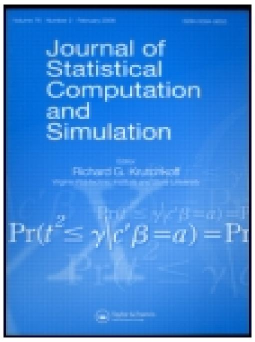

Journal of Statistical Computation and Simulation

Richard G. Krutchkoff

$\Pr(t^2 \leq \gamma | c'\beta = a) = \Pr$

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

The chi-square controversy: what if Pearson had R?

Luisa Canal & Rocco Micciolo

To cite this article: Luisa Canal & Rocco Micciolo (2014) The chi-square controversy: what if Pearson had R?, Journal of Statistical Computation and Simulation, 84:5, 1015-1021, DOI: 10.1080/00949655.2012.737793

To link to this article: https://doi.org/10.1080/00949655.2012.737793

Published online: 01 Nov 2012.

Journal of Statistical Computation and Simulation, 2014
Vol. 84, No. 5, 1015–1021, http://dx.doi.org/10.1080/00949655.2012.737793

The chi-square controversy: what if Pearson had R?

Luisa Canal* and Rocco Micciolo

Department of Cognitive Sciences and Education, University of Trento, Trento, Italy

(*Received 16 March 2012; final version received 4 October 2012*)

One of the most famous controversies in the history of Statistics regards the number of the degrees of freedom of a chi-square test. In 1900, Pearson introduced the chi-square test for goodness of fit without recognizing that the degrees of freedom depend on the number of estimated parameters under the null hypothesis. Yule tried an 'experimental' approach to check the results by a short series of 'experiments'. Nowadays, an open-source language such as R gives the opportunity to empirically check the adequateness of Pearson's arguments. Pearson paid crucial attention to the relative error, which he stated 'will, as a rule, be small'. However, this point is fallacious, as is made evident by the simulations carried out with R. The simulations concentrate on 2 × 2 tables where the fallacy of the argument is most evident. Moreover, this is one of the most employed cases in the research field.

Keywords: chi-square; degrees of freedom; R; simulation

1. Introduction

In a recent article, Hanley *et al.* [1] have wondered what Gosset could do if he could run extensive simulations to check the appropriateness of Student's z, t and s. One of the most famous controversies in the history of Statistics regards the number of the degrees of freedom of a chi-square test, a dispute which involved Pearson and Fisher. Stigler [2] and Baird [3] gave an exhaustive review of this debate.

In 1900, Pearson introduced the chi-square test for goodness of fit [4] without recognizing that the degrees of freedom depend on the number of estimated parameters under the null hypothesis. This error stood undetected for many years. In 1911, Yule proposed a test for comparing two proportions p_1 and p_2 as the ratio between $\hat{p}_1 - \hat{p}_2$ and its standard error under the null hypothesis $H_0{:}p_1 = p_2$ [5]. No mention of Pearson's chi-square test was made by Yule in the textbook where this test was published, even if the inferential problem is quite similar to that addressed by Pearson. In 1915, Greenwood and Yule [6] reported the discrepancy between the results given by the Pearson approach and those given by Yule's test. Finally, Fisher [7,8] definitively acknowledged that additional degrees of freedom have to be deducted for each additional parameter estimated from the sample data.

*Corresponding author. Email: luisa.canal@unitn.it

Although both Pearson and Fisher entirely relied on theoretical mathematical considerations, Yule [9], who was concerned with discrepant results when applying the Pearson approach to the comparison of two binomial counts, tried an ‘experimental’ approach to check the results by a short series of ‘experiments’ devising a number of procedures (based on mechanical devices, described in full detail, that had to guarantee randomness) to run what are now called ‘simulations’. Yule concluded: ‘It is worth noting how clearly, even if we had no theory to guide us, the experiment alone would exhibit the incorrectness of applying the same law in both cases’; the author refers to the comparison of results obtained by computing χ^2 ‘from the expected distribution given by uniformity’ and ‘from the independence values given by the row and column totals’.

Yule was able to carry out only a relatively ‘small’ number of experiments (350 2×2 tables and 100 4×4 tables) and was concerned with situations where ‘the assumption of normality can hardly be justified except as the roughest of approximations’.

