Computing the confidence levels for a root-mean-square test of goodness-of-fit

William Perkins, Mark Tygert*, Rachel Ward

Courant Institute of Mathematical Sciences, NYU, 251 Mercer St., New York, NY 10012

Abstract

The classic χ^2 statistic for testing goodness-of-fit has long been a cornerstone of modern statistical practice. The statistic consists of a sum in which each summand involves division by the probability associated with the corresponding bin in the distribution being tested for goodness-of-fit. Typically this division should precipitate rebinning to uniformize the probabilities associated with the bins, in order to make the test reasonably powerful. With the now widespread availability of computers, there is no longer any need for this. The present paper provides efficient black-box algorithms for calculating the asymptotic confidence levels of a variant on the classic χ^2 test which omits the problematic division. In many circumstances, it is also feasible to compute the exact confidence levels via Monte Carlo simulation.

Keywords: chi-square, goodness of fit, confidence, significance, test, Euclidean norm

1. Introduction

A basic task in statistics is to ascertain whether a given set of independent and identically distributed (i.i.d.) draws does not come from a specified probability distribution (this specified distribution is known as the "model"). In the present paper, we consider the case in which the draws are discrete random variables, taking values in a finite set. In accordance with the standard terminology, we will refer to the possible values of the discrete random variables as "bins" ("categories," "cells," and "classes" are common synonyms for "bins").

A natural approach to ascertaining whether the i.i.d. draws do not come from the specified probability distribution uses a root-mean-square statistic. To construct this statistic, we estimate the probability distribution over the bins using the given i.i.d. draws, and then measure the root-mean-square difference between this empirical distribution and the specified model distribution; see, for example, [1], page 123 of [2], or Section 2 below. If the draws do in fact arise from the specified model, then with high probability this root-mean-square is not large. Thus, if the root-mean-square statistic is large, then we can be confident that the draws do not arise from the specified probability distribution.

Let us denote by $x_{\rm rms}$ the value of the root-mean-square for the given i.i.d. draws; let us denote by $X_{\rm rms}$ the root-mean-square statistic constructed for different i.i.d. draws that definitely do in fact come from the specified model distribution. Then, the significance level

*Corresponding author.

α is defined to be the probability that $X_{\mathrm{rms}} \geq x_{\mathrm{rms}}$ (viewing X_{rms} — but not x_{rms} — as a random variable). The confidence level that the given i.i.d. draws do not arise from the specified model distribution is the complement of the significance level, namely $1 - \alpha$.

Unfortunately, the confidence levels for the simple root-mean-square statistic are different for different model probability distributions. To avoid this seeming inconvenience (at least asymptotically), one may weight the average in the root-mean-square by the inverses of the model probabilities associated with the various bins, obtaining the classic χ^2 statistic; see, for example, [3] or Remark 2.1 below. However, with the now widespread availability of computers, direct use of the root-mean-square statistic has become feasible (and actually turns out to be very convenient). The present paper provides efficient black-box algorithms for computing the confidence levels for any specified model distribution, in the limit of large numbers of draws. Calculating confidence levels for small numbers of draws via Monte Carlo can also be practical.

The simple statistic described above would seem to be more natural than the standard χ^2 statistic of [3], is typically easier to use (since it does not require any rebinning of data), and is more powerful in many circumstances, as we demonstrate both in Section 6 below and more extensively in a forthcoming paper. Even more powerful is the combination of the root-mean-square statistic and an asymptotically equivalent variation of the χ^2 statistic, such as the (log)likelihood-ratio or "G^2" statistic; the (log)likelihood-ratio and χ^2 statistics are asymptotically equivalent when the draws arise from the model, while the (log)likelihood-ratio can be more powerful than χ^2 for small numbers of draws (see, for example, [1]). The rest of the present article has the following structure: Section 2 details the statistic discussed above, expressing the confidence levels for the associated goodness-of-fit test in a form suitable for computation. Section 3 discusses the most involved part of the computation of the confidence levels, computing the cumulative distribution function of the sum of the squares of independent centered Gaussian random variables. Section 4 summarizes the method for computing the confidence levels of the root-mean-square statistic. Section 5 applies the method to several examples. Section 6 very briefly illustrates the power of the root-mean-square. Section 7 draws some conclusions and proposes directions for further research.

2. The simple statistic

This section details the root-mean-square statistic discussed briefly in Section 1, and determines its probability distribution in the limit of large numbers of draws, assuming that the draws do in fact come from the specified model. The distribution determined in this section yields the confidence levels (in the limit of large numbers of draws): Given a value x for the root-mean-square statistic constructed from i.i.d. draws coming from an unknown distribution, the confidence level that the draws do not come from the specified model is the probability that the root-mean-square statistic is less than x when constructed from i.i.d. draws that do come from the model distribution.

