A Better Chi-square Goodness-of-Fit Test for Continuous Distributions against a known Alternative.

Wolfgang Rolke[a] and Cristian Gutierrez Gongora[b]

[a]University of Puerto Rico, Mayaguez [b]University of Puerto Rico, Mayaguez

ARTICLE HISTORY
Compiled July 4, 2019

ABSTRACT
The chi square goodness-of-fit test is among the oldest known statistical tests, first proposed by Pearson in 1900 for the multinomial distribution. It has been a mainstay in many fields ever since. However, various studies have shown that when applied to continuous data it is generally inferior to other methods such as the Kolmogorov-Smirnov test. The performance, aka power, of the chi square test depends crucially on the way the data is binned. In this paper we describe a method that automatically finds a binning that is very good against a specific alternative. We show that then the chi square test is generally competitive and sometimes superior to other standard tests.

KEYWORDS
Kolmogorov-Smirnov; Anderson-Darling; Power; Monte Carlo Simulation

1. Introduction

A goodness-of-fit test is concerned with the question whether a data set may have been generated by a certain distribution. It has a null distribution of the form $H_0 : F = F_0$, where F_0 is a probability distribution. For example, one might wish to test whether a data set comes from a standard normal distribution. An obvious and often more useful extension is to the case where F_0 is a family of distributions but without specifying the parameters. So one might wish to test whether a data set comes from a normal distribution but without specifying the mean and standard deviation. If the null distribution is fully specified we have a simple hypothesis whereas if parameters are to be estimated it is called a composite hypothesis.

As described above a goodness-of-fit test is a hypothesis test in the Fisherian sense of testing whether the data is in agreement with a model. The main issue with this approach is that it does not allow us to decide which of two testing methods is better. To solve this problem Neyman and Pearson in the 1930s introduced the concept of an alternative hypothesis, and most tests done today follow more closely the Neyman-Pearson description, although they often are a hybrid of both. The original Fisherian test survives mostly in the goodness-of-fit problem, because here the obvious alternative is $H_0 : F \neq F_0$, a space so huge as to be useless for power calculations.

However, in almost all discussions of goodness-of-fit testing there is an immediate

CONTACT Wolfgang Rolke Email: wolfgang.rolke@upr.edu

pivot and a specific alternative is introduced. This is a necessary step in order to be able to say anything regarding the performance of a test. In our method we have taken advantage of this, and describe a way to find a good binning from among a large number of possible binnings. We will show that if such a binning is used the chi-square test can by quite competitive with other tests, and sometimes even better.

The goodness-of-fit test is one of the most studied problems in Statistics. For an introduction to Statistics and hypothesis testing in general see (Casella and Berger 2002). For discussions of the many goodness-of-fit tests available see (Raynor, Thas, and Best 2012), (D'Agostini and Stephens 1986) and (Thas 2010). Chi-square tests are the subject of (Greenwood and Nikulin 1996) and (Voinov and Nikulin 2013). (Thas 2010) has an extensive list of references on the subject.

2. Chi-square test

The original test by Pearson was designed to see whether an observed set of counts O was in agreement with a multinomial distribution with parameters $m, p_1, .., p_k$. This is done by calculating the expected counts $E_i = mp_i$ and the test statistic $X = \sum(O - E)^2/E$. Pearson showed that under mild conditions $X \sim \chi^2(k-1)$, a chi-square distribution with $k-1$ degrees of freedom. The test therefore rejects the null at the α level if X is larger than the $(1-\alpha)100\%$ quantile of said chi-square distribution. Later work showed that this test works well as long as the expected counts are not to small. $E > 5$ is often suggested although it has been shown that it can still work well even if some expected counts are smaller. In the context of goodness-of-fit testing for a continuous distribution it is generally possible to insure $E > 5$ for all bins, and we will do so in all that follows.

If the test is to be applied to a continuous distribution this distribution has to be discretized by defining a set of bins. In principle any binning (subject to $E > 5$) will work. Two standard methods often used are bins of equal size and bins with equal probabilities under the null hypothesis.