Today, we can run much more extensive simulations. Nowadays, reliable pseudo-random number generators, very fast CPU of our PC and powerful, sophisticated programming languages such as R (which allow for a very high degree of interactivity between the (simulated) data and the researcher) can be called into play to give to all of us the opportunity not only to reproduce Yule’s ‘experiments’ employing a much higher number of tables and a much higher number of observations for each table, but also to empirically check the adequateness of Pearson’s arguments.

Our examples concentrate on the case of 2×2 tables, which show the most evident discrepancy between the arguments of the contenders and are also one of the most employed cases in the research field.

Data collected in a 2×2 table can arise from three different sampling models (or experimental designs):

(1) two independent groups with (possibly) equal probability of ‘success’; in this case, only one of the two margins is fixed in advance by the researcher and the appropriate distribution is the binomial;

(2) cross-sectional design where only the overall total number of observations is fixed by the researcher; in this case, none of the total marginals is fixed and the appropriate distribution is the multinomial;

(3) both margins are fixed and the appropriate distribution is the hypergeometric; citing Yates [10], the famous ‘lady tasting tea experiment’ described by Fisher in the *Design of Experiments* [11] ‘provides a classic and somewhat rare example of the generation of 2×2 tables in which both margins are determined in advance’.

We provide an example for each of these three different designs beginning with a short theoretical note following what is reported by Stigler [2].

2. Theoretical note

In his original paper, Pearson [4] considered two different chi-square statistics: one, χ_s^2, based upon the expected frequencies (m_s) estimated from the observed frequencies (m') and the other, χ^2, based upon the expected theoretical frequencies (m) considered as known *a priori*:

$$\chi_s^2 = \sum \frac{(m' - m_s)^2}{m_s} \quad \chi^2 = \sum \frac{(m' - m)^2}{m}.$$

Employing Taylor series expansions about m_s, Pearson, after having discarded terms of the order $(\mu/m_s)^3$ or higher, correctly arrived at

$$\chi^2 - \chi_s^2 = -\sum\left\{\frac{\mu}{m_s}\frac{m'^2 - m_s^2}{m_s}\right\} + \sum\left\{\left(\frac{\mu}{m_s}\right)^2\frac{m'^2}{m_s}\right\}, \tag{1}$$

where $\mu = m - m_s$.

Pearson argued that the difference $\chi^2 - \chi_s^2$ was not large because the two terms on the right-hand side of Equation (1) were both small (eventually the first cancelling out part of the second term) so that χ_s^2 and χ^2 shared the same distribution. However, this line of reasoning is likely to be fallacious.

3. Simulations

Employing the R function *rmultinom* [12], we simulated 100,000 2 × 2 tables from a multinomial distribution with probabilities 3/8, 3/8, 1/8, 1/8; the sample size was 1000 for each table. Therefore, the expected theoretical frequencies (indicated with m in Equation (1)) were always 375, 375, 125, 125. On the other hand, for each simulated table, the expected frequencies (m_s in Equation (1)) were computed from the marginals of the table and the two terms on the right-hand side