To begin, we set notation and form the statistic X to be analyzed. Given n bins, numbered 1, 2, ..., $n-1$, n, we denote by p_1, p_2, ..., p_{n-1}, p_n the probabilities associated with the respective bins under the specified model; of course, $\sum_{k=1}^{n} p_k = 1$. To obtain a draw conforming to the model, we select at random one of the n bins, with probabilities p_1, p_2, ..., p_{n-1}, p_n. We perform this selection independently m times. For $k = 1, 2, \ldots, n-1, n$,

we denote by Y_k the fraction of times that we choose bin k (that is, Y_k is the number of times that we choose bin k, divided by m); obviously, $\sum_{k=1}^{n} Y_k = 1$. We define X_k to be $\sqrt{m}$ times the difference of Y_k from its expected value, that is,

$$X_k = \sqrt{m}\,(Y_k - p_k) \tag{1}$$

for $k = 1, 2, \ldots, n-1, n$. Finally, we form the statistic

$$X = \sum_{k=1}^{n} X_k^2, \tag{2}$$

and now determine its distribution in the limit of large m. (X is the square of the root-mean-square statistic $\sqrt{\sum_{k=1}^{n}(mY_k - mp_k)^2/m}$. Since the square root is a monotonically increasing function, the confidence levels are the same whether determined via X or via $\sqrt{X}$; for convenience, we focus on X below.)

Remark 2.1. The classic χ^2 test for goodness-of-fit of [3] replaces (2) with the statistic

$$\chi^2 = \sum_{k=1}^{n} \frac{X_k^2}{p_k}, \tag{3}$$

where X_1, X_2, $\ldots$, X_{n-1}, X_n are the same as in (1) and (2). χ^2 defined in (3) has the advantage that its confidence levels are the same for every model distribution, independent of the values of $p_1, p_2, \ldots, p_{n-1}, p_n$, in the limit of large numbers of draws. In contrast, using X defined in (2) requires computing its confidence levels anew for every different model.

The multivariate central limit theorem shows that the joint distribution of X_1, X_2, $\ldots$, X_{n-1}, X_n converges in distribution as $m \to \infty$, with the limiting generalized probability density proportional to

$$\exp\left(-\sum_{k=1}^{n} \frac{x_k^2}{2p_k}\right) \quad \delta\left(\sum_{k=1}^{n} x_k\right), \tag{4}$$

where δ is the Dirac delta; see, for example, [4] or Chapter 25 and Example 15.3 of [5]. The generalized probability density (4) is a centered multivariate Gaussian concentrated on a hyperplane passing through the origin (the hyperplane consists of the points such that $\sum_{k=1}^{n} x_k = 0$); the restriction of the generalized probability density (4) to the hyperplane through the origin is also a centered multivariate Gaussian. Thus, the distribution of X defined in (2) converges as $m \to \infty$ to the distribution of the sum of the squares of $n-1$ independent Gaussian random variables of mean zero whose variances are the variances of the restricted multivariate Gaussian distribution along its principal axes; see, for example, [4] or Chapter 25 of [5]. Given these variances, the following section describes an efficient algorithm for computing the probability that the associated sum of squares is less than any particular value; this probability is the desired confidence level, in the limit of large numbers of draws. See Sections 4 and 5 for further details.

To compute the variances of the restricted multivariate Gaussian distribution along its principal axes, we multiply the diagonal matrix D whose diagonal entries are $1/p_1$, $1/p_2$, $\ldots$,

$1/p_{n-1}$, $1/p_n$ from both the left and the right by the projection matrix P whose entries are

$$P_{j,k} = \begin{cases} 1 - \frac{1}{n}, & j = k \\ -\frac{1}{n}, & j \neq k \end{cases} \tag{5}$$

for $j, k = 1, 2, \ldots, n-1, n$ (upon application to a vector, P projects onto the orthogonal complement of the subspace consisting of every vector whose entries are all the same). The entries of this product $B = PDP$ are

$$B_{j,k} = \begin{cases} \frac{1}{p_k} - \frac{1}{n}\left(\frac{1}{p_j} + \frac{1}{p_k}\right) + \frac{1}{n^2}\sum_{l=1}^{n} \frac{1}{p_l}, & j = k \\ -\frac{1}{n}\left(\frac{1}{p_j} + \frac{1}{p_k}\right) + \frac{1}{n^2}\sum_{l=1}^{n} \frac{1}{p_l}, & j \neq k \end{cases} \tag{6}$$

for $j, k = 1, 2, \ldots, n-1, n$. Clearly, B is self-adjoint. By construction, exactly one of the eigenvalues of B is 0. The other eigenvalues of B are the multiplicative inverses of the desired variances of the restricted multivariate Gaussian distribution along its principal axes.

Remark 2.2. The $n \times n$ matrix B defined in (6) is the sum of a diagonal matrix and a low-rank matrix. The methods of [6, 7] for computing the eigenvalues of such a matrix B require only either $\mathcal{O}(n^2)$ or $\mathcal{O}(n)$ floating-point operations. The $\mathcal{O}(n^2)$ methods of [6, 7] are usually more efficient than the $\mathcal{O}(n)$ method of [7], unless n is impractically large.

Remark 2.3. It is not hard to accommodate homogeneous linear constraints of the form $\sum_{k=1}^{n} c_k\, x_k = 0$ (where $c_1, c_2, \ldots, c_{n-1}, c_n$ are real numbers) in addition to the requirement that $\sum_{k=1}^{n} x_k = 0$. Accounting for any additional constraints is entirely analogous to the procedure detailed above for the particular constraint that $\sum_{k=1}^{n} x_k = 0$. The estimation of parameters from the data in order to specify the model can impose such extra homogeneous linear constraints; see, for example, Chapter 25 of [5]. A detailed treatment is available in [8].