Among Statisticians the use of the chi-square test for continuous distributions has been discouraged for a long time. This is due to its lack of power when compared to other tests. However, it is still the go-to test in many applied fields, and this will be the case for a long time to come. The chi-square test does in fact have one advantage over most other test, namely that it deals with parameter estimation very easily. Fisher (Fisher 1922) showed that if the parameters are estimated by minimizing the chi-square statistic, X again has a chi-square distribution, now with $k-1-m$ degrees of freedom where m is the number of parameters estimated. In fact Fisher coined the term "degrees of freedom" in this seminal paper.

By contrast other standard tests do not easily generalize to allow parameter estimation, and usually the null distribution has to be found via simulation.

A number of alternative test statistics have been proposed, all of which also lead to a limiting chi-square distribution. For a survey of such statistics see (D'Agostini and Stephens 1986). In this paper we use only the classic Pearson formula, but the accompanying R routines allow a choice of several others. Their advantage usually lies in a faster convergence to the limiting chi-square distribution. We will however assure that this convergence is achieved by always requiring E¿5/.

3. Parameter estimation

If the test has a composite null hypothesis it is necessary to estimate the parameters from the data. In practice this is often done via maximum likelihood estimation using the unbinned data. That this leads to a test that is anti-conservative was shown long ago by Fisher (Fisher 1922). As an example consider the following case: We wish to test whether the data comes from a normal mixture, that is

$$F_0 = \lambda N(\mu_1, \sigma_1) + (1 - \lambda)N(\mu_2, \sigma_2) \tag{1}$$

and the parameters $\lambda, \mu_1, \sigma_1, \mu_2, \sigma_2$ are to be estimated. For our simulation we will generate 1000 observations with $\lambda = \frac{1}{3}, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 5, \sigma_2 = 2$ and use unbinned maximum likelihood to estimate the parameters. Then we apply the chi-square test with 10 equal probability bins at the 5% level. This is repeated 10000 times. We find a true type I error rate of over 9%, almost double the nominal one of 5%.

This simulation as well as all others discussed in this paper where done using R. An Rmarkdown file with all calculations as well as an R library with all routines is available at https://github.com/WolfgangRolke/betterchisquare.

Instead of maximum likelihood with the unbinned data one can use either maximum likelihood with binned data or a method called minimum chisquare. This is just what the name suggests, find the set of values of the parameters that minimizes the chi-square statistic. Using this estimation method in the above simulation yields a correct type I error rate of 5%. The same is true in every simulation study we performed.

This method of estimation has another advantage: by the way it is defined it is clear that if the null hypothesis is rejected, it would also be rejected for any other set of parameter values. Also, (Berkson 1980) argues in favor of minimum chi-square.

4. Other tests

Many other goodness-of-fit methods have been developed over time. The most commonly used are the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests, and we will employ these for comparison. For the case where parameters need to be estimated we will use maximum likelihood estimation and simulation to find the null distributions of these tests. For discussions of the relative merits of these tests see (Goodman 1954), (Birnbaum 1952), (Thas 2010) and (Massey 1951).

5. Binning

5.1. Bin types

As stated above, the two standard binning methods are bins of equal size and bins with equal probabilities. In the case of equal size bins often some adjustment is necessary to insure that all expected counts are at least 5.

Let us say that it was decided to use k bins, and the equal probability method yielded bins with endpoints B_i^0 whereas the equal size bins have endpoints B_i^1, $i = 0, .., k$. Here possibly $B_0^0 = B_0^1 = -\infty$ and $B_k^0 = B_k^1 = \infty$. We define a new set of bins by interpolating between these, so a κ bin set has endpoints

$$B_i^\kappa = (1-\kappa)B_i^0 + \kappa B_i^1 \tag{2}$$

for any $0 \leq \kappa \leq 1$. Note that this includes equal probability bins ($\kappa = 0$) as well as equal size bins ($\kappa = 1$).