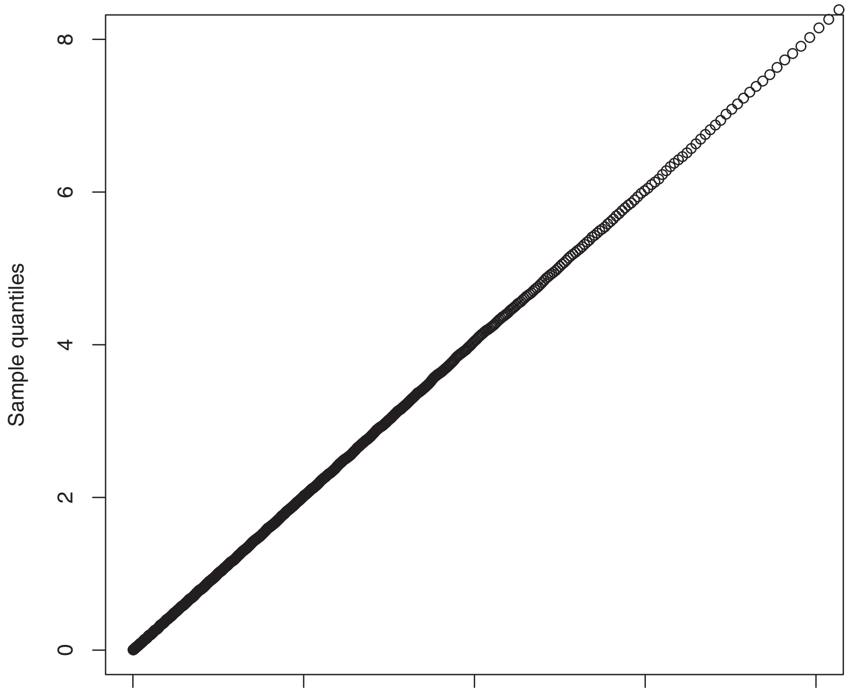

Figure 1. The results of 100,000 simulations (sample size for each simulation = 1000) from a multinomial distribution with probabilities 3/8, 3/8, 1/8, 1/8. Vertical axis: sample quantiles of the second term on the right-hand side of Equation (1). Horizontal axis: theoretical quantiles for a χ_2^2 distribution.

of Equation (1) were also computed. As stated by Pearson, the first term on the right-hand side of Equation (1), averaged on the 100,000 simulated tables, was negligible: −0.00572 (median: −0.00034). The fact that the first term on the right-hand side of Equation (1) is near zero is due to the use of maximum-likelihood estimates for the estimated parameters and the simulation also verifies this.

However, the second term on the right-hand side of Equation (1) was far from being negligible; as a matter of fact, the average was 2.01 (median: 1.40) despite what Pearson expected, and the extremes of the range of the middle 50% of the simulated values were 0.576 and 2.79. The distribution of the 100,000 values was highly skewed and resembled a chi-square distribution with two degrees of freedom (see the quantile–quantile plot in Figure 1) and so it is certainly not negligible. With a computer simulation, this is highly evident.

According to Stigler [2], the main fallacy in the arguments of Pearson was to pay crucial attention to the relative error μ/m_s, which Pearson told 'will, as a rule, be small'. Unfortunately, it was not so small to make the product $(\mu/m_s)^2 \times (m'^2/m_s)$ vanish.

According to Pearson, the chi-square test statistic $\sum (m' - m_s)^2/m_s$, based on the expected frequencies calculated from the marginals of the 2×2 simulated table, would be approximated by a χ^2 distribution with three degrees of freedom. Figure 2 shows the quantile–quantile plot where the theoretical quantiles are those of a χ^2 distribution with three degrees of freedom.

The systematic differences between empirical and theoretical quantiles are quite evident. On the other hand, the theoretical quantiles of a χ^2 distribution with one degree of freedom fit the empirical quantiles well. If Pearson had R, a quick look at Figure 2 would have shown to him that something in Equation (1) was wrong.

Two other simulations were performed. The first 'inspired' from Yule's [9] paper generated 100,000 2×2 tables sampling from two binomials with the same parameters ($n = 500; p = 0.75$) employing the *rbinom* R function and the second sampling from a hypergeometric distribution

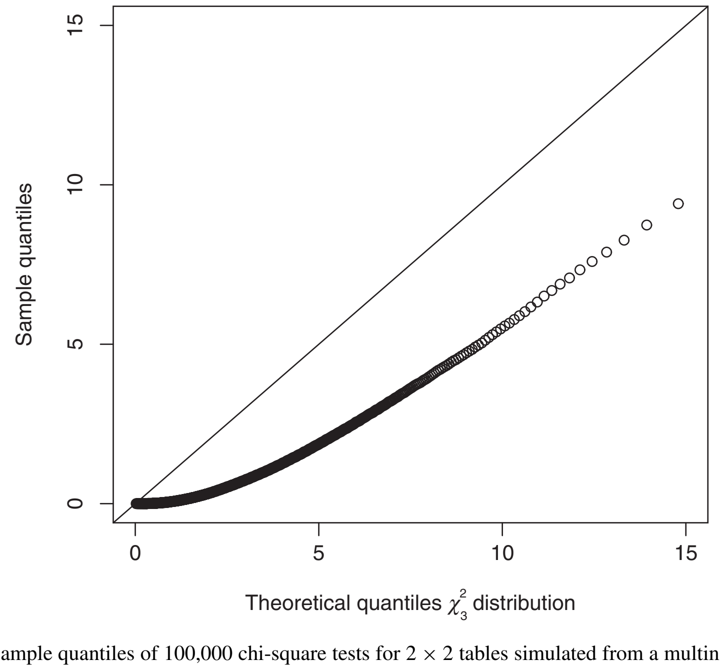

Figure 2. The sample quantiles of 100,000 chi-square tests for 2×2 tables simulated from a multinomial distribution (sample size for each simulation = 1000) with probabilities 3/8, 3/8, 1/8, 1/8 plotted against the theoretical quantiles of a χ^2_3 distribution.