3. The sum of the squares of independent centered Gaussian random variables

This section describes efficient algorithms for evaluating the cumulative distribution function (cdf) of the sum of the squares of independent centered Gaussian random variables. The principal tool is the following theorem, expressing the cdf as an integral suitable for evaluation via quadratures (see, for example, Remark 3.4 below).

Theorem 3.1. *Suppose that n is a positive integer, $X_1, X_2, \ldots, X_{n-1}, X_n$ are i.i.d. Gaussian random variables of zero mean and unit variance, and $\sigma_1, \sigma_2, \ldots, \sigma_{n-1}, \sigma_n$ are positive real numbers. Suppose in addition that X is the random variable*

$$X = \sum_{k=1}^{n} |\sigma_k\, X_k|^2. \tag{7}$$

Then, the cumulative distribution function (cdf) P of X is

$$P(x) = \int_0^\infty \mathrm{Im}\left(\frac{e^{1-t}\, e^{it\sqrt{n}}}{\pi\left(t - \frac{1}{1-i\sqrt{n}}\right) \prod_{k=1}^{n} \sqrt{1 - 2(t-1)\sigma_k^2/x + 2it\sigma_k^2\sqrt{n}/x}} \right) dt \tag{8}$$

for any positive real number x, and $P(x) = 0$ for any nonpositive real number x. The square roots in (8) denote the principal branch, and Im *takes the imaginary part.*

Proof. For any $k = 1, 2, \ldots, n-1, n$, the characteristic function of $|X_k|^2$ is

$$\varphi_1(t) = \frac{1}{\sqrt{1-2it}}, \tag{9}$$

using the principal branch of the square root. By the independence of $X_1, X_2, \ldots, X_{n-1}, X_n$, the characteristic function of the random variable X defined in (7) is therefore

$$\varphi(t) = \prod_{k=1}^{n} \varphi_1(t\sigma_k^2) = \frac{1}{\prod_{k=1}^{n} \sqrt{1-2it\sigma_k^2}}. \tag{10}$$

The probability density function of X is therefore

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx}\, \varphi(t)\, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itx}}{\prod_{k=1}^{n} \sqrt{1-2it\sigma_k^2}}\, dt \tag{11}$$

for any real number x, and the cdf of X is

$$P(x) = \int_{-\infty}^{x} p(y)\, dy = \frac{1}{2} + \frac{i}{2\pi}\, \mathrm{PV} \int_{-\infty}^{\infty} \frac{e^{-itx}}{t\, \prod_{k=1}^{n} \sqrt{1-2it\sigma_k^2}}\, dt \tag{12}$$

for any real number x, where PV denotes the principal value.

It follows from the fact that X is almost surely positive that the cdf $P(x)$ is identically zero for $x \le 0$; there is no need to calculate the cdf for $x \le 0$. Substituting $t \mapsto t/x$ in (12) yields that the cdf is

$$P(x) = \frac{1}{2} + \frac{i}{2\pi}\, \mathrm{PV} \int_{-\infty}^{\infty} \frac{e^{-it}}{t\, \prod_{k=1}^{n} \sqrt{1-2it\sigma_k^2/x}}\, dt \tag{13}$$

for any positive real number x, where again PV denotes the principal value. The branch cuts for the integrand in (13) are all on the lower half of the imaginary axis.

Though the integration in (13) is along $(-\infty, \infty)$, we may shift contours and instead integrate along the rays

$$\left\{(-\sqrt{n} - i)t + i \ : \ t \in (0, \infty)\right\} \tag{14}$$

and

$$\left\{(\sqrt{n} - i)t + i \ : \ t \in (0, \infty)\right\}, \tag{15}$$

obtaining from (13) that

$$P(x) = \frac{i}{2\pi} \int_0^{\infty} \left(\frac{e^{1-t}\, e^{-it\sqrt{n}}}{\left(t - \frac{1}{1+i\sqrt{n}}\right) \prod_{k=1}^{n} \sqrt{1-2(t-1)\sigma_k^2/x - 2it\sigma_k^2\sqrt{n}/x}} - \frac{e^{1-t}\, e^{it\sqrt{n}}}{\left(t - \frac{1}{1-i\sqrt{n}}\right) \prod_{k=1}^{n} \sqrt{1-2(t-1)\sigma_k^2/x + 2it\sigma_k^2\sqrt{n}/x}} \right) dt \tag{16}$$

for any positive real number x. Combining (16) and the definition of "Im" yields (8). □

Remark 3.2. We chose the contours (14) and (15) so that the absolute value of the expression under the square root in (8) is greater than $\sqrt{n/(n+1)}$. Therefore,

$$\left| \prod_{k=1}^{n} \sqrt{1 - 2(t-1)\sigma_k^2/x + 2it\sigma_k^2\sqrt{n}/x} \right| > \left(\frac{n}{n+1} \right)^{n/4} > \frac{1}{e^{1/4}} \tag{17}$$

for any $t \in (0, \infty)$ and any $x \in (0, \infty)$. Thus, the integrand in (8) is never large for $t \in (0, \infty)$.

Remark 3.3. The integrand in (8) decays exponentially fast, at a rate independent of the values of σ_1, σ_2, ..., σ_{n-1}, σ_n, and x (see the preceding remark).