5.2. Number of bins

Many studies in the past have focused on the number of bins to use. A default formula used in many software programs is $k = 1 + \log_2(n)$, where n is the sample size. One of the first suggestions was $k = 4\left[\frac{2(n-1)^2}{c^2}\right]^{\frac{1}{5}}$ (Mann and Wald 1942). Others can be found in (Williams 1950), (Koehler and Gann 1990), (Oosterhoff 1985), (Dahiya and Gurland 1973), (Kallenberg, Oosterhoff, and Schriever 1985), (Kallenberg 1985), (Quine and Robinson 1985), (Bogdan 1995), (Harrison 1985) and (D'Agostini and Stephens 1986). (Mineo 1979) suggests a different binning scheme.

All of these suggestions have in common that that the number of bins increases as the sample size increases. However, a simple example shows that this need not be true. Say we wish to test $H_0 : F = U[0,1]$ vs $H_1 : F =$Linear(slope=0.2). Here under the alternative the density is $f(x) = 0.4x + 0.8$; $0 < x < 1$. In this case equal size and equal probability bins are the same. We run simulations for the cases of 2 to 21 bins and sample sizes 100, 250, 500 100 and 2000. The resulting powers are shown in figure 1. The power increases as the sample size increases but in each case just two bins yields the highest power.

5.3. Finding the optimal bin set

Optimal in the context of hypothesis testing always means having the highest power, and the power of a test can only be found when an alternative is specified. So let us consider the following problem: we wish to test $H_0 : F = F_0$ vs $H_1 : F = F_1$. A standard way to estimate the power would proceed as follows: generate a data set of the desired size from F_1, apply the test to the data and see whether it rejects the null at the desired level. Repeat many times, and the percentage of rejections is the power of the test.

In the case of a chi-square test this means one has to find the bin counts for the generated data set. However, we already know that these counts have a multinomial distribution with probabilities $p_i = F_1(B_{i+1}^\kappa) - F_1(B_i^\kappa)$. We can therefore run the simulation by generating variates from a multinomial distribution with these probabilities directly. In our simulation studies we have seen a speedup on the order of 100 and more using this approach. An additional advantage is that we need not be able to generate variates from any F_1 directly.

There is another issue when trying to find an optimal binning. How are we to choose between two binnings that both have a power of 1? Moreover, using (say) k and k+1 bins will often result in tests with almost equal power, well within simulation error. To always find a single best binning (from among our sets) we will use the idea of a *perfect* data set. This is an artificial data set that has its observations at the exact right spot, under the alternative. These of course are simply the quantiles of the distribution under the alternative hypothesis. Applying the test to this perfect data set and using

k, κ bins we will use as a figure of merit $M_{k,\kappa} = X/\text{qchisq}(0.95, k-1-m)$, where X is the value of the test statistic and $\text{qchisq}(0.95, k-1-m)$ is the 95% critical value of a chisquare test. 95% was chosen here not because we wish to test at the 5% level but to account (roughly) for the increase in the critical values.

The idea here is simple: the binning that yields the highest value of $M_{k,\kappa}$ would be best in rejecting the null hypothesis for the perfect data set at the 95% level, and one would expect this to be quite good for testing a random data set.

Let us apply this idea to the following case: we wish to test $H_0 : F = \text{Linear}(\text{slope} = -0.5)$ vs $H_a : F =$ truncExp(1), an exponential rate 1 truncated to [0, 1].

Assuming a sample size of 10000, in the left pane of figure 2 we have $M_{k,\kappa}$ for $k = 2, .., 21$ and $\kappa = 0, 0.25, 0.5, 0.75$ and 1. The highest value is achieved for $k = 4, \kappa = 1$. In the right pane we see the actual powers, and indeed (within simulation error) $k = 4, \kappa = 1$ is best.

So our test proceeds as follows: it searches through a grid of number and type of bins. By default these are the 2+number of estimated parameters to $2(1+\log_2(n))$ and $\kappa = 0, 0.25, 0.5, 0.75, 1$, but these can be adjusted by the user. It finds the combination that maximizes $M_{k,\kappa}$ and applies the chi-square test with these bins to the data. Therefore the choice of number of bins and type of bins is automatic.

