation is considered accurate when the degrees of freedom, v , is large, usually $v > 100$. When this is the case $\chi_{\alpha,v}^2$ can be approximated by *Fisher's approximation*

$$\chi_{\alpha,v}^2 \approx \frac{1}{2} \left(z_{\alpha} + \sqrt{2v-1} \right)^2 \quad (2)$$

while [Heyworth \(1976\)](#) suggested that, near the tails of the chi-squared distribution (eg, when α is small and v is large), (2) may be modified to

$$\chi_{\alpha,v}^2 \approx \frac{1}{2} \left(z_{\alpha} + \sqrt{2v} \right)^2 .$$

A study of this approximation in comparison with other approximations (including that of [Hoaglin \(1977\)](#) which we shall soon discuss), reveals that Fisher's approximation underestimates $\chi_{\alpha,v}^2$ by as much as 10% for $v = 3$ and $\alpha = 0.005$.

4.3. Wilson and Hilferty's approximation

In response to the results obtained by [Fisher \(1928\)](#), [Wilson and Hilferty \(1931\)](#) presented a number of alternative approximations to approximate $\chi_{\alpha,v}^2$. One of the more popular approximation is

$$\chi_{\alpha,v}^2 \approx v \left(z_{\alpha} \sqrt{\frac{2}{9v}} + \left(1 - \frac{2}{9v} \right) \right)^2 .$$

[Wilson and Hilferty \(1931\)](#), on page 688, remarked that this result is "somewhat remarkable". Other approximations they considered lead to the following results

$$\chi_{\alpha,v}^2 \approx \frac{1}{2} \left(z_{\alpha} + \sqrt{2v-2} \right)^2$$

which is akin to that of [Fisher \(1928\)](#), and the following equations proposed by [Heyworth \(1976\)](#)

$$\begin{aligned} \chi_{\alpha,v}^2 &\approx 2 \left(\frac{z_{\alpha}}{3} \left(\frac{v}{2} - \frac{1}{3} \right)^{-1/6} + \left(\frac{v}{2} - \frac{1}{3} \right)^{1/3} \right)^3 , \\ \chi_{\alpha,v}^2 &\approx \left(\frac{z_{\alpha} \sqrt{2}}{3} \left(v - \frac{2}{3} \right)^{-1/6} + \left(v - \frac{2}{3} \right)^{1/3} \right)^3 . \end{aligned}$$

4.4. Hoaglin's approximation

In all of the approximations of the chi-squared quantile described above, it is important to first calculate the quantile of the standard normal distribution, z_{α} , for the given level of significance, α . To avoid having to determine this quantity, one way to approximate $\chi_{\alpha,v}^2$ is to consider *Hoaglin's approximation* ([Hoaglin 1977](#))

$$\sqrt{\chi_{\alpha,v}^2} \approx a + b\sqrt{v} + (c + d\sqrt{v}) \sqrt{-\log_{10} \alpha} \quad (3)$$

where a , b , c and d are constants. After considering various specifications of these constants, [Hoaglin \(1977\)](#) showed that

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

provides an extremely good approximation to the chi-squared quantile; stating (on page 512) that only 8 out of the 140 simulations that were carried out were not within 0.4% of the theoretical quantile. Such an approximation is therefore both versatile and accurate. For this reason, we shall be using (3) – (4) to propose, and adapt, a very simple approximation of the p-value of a given chi-squared statistic with known degrees of freedom. Note that, since Hoaglin (1977) considered these values of a , b , c and d up to five decimal places, we shall also consider the same level of precision of the constants in in the following sections.