Remark 3.4. An efficient means of evaluating (8) numerically is to employ adaptive Gaussian quadratures; see, for example, Section 4.7 of [9]. To attain double-precision accuracy (roughly 15-digit precision), the domain of integration for t in (8) need be only $(0, 40)$ rather than the whole $(0, \infty)$. Good choices for the lowest orders of the quadratures used in the adaptive Gaussian quadratures are 10 and 21, for double-precision accuracy.

Remark 3.5. For a similar, more general approach, see [10]. For alternative approaches, see [11]. Unlike these alternatives, the approach of the present section has an upper bound on its required number of floating-point operations that depends only on the number n of bins and on the precision ε of computations, not on the values of σ_1, σ_2, ..., σ_{n-1}, σ_n, or x. Indeed, it is easy to see that the numerical evaluation of (8) theoretically requires $\mathcal{O}(n \ln^2(\sqrt{n}/\varepsilon))$ quadrature nodes: the denominator of the integrand in (8) cannot oscillate more than $n+1$ times (once for each "pole") as t ranges from 0 to ∞, while the numerator of the integrand cannot oscillate more than $\sqrt{n} \ln(2\sqrt{n}/\varepsilon)$ times as t ranges from 0 to $\ln(2\sqrt{n}/\varepsilon)$; furthermore, the domain of integration for t in (8) need be only $(0, \ln(2\sqrt{n}/\varepsilon))$ rather than the whole $(0, \infty)$. In practice, using several hundred quadrature nodes produces double-precision accuracy (roughly 15-digit precision); see, for example, Section 5 below. Also, the observed performance is similar when subtracting the imaginary unit i from the contours (14) and (15).

4. A procedure for computing the confidence levels

An efficient method for calculating the confidence levels in the limit of large numbers of draws proceeds as follows. Given i.i.d. draws from any distribution — not necessarily from the specified model — we can form the associated statistic X defined in (2) and (1); in the limit of large numbers of draws, the confidence level that the draws do not arise from the model is then just the cumulative distribution function P in (8) evaluated at $x = X$, with σ_k^2 in (8) being the inverses of the positive eigenvalues of the matrix B defined in (6) — after all, $P(x)$ is then the probability that x is greater than the sum of the squares of independent centered Gaussian random variables whose variances are the multiplicative inverses of the positive eigenvalues of B. Remark 3.4 above describes an efficient means of evaluating $P(x)$ numerically.

5. Numerical examples

This section illustrates the performance of the algorithm of Section 4, via several numerical examples.

Below, we plot the complementary cumulative distribution function of the square of the root-mean-square statistic whose probability distribution is determined in Section 2, in the limit of large numbers of draws. This is the distribution of the statistic X defined in (2) when the i.i.d. draws used to form X come from the same model distribution $p_1, p_2, \ldots, p_{n-1}, p_n$ used in (1) for defining X. In order to evaluate the cumulative distribution function (cdf) P, we apply adaptive Gaussian quadratures to the integral in (8) as described in Section 3, obtaining σ_k in (8) via the algorithm described in Section 2.

In applications to goodness-of-fit testing, if the statistic X from (2) takes on a value x, then the confidence level that the draws do not arise from the model distribution is the cdf P in (8) evaluated at x; the significance level that the draws do not arise from the model distribution is therefore $1 - P(x)$. Figures 1 and 2 plot the significance level $(1 - P(x))$ versus x for six example model distributions (examples a, b, c, d, e, f). Table 3 provides formulae for the model distributions used in the six examples. Tables 1 and 2 summarize the computational costs required to attain at least 9-digit absolute accuracy for the plots in Figures 1 and 2, respectively. Each plot displays $1 - P(x)$ at 100 values for x. Figure 2 focuses on the tails of the distributions, corresponding to suitably high confidence levels.

The following list describes the headings of the tables:

- n is the number of bins/categories/cells/classes in Section 2 ($p_1, p_2, \ldots, p_{n-1}, p_n$ are the probabilities of drawing the corresponding bins under the specified model distribution).
- l is the maximum number of quadrature nodes required in any of the 100 evaluations of $1 - P(x)$ displayed in each plot of Figures 1 and 2.
- t is the total number of seconds required to perform the quadratures for all 100 evaluations of $1 - P(x)$ displayed in each plot of Figures 1 and 2.
- p_k is the probability associated with bin k ($k = 1, 2, \ldots, n-1, n$) in Section 2. The constants $C_{(\mathrm{a})}, C_{(\mathrm{b})}, C_{(\mathrm{c})}, C_{(\mathrm{d})}, C_{(\mathrm{e})}, C_{(\mathrm{f})}$ in Table 3 are the positive real numbers chosen such that $\sum_{k=1}^{n} p_k = 1$. For any real number x, the floor $\lfloor x \rfloor$ is the greatest integer less than or equal to x; the probability distributions for examples (c) and (d) involve the floor.

We used Fortran 77 and ran all examples on one core of a 2.2 GHz Intel Core 2 Duo microprocessor with 2 MB of L2 cache. Our code is compliant with the IEEE double-precision standard (so that the mantissas of variables have approximately one bit of precision less than 16 digits, yielding a relative precision of about 2E–16). We diagonalized the matrix B defined in (6) using the Jacobi algorithm (see, for example, Chapter 8 of [12]), not taking advantage of Remark 2.2; explicitly forming the entries of the matrix B defined in (6) can incur a numerical error of at most the machine precision (about 2E–16) times $\max_{1 \le k \le n} p_k / \min_{1 \le k \le n} p_k$, yielding 9-digit accuracy or better for all our examples. A future article will exploit the interlacing properties of eigenvalues, as in [6], to obtain higher precision. Of course, even 5-digit precision would suffice for most statistical applications; however, modern computers can produce high accuracy very fast, as the examples in this section illustrate.