6. Other circumstances

Our routines also allow for two situations sometimes encountered in practice:

6.1. Already binned data

In some fields it is common that the data, although coming from a continuous distribution, is already binned. This is typically the case, for example, in high energy physics experiments because of finite detector resolution. If so our routine finds the optimal binning as described above and then finds the combination of the data bins that comes closest to the optimal one.

Binned data also causes issues with the Kolmogorov-Smirnov and the Anderson-Darling tests, neither of which works well in the presence of ties. In order to be able to include those tests our routine spreads out the observations over each bin.

6.2. Random sample size

Another feature often encountered is that the sample size itself is random. This is the case, for example, if the determining factor was the time over which an experiment was run. Our routine allows this if the sample size is a variate from a Poisson distribution with known rate λ, as is often the case.

One consequence of such a random sample size is that the bin counts no longer have a multinomial distribution, but instead are independent Poisson with rates that depend on the bin probabilities and λ. In turn this implies that the chi-square statistic now has $k - m$ instead of $k - m - 1$ degrees of freedom. Another consequence is that for the Kolmogorov-Smirnov and the Anderson-Darling tests the null distribution has to be found via simulation, even if no parameters are estimated.

7. Performance

In this section we will discuss a number of cases. For each we will find the power of the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests. We also include chi-square tests with equal bin sizes (Equal Size) and equal bin probabilities (Equal Prob) and the number of bins found with the often used formula $k = 1 + \log_2(n)$. A binning often done in real life by practitioners uses the data as it would be used to do a histogram, that is with a fairly large number of essentially equal size bins. In the examples below we start with 50 bins but may have a little less after combining them to achieve E>5 (Histogram). Finally we include our new test, which is always highlighted in the graphs to make it easier to compare to all others (Best Bin).

7.1. Normal(0, 1) vs t(df)

We have $H_0 : F = N(0, 1)$ vs $H_a : F = t(df)$. We use a sample size of $n = 1000$ and $B = 10000$ simulation runs. This will be the case for all future simulations as well, unless stated otherwise.

The chi-square test with the new binning has the highest power. It uses 3 bins and $\kappa = 0.75$ regardless of sample size, see figure 3.

In our studies (not shown here) we always also found the powers of the tests directly and verified that the combination of k and κ that maximized $M_{k,\kappa}$ also had the highest power.

7.2. Normal vs t(df)

Here we have $H_0 : F = N$ vs $H_a : F = t(df)$, so now the mean and the standard deviation are estimated. For the KS and AD tests maximum likelihood estimation is used, and the null distribution is found via simulation.

In figure 4 we see that the AD test performs best for a low number of degrees of freedom and the new test is best for larger ones. It uses 5 equal size bins regardless of sample size. Notice that because two parameters are estimated, four bins is the least possible.

7.3. Flat (Uniform) vs Linear

Next we have $H_0 : F = U[0, 1]$ vs $H_a : F = \text{Linear}(s)$.

Here the AD test performs best, but new test performs quite well, and much better than the chi-square tests with more bins. It uses 2 or 3 equal size bins regardless of sample size, see figure 5

7.4. Exponential with normal bump

Here the null hypothesis specifies an exponential and the alternative is a mixture of 90% exponential rate 1 and 10 normal mean 1.5 and the standard deviation varies. The rate of the exponential is estimated.

The AD test performs best but the new test is almost as good. It uses four bins with the κ depending on the alternative as well as the sample size. (figure 6)

7.5. Linear vs Linear+Sign Wave

As our last example we consider the case of a linear density $f_0(x) = 2/3(1 + x)$ vs $f_1(x) = 2/3(1 + x) + \lambda \sin(5\pi x)$. While admittedly a bit artificial, as we see in figure 9 it is an example where any of the chi-square tests beats both KS and AD handily! (figure 7)

8. Computational issues

All the calculations and simulations discussed in this paper were done using R. An R library with all the routines as well as an RMarkdown file with the routines to do all the simulations discussed here is available atis available at https://github.com/WolfgangRolke/betterchisquare.