5. The simple approximation of the p-value

Since (3) can be used to approximate, given a level of significance, the quantile of the chi-squared distribution, it can also be used to approximate the p-value of a chi-squared statistic. To do so, we consider

$$\sqrt{X^2} \approx a + b\sqrt{v} + (c + d\sqrt{v}) \sqrt{-\log_{10} P(\chi^2 > X^2)} \quad (5)$$

where X^2 is the chi-squared statistic and $P(\chi^2 > X^2)$ is its p-value. Therefore, by rearranging (5), we obtain a very simple approximation of the p-value of X^2 ,

$$P(\chi^2 > X^2) \approx \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{X^2} - (a + b\sqrt{v})}{(c + d\sqrt{v})}\right)^2}. \quad (6)$$

Of course the approximation depends on the choice of a , b , c and d . To attain similar levels of accuracy as Hoaglin's approximation of the chi-squared quantile, a reasonable choice of the constants are those given in (4). Thus, we consider

$$P(\chi^2 > X^2) \approx \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{X^2} - (-1.37266 + 1.06807\sqrt{v})}{(2.13161 - 0.04589\sqrt{v})}\right)^2} \quad (7)$$

Just as a simple example, when $v = 1$ (as is the case for any $2 \times c$ table, where $c \geq 2$), (7) shows that an approximation of the p-value of a chi-squared statistic, X^2 , is

$$P(\chi^2 > X^2) \approx \left(\frac{1}{10}\right)^{(0.229873(\sqrt{X^2} + 0.304059))^2}$$

Therefore, for $X^2 = 3.841459$, its approximate p-value is $P(X^2) = 0.066$, which compares quite well with its theoretic p-value of 0.05. When we consider the chi-squared statistic of 18.30704 with 10 degrees of freedom the approximate p-value is 0.048 – within 4% of its theoretic p-value of 0.05. A more comprehensive examination of the approximation is given in Section 6.

Despite the simplicity of the approximation (7), $P(\chi^2 > X^2)$ is a quadratic function with a maximum of 1 achieved at the turning point $\sqrt{X^2} - (-1.37266 + 1.06807\sqrt{v}) = 0$. This is problematic for those values of $\sqrt{X^2} - (-1.37266 + 1.06807\sqrt{v}) < 0$. This is because, in this region, the approximated p-value will approach zero as $X^2 \rightarrow 0$ when it should be approaching 1. Therefore one may adapt (7) so that

$$P(\chi^2 > X^2) \approx \begin{cases} \left(\frac{1}{10}\right)^{\left(\frac{\sqrt{X^2} - (-1.37266 + 1.06807\sqrt{v})}{(2.13161 - 0.04589\sqrt{v})}\right)^2}, & X^2 \geq (-1.37266 + 1.06807\sqrt{v})^2 \\ 1, & X^2 < (-1.37266 + 1.06807\sqrt{v})^2 \end{cases} \quad (8)$$

6. An evaluation of the approximation of the p-value

6.1. An empirical evaluation

Let us revisit Table 2. Pearson's chi-squared statistic (without Yates' continuity correction) was easily found to be 8.548 with a p-value of 0.0035. Equation (8) approximates this p-value to be

$$P(\chi^2 > X^2 = 8.548) \approx \left(\frac{1}{10}\right)^{(0.229873(\sqrt{8.548}+0.304059))^2} = 0.004$$

which compares very well with its theoretical value. When Yates' continuity correction is included in the calculation of the chi-squared statistic, this statistic was found to be 6.874 with a p-value of 0.0087. Its approximate p-value compares very well with this theoretical result and is

$$P(\chi^2 > X_Y^2 = 6.874) \approx \left(\frac{1}{10}\right)^{(0.229873(\sqrt{6.874}+0.304059))^2} = 0.0107$$

For both test statistics, the theoretical and approximate p-values still leads to the same conclusion concerning the test of association between the variables of hydronephrosis and gender. Note that by making use of the approximation given by (8), these two p-values can be obtained without resorting to statistical packages and can be easily calculated using a hand held calculator.

In this illustrative example, the degrees of freedom is very small; $v = 1$. It also highlights the how the approximation works in a single case. A more thorough examination of the approximation with the p-value is now discussed.

6.2. A computational evaluation

To assess the accuracy of (8), we approximate the p-value for a series of chi-squared statistics ranging from 0.5 to 50 in intervals of 0.5. Doing so gives a series of 100 chi-squared statistics. We shall be comparing these p-value approximations with the theoretical p-values calculated from R using the `pchisq()` function which utilises the algorithm of Ding (1992). The degrees of freedom we shall consider will be 1, 5, 10 and 20. Of course, in situations where there are a large number of degrees of freedom under consideration simulation studies of the type described in this section can be easily extended.