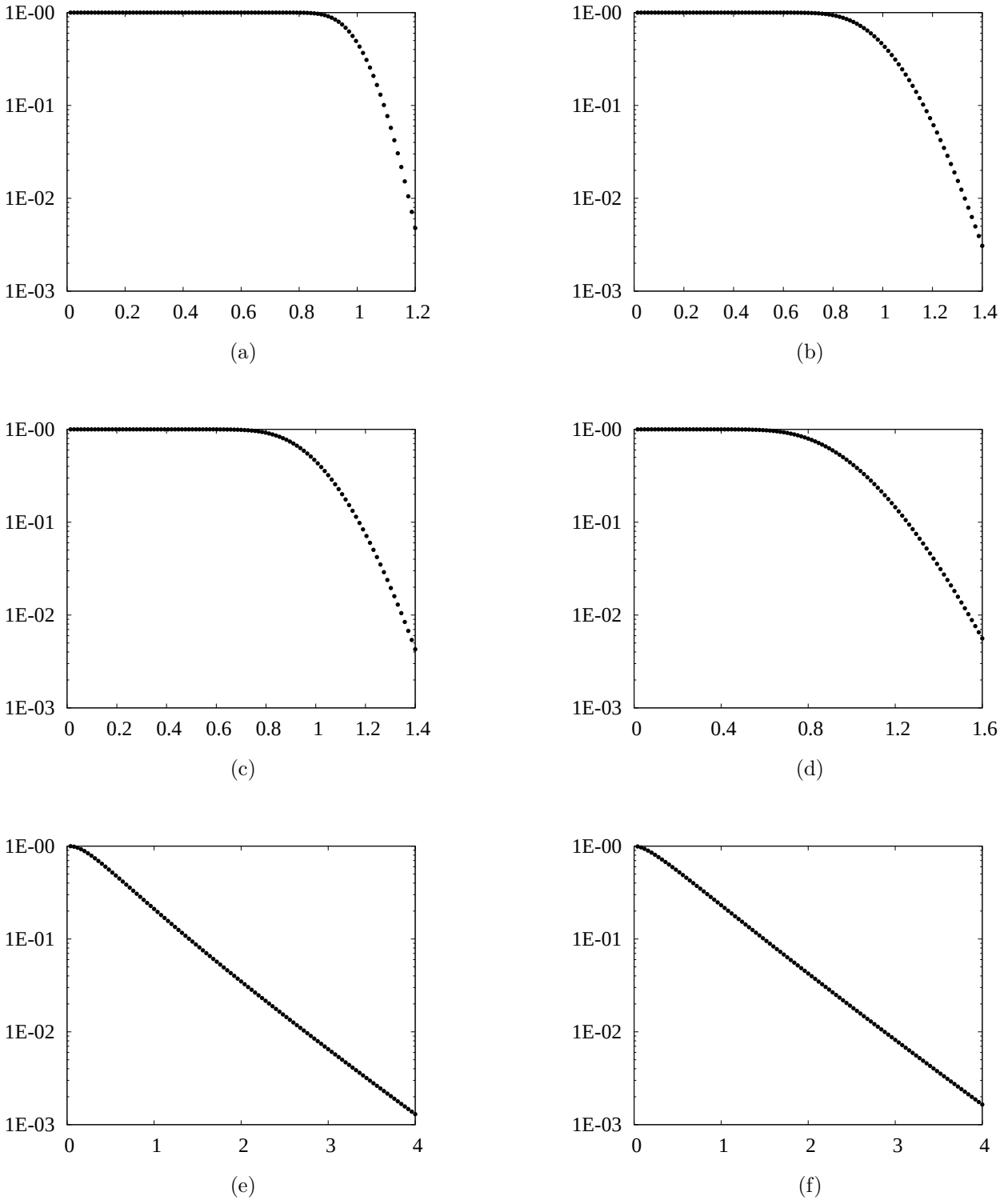
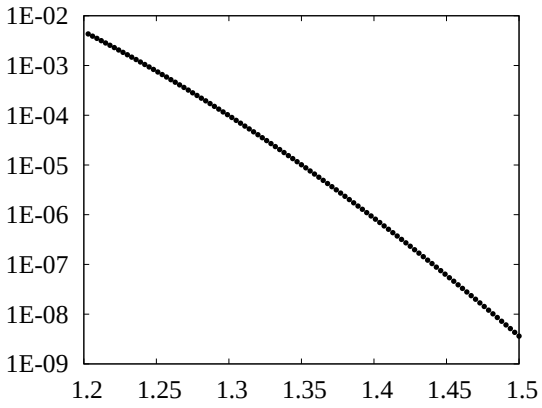
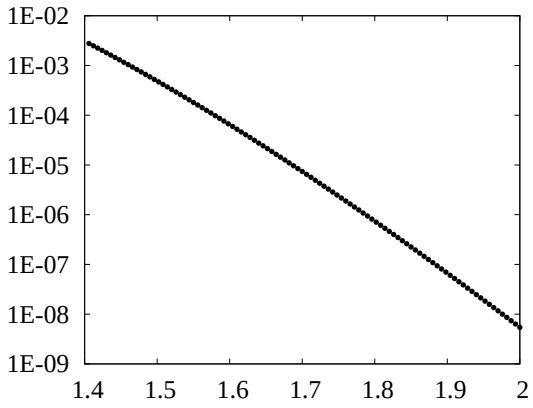