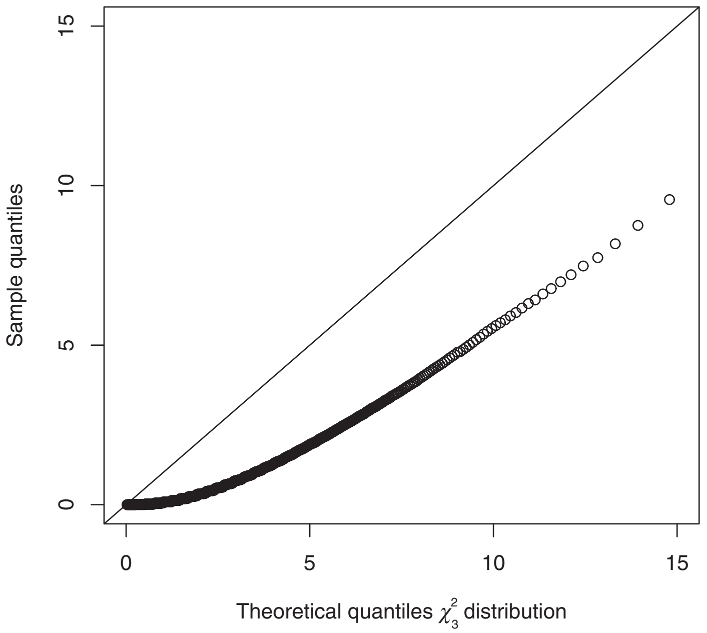

Figure 3. The sample quantiles of 100,000 chi-square tests for 2 × 2 tables simulated from two binomial distributions with parameters $n = 500$, $p = 0.75$ plotted against the theoretical quantiles of a χ^2_3 distribution.

with row marginals fixed at 750 and 250 and column marginals fixed at 500 and 500 employing the *rhyper* R function. For each table, the chi-square statistic was calculated (without the continuity correction), and the sample quantiles were plotted against the theoretical quantiles of a χ^2 distribution with three degrees of freedom. The results are shown in Figures 3 and 4. Obviously, in both cases, the sample and the theoretical quantiles are entirely different.

The R script to run the simulation based on the multinomial sampling scheme is reported below:

```
nrep <-  100000
tot <-  1000 # total number of observations
x <-  rmultinom(nrep,tot,prob=c(3,3,1,1)/8)
out <-  t(x) # a,b,c,d
job <-  function(x) {
    tbl  <-  matrix(x,ncol=2,byrow=TRUE)
    att  <-  chisq.test(tbl)$exp
    cq   <-  sum((tbl-att)^2/att)
    cq
}
pquant <-  qchisq(seq(0.001,0.999,0.001),3)
ris <-  apply(out,1,job)
squant <-  quantile(ris,seq(0.001,0.999,0.001))
plot(pquant,squant,xlab="Theoretical Quantiles",
    ylab="Sample Quantiles")
abline(0,1)
```