We also created a R Shiny app running online at http://drrolke.shinyapps.io/betterchisquare that allows the user to upload their data and run the test without knowledge of R.

9. Conclusions

We have presented a new method for binning continuous data for a chi-square test. Our method uses the idea of a perfect data set to quickly search through a large number of binnings and find the one with the highest power against a specified alternative. Our simulation studies show that this method is quite competitive with and sometimes better than either the Kolmogorov-Smirnov and the Anderson Darling tests.

Our results also are of interest if there is no alternative hypothesis. Unlike most published results we find that a small number of bins is generally best, regardless of the sample size. Certainly the practice in many fields to draw a histogram with many bins, and then apply the chi-square test using the same binning leads to a badly underpowered test.

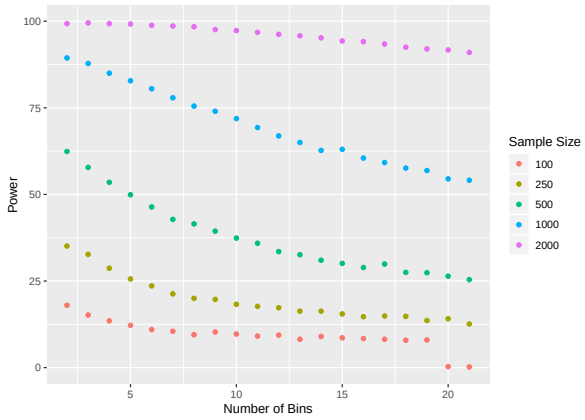
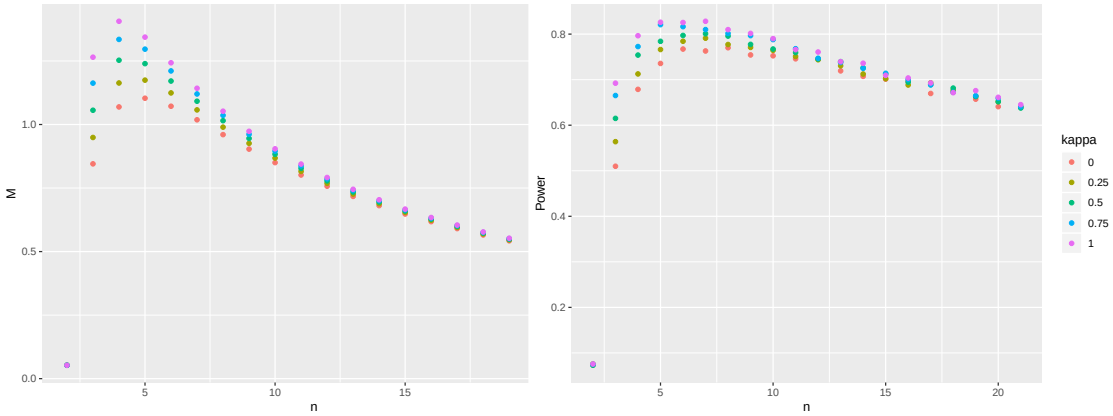


Figure 1. Power of testing uniform vs. linear. While power increases as sample size increases, in each case just two bins yield the highest power.

10. Figures and Tables



(a) The maximum value of $M_{k,\kappa}$ is attained for $k =$ (b) The maximum power is attained for $k = 4, \kappa = 1$
$4, \kappa = 1$

Figure 2. $M_{k,\kappa}$ and Power of a test of linear vs truncated exponential. Both are maximized at $k = 4, \kappa = 1$ (within simulation error)

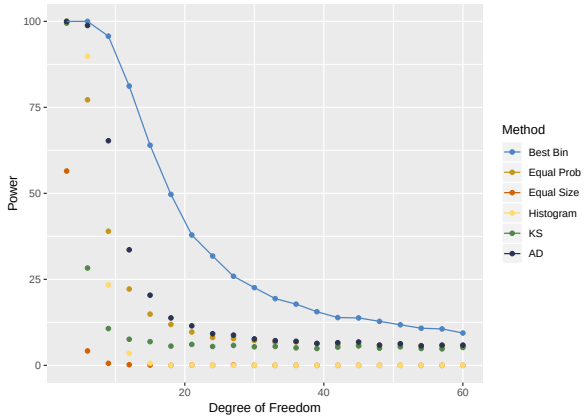


Figure 3. Powers of tests for $H_0 : F_0 = N(0, 1)$ vs $H_a : F_1 = t(df)$. Chi-square test with new binning scheme is best.