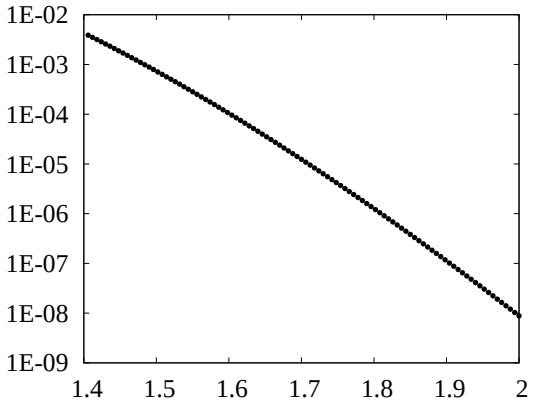
Fig. 1: The vertical axis is $1 - P(x)$ from (8); the horizontal axis is x.



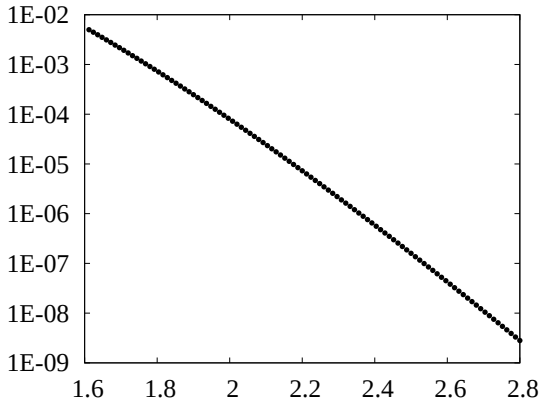
(a)



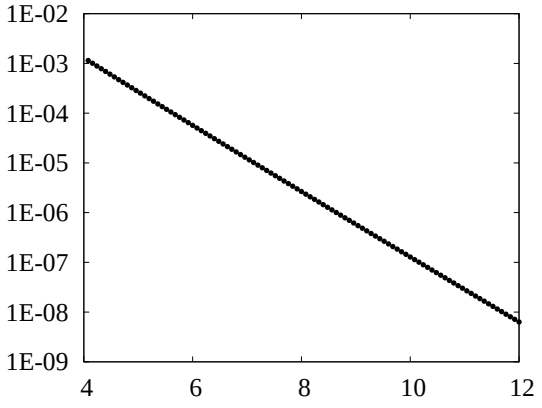
(b)



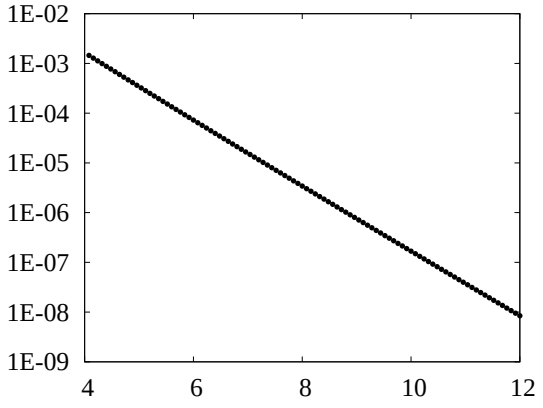
(c)



(d)



(e)



(f)

Fig. 2: The vertical axis is $1 - P(x)$ from (8); the horizontal axis is x.

Table 1: Values for Figure 1

	n	l	t
(a)	500	310	5.0
(b)	250	270	2.4
(c)	100	250	0.9
(d)	50	250	0.5
(e)	25	330	0.3
(f)	10	270	0.1

Table 2: Values for Figure 2

	n	l	t
(a)	500	310	5.7
(b)	250	330	3.0
(c)	100	270	1.0
(d)	50	290	0.6
(e)	25	350	0.4
(f)	10	270	0.2

Table 3: Values for both Figure 1 and Figure 2

	n	p_k
(a)	500	$C_{(\mathrm{a})} \cdot (300 + k)^{-2}$
(b)	250	$C_{(\mathrm{b})} \cdot (260 - k)^{3}$
(c)	100	$C_{(\mathrm{c})} \cdot \lfloor (40 + k)/40 \rfloor^{-1/6}$
(d)	50	$C_{(\mathrm{d})} \cdot (1/2 + \ln \lfloor (61 - k)/10 \rfloor)$
(e)	25	$C_{(\mathrm{e})} \cdot \exp(-5k/8)$
(f)	10	$C_{(\mathrm{f})} \cdot \exp(-(k-1)^2/6)$

6. The power of the root-mean-square

This section very briefly compares the statistic defined in (2) and the classic χ^2 statistic defined in (3). This abbreviated comparison is in no way complete; a much more comprehensive treatment constitutes a forthcoming article.