Suppose we consider the case where the degrees of freedom, v , is 1 and 5. A plot of the approximate p-value, using (8), against its theoretic p-value is given in Figure 1.

In Figure 1a) we can see that there is a very good linear comparison of the two p-values for $v = 1$. Although, in this case, the approximate values overestimate the theoretic p-values, however this occurs only when the theoretic and approximate p-values are large (exceeding about 0.1). The vertical and horizontal dashed lines coincide with a nominal level of significance of 0.05. We can see that, at this level, only one of the approximated p-values (of the 92) would lead to a different conclusion. This is where the chi-squared statistic is 4; the theoretic p-value is 0.0455 and the approximate p-value is 0.0601. However, even when performing a test at the 5% level of significance, obtaining either p-value does not provide compelling evidence for or against the null hypothesis under consideration.

Suppose we now consider the case where the degrees of freedom is $v = 5$. Figure 1b) provides a comparison of the approximate and theoretic p-values. Like we saw with Figure 1a), the approximate p-values are exceptionally accurate for small p-values; for p-values between 0.001 and 0.05 the accuracy of (8) is between 0.2% and 0.7% of the theoretic p-value – practically negligible.

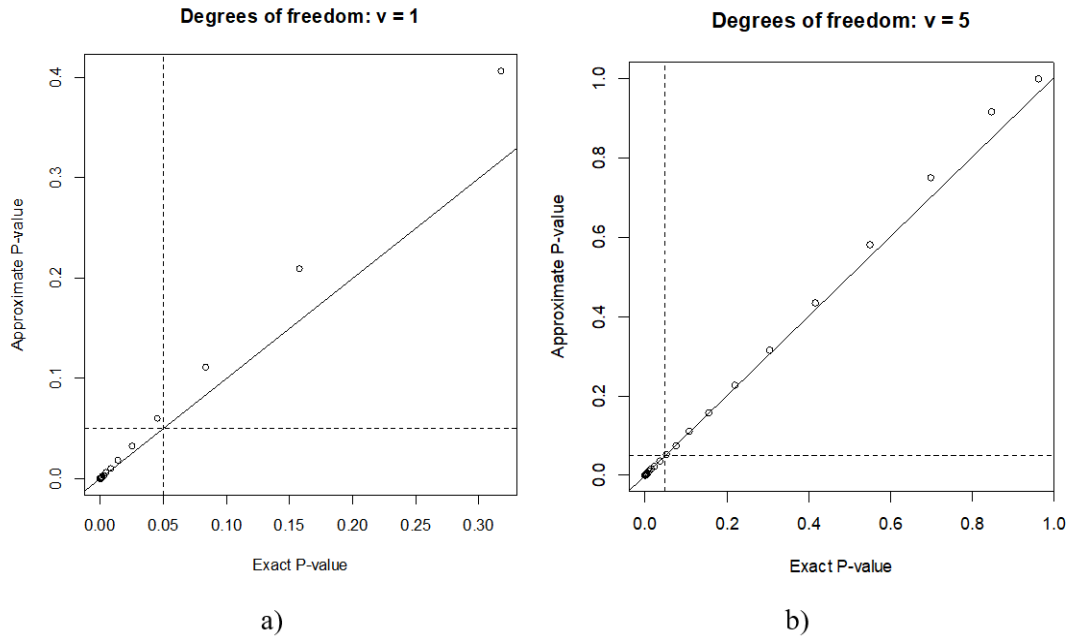


Figure 1: Approximate versus theoretic p-values. Approximate p-values are calculated from (8) and the theoretic p-value is calculated using R. Figures 1a) and 1b) show this comparison for 1 degree of freedom and 5 degrees of freedom, respectively.