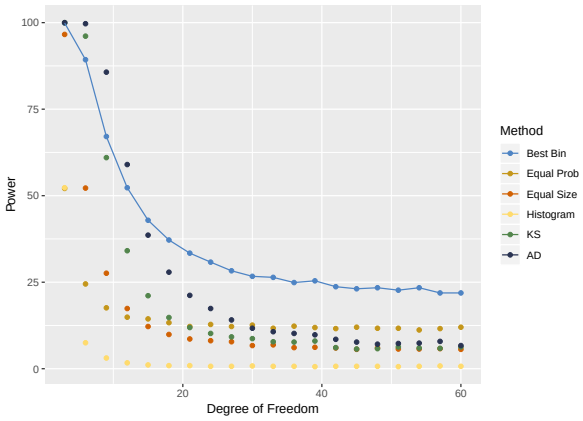


Figure 4. Powers of tests for H_0 : F_0 =Normal vs H_a : $F_1 = t(df)$. Mean and standard deviation are estimated from data. For low degrees of freedom AD is best, for higher df's Chi-square test with new binning scheme is best.

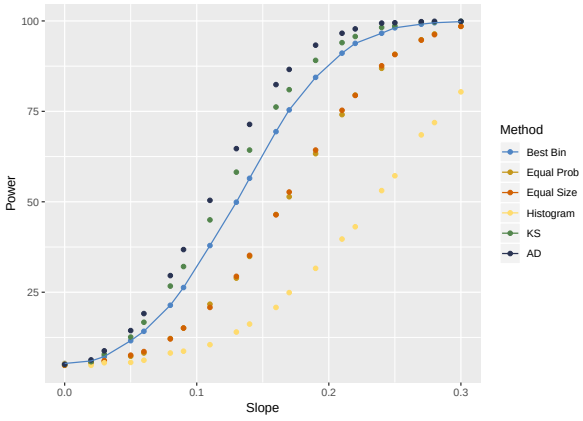


Figure 5. Powers of tests for $H_0 : F_0 = U[0,1]$ vs $H_a : F_1$ =Linear. AD is best but Chi-square test with new binning scheme is almost as good.

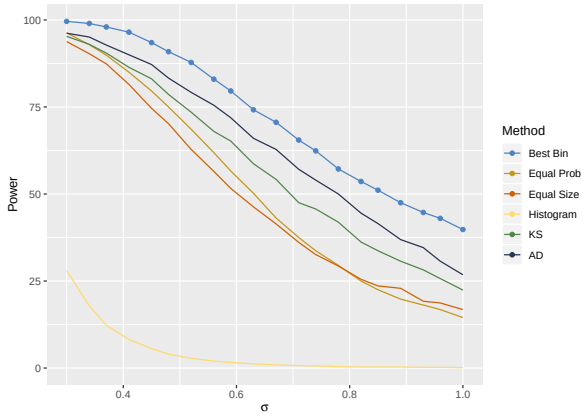


Figure 6. Powers of tests for $H_0 : F_0 = Exp(1)$ vs exponential with a normal bump. AD is best but Chi-square test with new binning scheme is almost as good.