We will discuss four statistics in all — the root-mean-square, χ^2, the (log)likelihood-ratio, and the Freeman-Tukey or Hellinger distance. We use p_1, p_2, ..., p_{n-1}, p_n to denote the expected fractions of the m i.i.d. draws falling in each of the n bins, and Y_1, Y_2, ..., Y_{n-1}, Y_n to denote the observed fractions of the m draws falling in the n bins. That is, p_1, p_2, ..., p_{n-1}, p_n are the probabilities associated with the n bins in the model distribution, whereas Y_1, Y_2, ..., Y_{n-1}, Y_n are the fractions of the m draws falling in the n bins when we take the draws from a certain "actual" distribution that may differ from the model.

With this notation, the square of the root-mean-square statistic is

$$X = m \sum_{k=1}^{n} (Y_k - p_k)^2. \tag{18}$$

We use the designation "root-mean-square" to label the lines associated with X in the plots below.

The classic Pearson χ^2 statistic is

$$\chi^2 = m \sum_{k=1}^{n} \frac{(Y_k - p_k)^2}{p_k}. \tag{19}$$

We use the standard designation "χ^2" to label the lines associated with χ^2 in the plots below.

The (log)likelihood-ratio or "G^2" statistic is

$$G^2 = 2m \sum_{k=1}^{n} Y_k \ln\left(\frac{Y_k}{p_k}\right), \tag{20}$$

under the convention that $Y_k \ln(Y_k/p_k) = 0$ if $Y_k = 0$. We use the common designation "G^2" to label the lines associated with G^2 in the plots below.

The Freeman-Tukey or Hellinger-distance statistic is

$$H^2 = 4m \sum_{k=1}^{n} (\sqrt{Y_k} - \sqrt{p_k})^2. \tag{21}$$

We use the well-known designation "Freeman-Tukey" to label the lines associated with H^2 in the plots below.

In the limit that the number m of draws is large, the distributions of χ^2 defined in (19), G^2 defined in (20), and H^2 defined (21) are all the same when the actual distribution of the draws is identical to the model (see, for example, [1]). However, when the number m of draws is not large, then their distributions can differ substantially. In this section, we compute confidence levels via Monte Carlo simulations, without relying on the number m of draws to be large.

Remark 6.1. Below, we say that a statistic based on given i.i.d. draws "distinguishes" the actual underlying distribution of the draws from the model distribution to mean that the computed confidence level is at least 99% for 99% of 40,000 simulations, with each simulation generating m i.i.d. draws according to the actual distribution. We computed the confidence levels by conducting 40,000 simulations, each generating m i.i.d. draws according to the model distribution.

6.1. First example

Let us first specify the model distribution to be

$$p_1 = \frac{1}{4}, \tag{22}$$

$$p_2 = \frac{1}{4}, \tag{23}$$

$$p_k = \frac{1}{2n-4} \tag{24}$$

for $k = 3, 4, \ldots, n-1, n$. We consider m i.i.d. draws from the distribution

$$\tilde{p}_1 = \frac{3}{8}, \tag{25}$$

$$\tilde{p}_2 = \frac{1}{8}, \tag{26}$$

$$\tilde{p}_k = p_k \tag{27}$$

for $k = 3, 4, \ldots, n-1, n$, where $p_3, p_4, \ldots, p_{n-1}, p_n$ are the same as in (24).

Figure 3 plots the percentage of 40,000 simulations, each generating 200 i.i.d. draws according to the actual distribution defined in (25)–(27), that are successfully detected as not arising from the model distribution at the 1% significance level (meaning that the associated statistic for the simulation yields a confidence level of 99% or greater). We computed the significance levels by conducting 40,000 simulations, each generating 200 i.i.d. draws according to the model distribution defined in (22)–(24). Figure 3 shows that the root-mean-square is successful in at least 99% of the simulations, while the classic χ^2 statistic fails often, succeeding in only 81% of the simulations for $n = 16$, and less than 5% for $n \geq 256$.

Figure 4 plots the number m of draws required to distinguish the actual distribution defined in (25)–(27) from the model distribution defined in (22)–(24). Remark 6.1 above specifies what we mean by "distinguish." Figure 4 shows that the root-mean-square requires only about $m = 185$ draws for any number n of bins, while the classic χ^2 statistic requires 90% more draws for $n = 16$, and greater than 300% more for $n \geq 128$. Furthermore, the classic χ^2 statistic requires increasingly many draws as the number n of bins increases, unlike the root-mean-square.

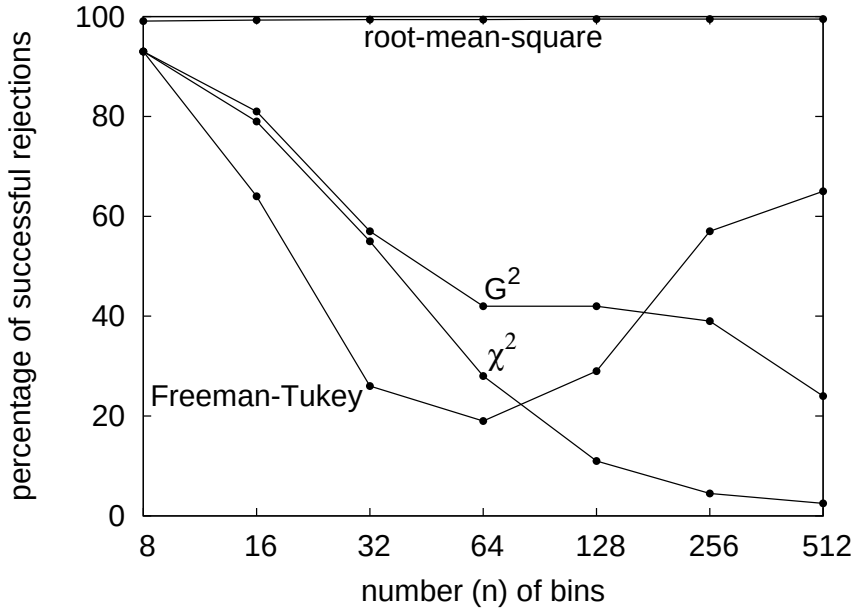


Fig. 3: First example (rate of success); see Subsection 6.1.