Suppose we now consider Figure 2a) which compares the theoretic p-value with the approximated p-value of (8) for 10 degrees of freedom – for theoretic p-values of between 0.001 and 0.05 the approximation was accurate to within 0.5% and 2%. At the top end (where the true p-value exceeded 0.996), there were only 5 cases where the level of accuracy exceeded 2.5%. Similarly, Figure 2b) compares the theoretic p-value with the approximate p-value and shows that, for theoretic p-values between 0.001 and 0.05, their approximations were in error of between 0.01% and 3.8%.

To monitor the accuracy of the approximation (8) as the chi-squared statistic increases, Figure's 3, 4, 5 and 6 plot the chi-squared statistic versus its theoretic and approximate p-value, for 1, 5, 10 and 20 degrees of freedom, respectively. The solid line in each figure summarises the chi-squared statistic versus the theoretic p-value, while the dotted line shows the chi-squared statistic versus the approximate p-value. We can see in all cases that the approximation is extremely good for all degrees of freedom, especially when X^2 exceeds 15 – 20.

Figure 3 compares the theoretic and approximate p-values of the chi-squared statistics ranging from 0.5 and 100 for 1 degree of freedom. We can see here that in all cases the approximation (8) is extremely accurate, especially for chi-squared statistics exceeding 10. For those chi-squared statistics that are less than 10, the error between the theoretic and approximate p-values does not exceed about 33%. As an example of this, the worst case scenario here occurs when $X^2 = 3.5$. The theoretic p-value of this statistic, for 1 degree of freedom, is 0.0614 while the approximate p-value is 0.0817. While the difference may seem large, the approximate p-value does not lead to a contradictory conclusion about a hypothesis test when compared with the theoretic p-values, even if a level of significance of 1% or 5% were considered.

A similar demonstration of the accuracy of (8) can also be made for the case when we have 5 degrees of freedom; see Figure 4. For chi-squared values of between 10 and 30, the p-value approximation is less than 2% in error of the theoretic p-value. For chi-squared values less than 10, the error does not exceed 8% although in these cases, the theoretic and approximate p-value exceed 0.05. For chi-squared statistics that exceed 50, the error between the theoretic and approximate p-value increases. However for such values, the approximate and theoretic p-values are effectively zero until the 7th decimal place.

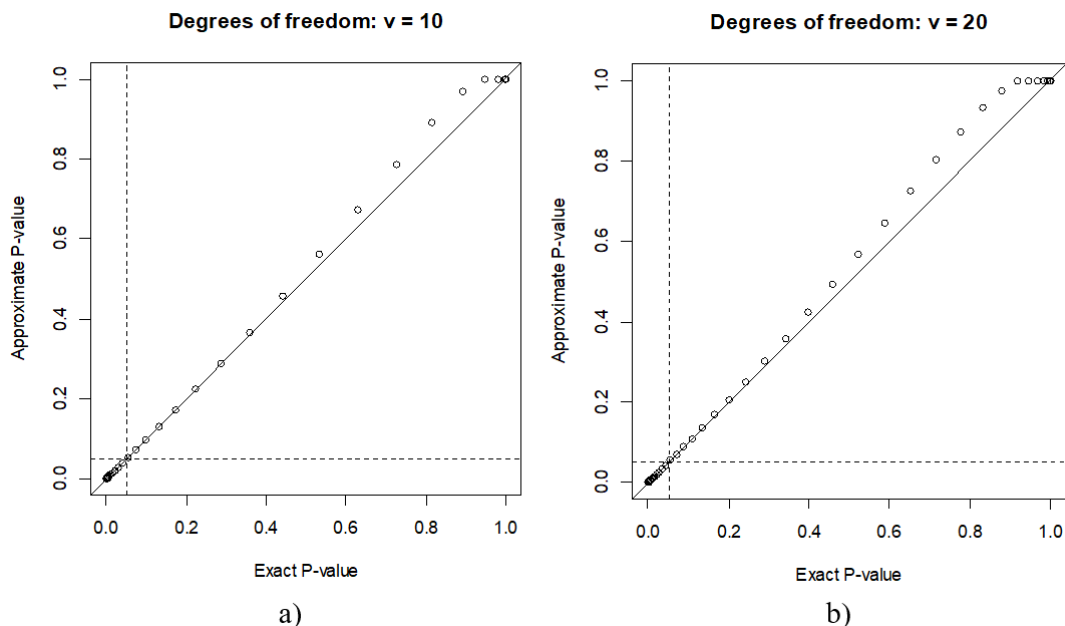


Figure 2: Approximate versus theoretic p-values. Approximate p-values are calculated from (8) and the theoretic p-value is calculated using R. Figures 2a) and 1b) show this comparison for 10 degree of freedom and 20 degrees of freedom, respectively.