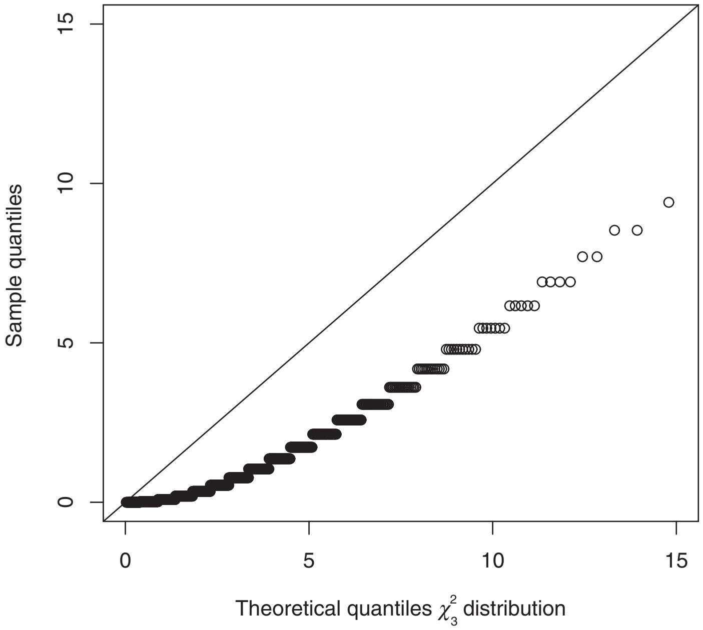

Figure 4. The sample quantiles of 100,000 chi-square tests for 2 × 2 tables simulated from one hypergeometric distribution with row marginals fixed at 750 and 250 and column marginals fixed at 500 and 500 plotted against the theoretical quantiles of a χ^2_3 distribution.

4. Conclusions

In the twenty-first century, where computation is fast and cheap, new computer-based methods can be developed to numerically analyse the properties of statistical theory. Monte Carlo and bootstrap methods are widely employed to replace standard assumptions about data with massive calculations [13]. In this context, the R programming language is approaching the status of the default statistical package for universities and research organizations and is making steady inroads into the commercial market [14]. Besides the wide variety of statistical techniques, R is an object-oriented programming language, implementing quite powerful and computationally fast vectorization algorithms well suited for doing simulations.

If Pearson had R, the use of simulation would have led him to the conclusion that his expectations, concerning formula (1), were not valid. However, the correct number of degrees of freedom was yet to be determined, and we do not know how Pearson would have corrected that error. Fisher did correct the error, and his achievement is a remarkable contribution to the development of statistics. Notwithstanding this, the chi-square test was a very valuable achievement in statistical methodology. In a special issue of *Science* dedicated to the top 20 discoveries since 1900, considering all branches of science and technology that have changed our world, Hacking [15] says that Karl Person's chi-square test ushered 'in a new kind of decision making'.

References

[1] J.A. Hanley, M. Julien, and E.E.M. Moodie, *Student's z, t, and s: What if Gosset had R*? Am. Stat. 62 (2008), pp. 64–69.

[2] S.M. Stigler, *Karl Pearson's theoretical errors and the advances they inspired*, Stat. Sci. 23 (2008), pp. 261–271.

[3] D. Baird, *The Fisher/Pearson chi-squared controversy: A turning point for inductive inference*, Br. J. Philos. Sci. 34 (1983), pp. 105–118.

[4] K. Pearson, *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*, Philos. Magazine, 5th Series 50 (1900), pp. 157–175. Reprinted in Karl Pearson's Early Statistical Papers 339–357, Cambridge University Press, Cambridge, 1948.

[5] G.U. Yule, *An Introduction to the Theory of Statistics*, Griffin, London, 1911

[6] M. Greenwood and G.U. Yule, *The statistics of antityphoid and anti-cholera inoculations, and the interpretation of such statistics in general*, Proc. R. Soc. Med. 8 (1915), pp. 113–190.

[7] R.A. Fisher, *On the interpretation of* χ^2 *from contingency tables, and the calculation of P*, J. R. Stat. Soc. 85 (1922), pp. 87–94

[8] R.A. Fisher, *Conditions under which* χ^2 *measures the discrepancy between observation and hypothesis*, J. R. Stat. Soc. 87 (1924), pp. 442–450.

[9] G.U. Yule, *On the application of the* χ^2 *method to association and contingency tables, with experimental illustrations*, J. R. Stat. Soc. 85 (1922), pp. 95–104.

[10] F. Yates, *Tests of significance for* 2×2 *contingency tables*, J. R. Stat. Soc. A 147 (1984), pp. 426–463.

[11] R.A. Fisher, *The Design of Experiments*, Oliver and Boyd, Edinburgh, 1935.

[12] R Development Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0. Available at http://www.R-project.org/.

[13] P. Diaconis and B. Efron, *Computer-intensive methods in statistics*, Scientific American, May 1983, pp. 116–130.

[14] M. Aitkin, B. Francis, J. Hinde, and R. Darnell, *Statistical Modelling in R*, Oxford University Press, Oxford, 2009.

[15] I. Hacking, *Trial by number*, Science 84 (1984), pp. 69–70.