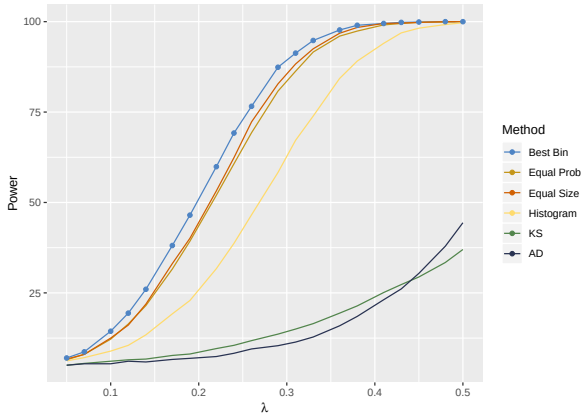


Figure 7. Powers of tests for $H_0 : F_0$ =Linear vs linear plus a sign wave. Chi-square test with new binning scheme is best, and much better than either KS or AD..

References

[1] Berkson J. "Minimum chi-square, not maximum likelihood." *Ann. Math. Stat.*, 1980. 8(3): 457-487.

[2] Birnbaum Z.W. "Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size." *JASA*, 1952. 47: 425-441.

[3] Bogdan M. "Data Driven Version of Pearson's Chi-Square Test for Uniformity." *Journal of Statistical Computation and Simulation*, 1995 52:217-237.

[4] Casella G., Berger R. *Statistical Inference*. Duxbury Advanced Series in Statistics and Decision Sciences. Thomson Learning, 2002.

[5] D'Agostini R.B, Stephens M.A *Goodness-of-Fit Techniques*. Statistics: Textbooks and Monographs. Marcel Dekker, 1986.

[6] Dahiya R.C., Gurland J. "How Many Classes in the Pearson Chi-Square Test?" *Journal of the American Statistical Association*, 1973 68:707-712.

[7] Fisher R.A. "On the Interpretation of Chi-Square of Contingency Tables and the Calculation of P." *Journal of the Royal Statistical Society*, 1922 85.

[8] Greenwood P.E., Nikulin M.S. *A Guide to Chi-Square Testing*, Wiley, 1996.

[9] Goodman L.A. "Kolmogorov-Smirnov Testsfor Psychological Research." *Psychological Bull*, 1954 51: 160-168.

[10] Harrison R.H. "Choosing the Optimum Number of Classes in the Chi-Square Test for Arbitrary Power Levels." *Indian J. Stat.*, 1985 47(3):319-324

[11] Kallenberg W. "On Moderate and Large Deviations in Multinomial Distributions." *The Annals of Statistics*, 1985 13:1554–1580.

[12] Kallenberg W., Ooosterhoff J., and Schriever B. "The Number of Classes in Chi-Squared Goodness-of-Fit Tests." *Journal of the American Statistical Association*, 1985 80:959–968.

[13] Koehler K., Gann F. "Chi-Squared Goodness-of-Fit Tests: Cell Selection and Power." *Communications in Statistics-Simulation*, 1990 19:1265-1278.

[14] Mann H., Wald A. "On the Choice of the Number and Width of Classes for the Chi-Square Test of Goodness of Fit." *Annals of Mathematical Statistics*, 1942 13:306-317.

[15] Massey F.J. "The Kolmogorov-Smirnov Test for Goodness-of-Fit." *JASA*, 1951. 46: 68-78.

[16] Mineo A. "A New Grouping Method for the Right Evaluation of the Chi- Square Test of Goodness-of-Fit." *Scand. J. Stat.*, 1979 6(4):145-153.

[17] Ooosterhoff J. "The Choice of Cells in Chi-Square Tests." *Statistica Neerlandia*, 1985 39:115-128.

[18] Quine M., Robinson J. "Efficiencies of Chi-Square and Likelihoodratio Goodness-of-Fit Tests." *Annals of Statistics*, 1985 13: 727-742.

[19] Raynor J.C., Thas O., and Best D.J., *Smooth Tests of Goodness of Fit.*, 2012.

[20] Thas O *Continuous Distributions*. Springer Series in Statistics. Springer, 2010

[21] Voinov N.B., Nikulin M., *Chi-Square Goodness of Fit Test With Applications.*, Academic Press, 2013

[22] Williams CA. "On the Choice of the Number and Width of Classes for the Chi-Square Test of Goodness of Fit." *Journal of the American Statistical Association*, 1950, 45:77-86.