Suppose we now consider Figure 5; it compares the theoretic and approximate p-values as the chi-squared statistic ranges between 0.5 and 100 for 10 degrees of freedom. For chi-squared statistics between 10 and 30 the approximate p-value is within 5% of the theoretic p-value. This reduces down to less than 2% for chi-squared statistics ranging between 12 and 35 (inclusive). For chi-squared statistics less than 10, the most extreme difference occurs when X^2 is 5.5; the theoretic p-value is 0.8554 and the approximate p-value is 0.9346. For those chi-squared values greater than 40, the relative error between the theoretic and approximate p-values exceeds 10%. However, in these cases both p-values are effectively zero to the 5th decimal place.

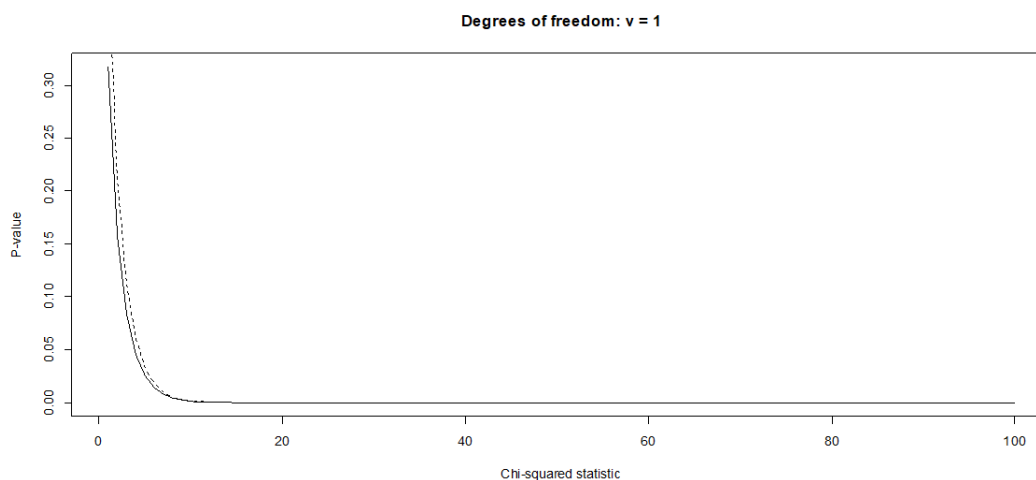


Figure 3: Plot of the chi-squared statistic versus its p-value for 1 degree of freedom. The solid line is the theoretic p-value, the dashed line is approximate p-value using (8).

For 20 degrees of freedom, we obtain approximations of the chi-squared p-value that are just as accurate as for those degrees of freedom that are smaller; see Figure 6. For $X^2 < 50$. The

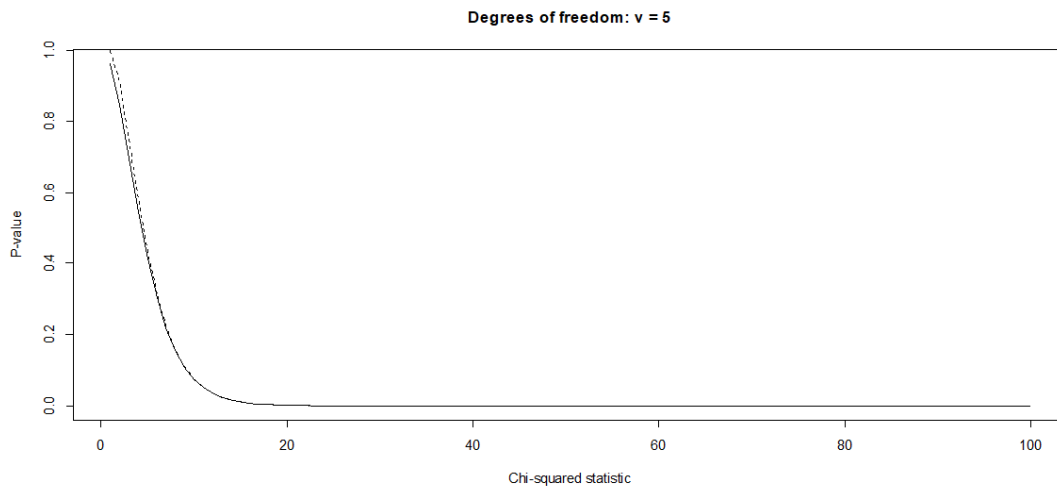


Figure 4: Plot of the chi-squared statistic versus its p-value for 5 degrees of freedom. The solid line is the theoretic p-value, the dashed line is approximate p-value using (8).

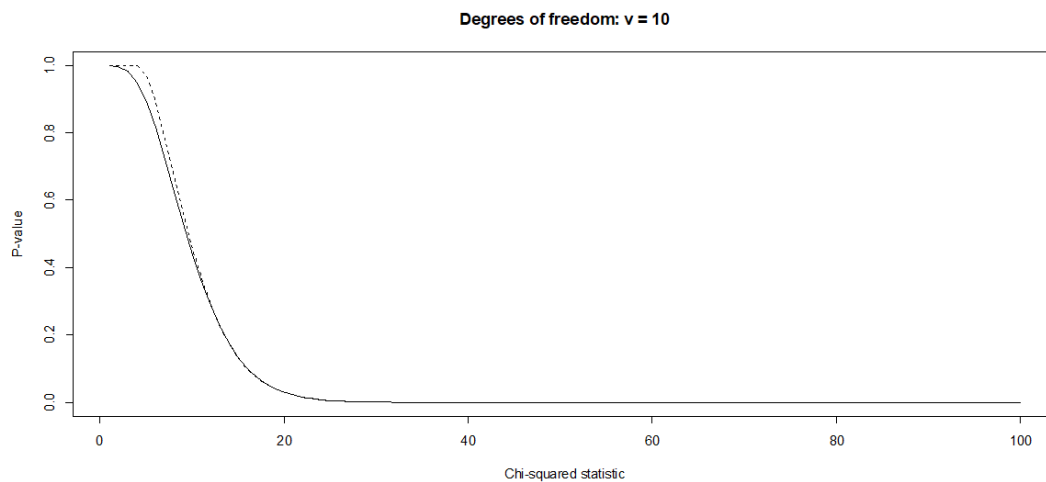


Figure 5: Plot of the chi-squared statistic versus its p-value for 10 degrees of freedom. The solid line is the theoretic p-value, the dashed line is approximate p-value using (8).

largest relative difference in the p-values arises when $X^2 = 14.5$ and is around 12%. In this case the theoretic p-value is 0.8043 and the approximate p-value is 0.9037. For chi-squared statistics of between 25 and 40, the approximate p-value is within 2% of the theoretic p-value. For $X^2 > 50$, the largest difference in the approximate p-value, when compared with its theoretic value, is about 50%. However, in this case both p-values are zero to the fourth decimal place. For values of $X^2 > 60$, these theoretic and approximate p-values are zero to the sixth decimal place. Therefore, for all values of X^2 here, the approximate p-values provide a very accurate and practical approximation of the chi-squared p-values.

7. Discussion

This paper has derived and explored the suitability and applicability of a very simple approximation of the p-value for a chi-squared statistic given its degrees of freedom; see equation (8). This approximation is based on the approximation of (the square root of) the chi-squared statistic that Hoaglin (1977) presented. We have shown that, for chi-squared values ranging from 0.5 to 50, the approximations are very accurate (generally within 5%) of the theoretic p-values. This level of accuracy is comparable to the accuracy that Hoaglin found for his

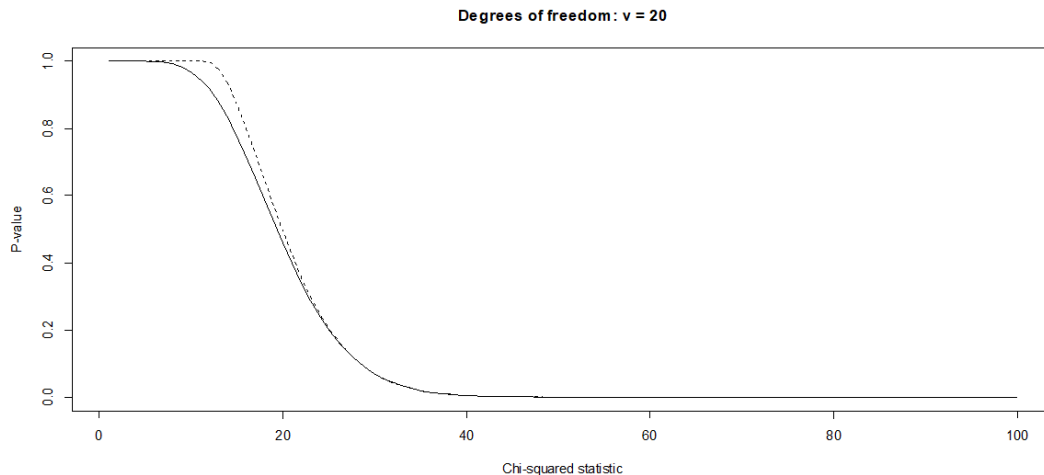


Figure 6: Plot of the chi-squared statistic versus its p-value for 20 degrees of freedom. The solid line is the theoretic p-value, the dashed line is approximate p-value using (8).

approximations of X^2 . Approximations exceed this level when both the theoretic and approximate p-values are very large (exceeding 0.8). The approximation (8) works especially well when for all large chi-squared values irrespective of the degrees of freedom considered – while we have not documented further degrees of freedom (for the sake of brevity) such conclusions apply. In all cases, the approximations work extremely well for small p-values not exceeding 0.2. Such results are therefore promising since rarely are conclusions from statistical inferences made using p-values that exceed 0.1. Further improvements on the approximation can be made by considering improved estimators of the Hoaglin’s constants (a , b , c and d of (4)) but this can be investigated at a later date. Further research can explore more comprehensively Hoaglin’s approximation of the chi-squared statistic and the p-value approximation for all degrees of freedom.

The purpose of this paper is not to provide an alternative to the conventional, and computational, means of determining the p-value of a chi-squared statistic. Instead it provides the analyst with a simple means of approximating the p-value using, something as technologically simple as a standard desk calculator. Of course, when more precise calculations of a chi-squared p-value are needed, one can use the tools found in many of the programming packages that are available to evaluate such values.

References

- Aroian LA (1943). “A New Approximation to the Levels of Significance of the Chi-square Distribution.” *Annals of Mathematical Statistics*, **14**, 93 – 95.
- Beh EJ, Lombardo R (2014). *Correspondence Analysis: Theory, Practice and New Strategies*. Wiley, Chichester.
- 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.
- Cressie N, Read TRC (1984). “Multinomial Goodness-of-Fit Tests.” *Journal of the Royal Statistical Society, Series B*, **46**, 440 – 464.

- Delucchi KL (1983). "The Use and Misuse of Chi-square: Lewis and Burke Revisited." *Psychological Bulletin*, **94**, 166 – 176.
- Dijkers MP (2005). "Misuse of the Pearson Chi-Square Test of Association." *Archives of Physical Medicine and Rehabilitation*, **86**, 602.
- Ding CG (1992). "Algorithm AS275: Computing the Non-central Chi-squared Distribution Function." *Applied Statistics*, **41**, 478 – 482.
- Edwards AL (1950). "On 'The Use and Misuse of the Chi-square Test' – the Case of the 2×2 Contingency Table." *Psychological Bulletin*, **47**, 341 – 346.
- Elderton WP (1902). "Tables for Testing the Goodness of Fit of Theory to Observation." *Biometrika*, **1**, 155 – 163.
- Fienberg SE (1979). "The Use of Chi-squared Statistics for Categorical Data Problems." *Journal of the Royal Statistical Society, Series B*, **41**, 54 – 64.
- Fisher RA (1928). *Statistical Methods for Research Workers (2nd ed.)*. Oliver and Boyd.
- Fleiss JL, Levin B, Paik MC (2003). *Statistical Methods for Rates and Proportions (3rd edn)*. Wiley.
- 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.
- Hald A, Sinkbaek SA (1950). "A Table of Percentage Points of the X^2 Distribution." *Scandinavian Actuarial Journal*, **33**, 168 – 175.
- Haldane JBS (1940). "The Mean and Variance of χ^2 when Used as a Test of Homogeneity, when Expectations are Small." *Biometrika*, **31**, 346 – 355.
- Harter HL (1964). "New Tables of the Incomplete Gamma-Function Ratio and the Percentage Points of the Chi-square and Beta Distributions."
- Heyworth MR (1976). "Approximation to Chi-square." *The American Statistician*, **30**, 204.
- Hoaglin DC (1977). "Approximations for Chi-squared Percentage Points." *Journal of the American Statistical Association*, **72**, 508 – 515.
- Khamis S, Rudert W (1965). *Tables of the Incomplete Gamma Function Ratio: Chi-square Integral, Poisson Distribution*. Justus von Leibig, Darmstad.
- Koehler KJ (1983). "A Simple Approximation for the Percentiles of the t Distribution." *Technometrics*, **25**, 103 – 105.
- Krauth J, Steinebach J (1976). "Extended Tables of the Percentage Points of the Chi-square Distribution for at Most Ten Degrees of Freedom." *Biometrische Zeitschrift*, **18**, 13 – 22.
- Lancaster HO (1969). *The Chi-squared Distribution*. Wiley.
- Lewis D, Burke CJ (1949). "The Use and Misuse of the Chi-square Test." *Psychological Bulletin*, **46**, 433 – 489.
- Lin JT (1988). "Approximating the Cumulative Chi-square Distribution and its Inverse." *The Statistician*, **37**, 3 – 5.
- Lin JT (1990). "Alternative to Koehler's Approximation for the Percentiles of the t Distributions." *Probability in the Engineering and Informational Sciences*, **4**, 535 – 537.
- Merrington M (1941). "Numerical Approximations to the Percentage Points of the χ^2 Distribution." *Biometrika*, **32**, 200 – 202.

- Pearson K (1904). "On the Theory of Contingency and its Relation to Association and Normal Correlation." *Drapers Memoirs, Biometric Series*, **1**.
- Pearson K (1922). *Tables of the Incomplete Gamma-Function*. Cambridge University Press.
- Russell W, Lal M (1969). "Tables of Chi-square Probability Function."
- Terrell GR (1984). "Chi-squared Left-tail Probabilities." *Journal of Statistical Computation and Simulation*, **28**, 264 – 266.
- Thompson CM (1941). "Chi-squared Left-tail Probabilities." *Table of percentage points of the incomplete beta-function distribution*, **32**, 151 – 181.
- Vanderbeck JP, Cooke JR (1961). "Extended Table of Percentage Points of the Chi-Square Distribution."
- Wilson EB, Hilferty MM (1931). "The Distribution of Chi-square." *Proceedings of the National Academy of Sciences (USA)*, **17**, 684 – 688.
- Yates F (1934). "Contingency Tables Involving Small Numbers and the χ^2 Test." *Journal of the Royal Statistical Society, Supplement* **1**, 217 – 235.

Affiliation:

Eric J. Beh

School of Mathematical & Physical Sciences

University of Newcastle Callaghan, NSW, 2308, Australia

E-mail: eric.beh@newcastle.edu.au