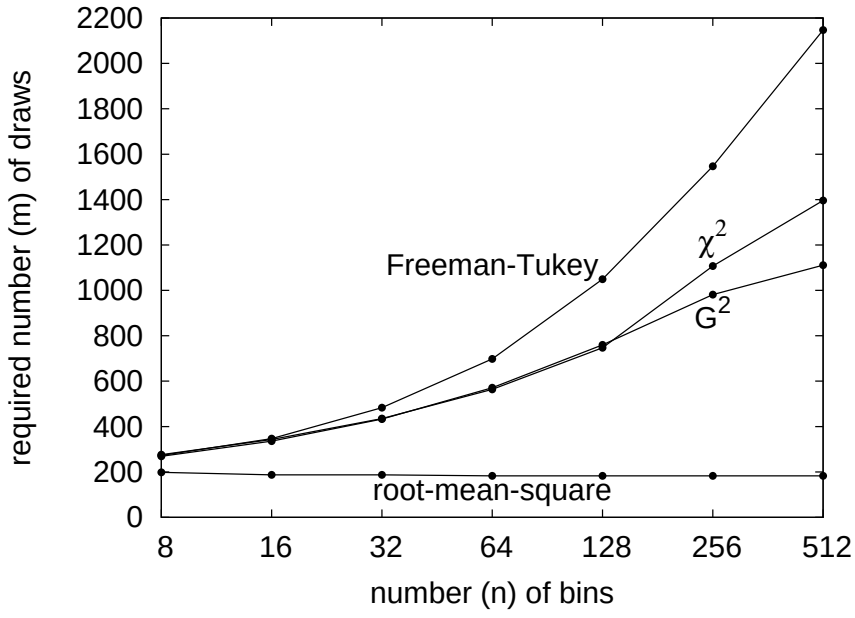


Fig. 4: First example (statistical "efficiency"); see Subsection 6.1.

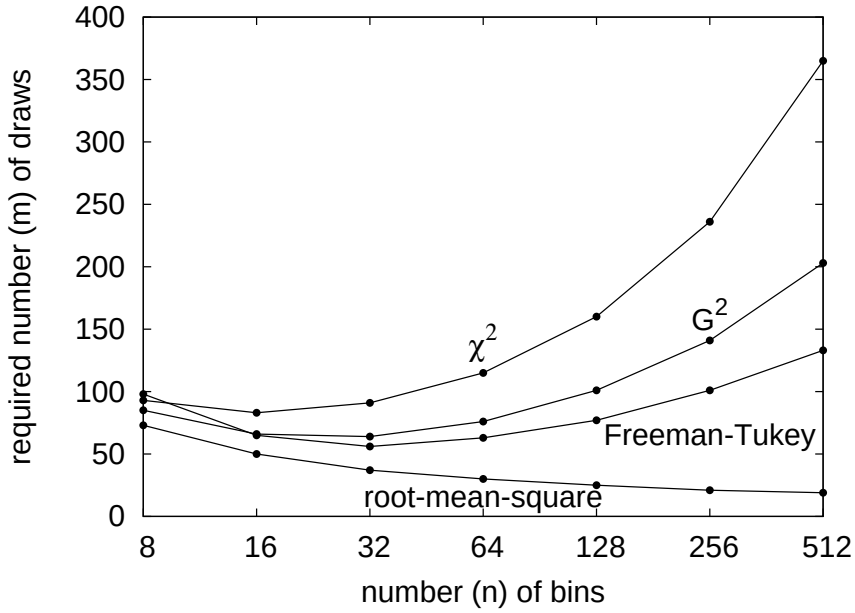


Fig. 5: Second example; see Subsection 6.2.

6.2. Second example

Next, let us specify the model distribution to be

$$p_k = \frac{C_1}{k} \tag{28}$$

for $k = 1, 2, \ldots, n-1, n$, where

$$C_1 = \frac{1}{\sum_{k=1}^{n} 1/k}. \tag{29}$$

We consider m i.i.d. draws from the distribution

$$\tilde{p}_k = \frac{C_2}{k^2} \tag{30}$$

for $k = 1, 2, \ldots, n-1, n$, where

$$C_2 = \frac{1}{\sum_{k=1}^{n} 1/k^2}. \tag{31}$$

Figure 5 plots the number m of draws required to distinguish the actual distribution defined in (30) and (31) from the model distribution defined in (28) and (29). Remark 6.1 above specifies what we mean by "distinguish." Figure 5 shows that the classic χ^2 statistic requires increasingly many draws as the number n of bins increases, while the root-mean-square exhibits the opposite behavior.

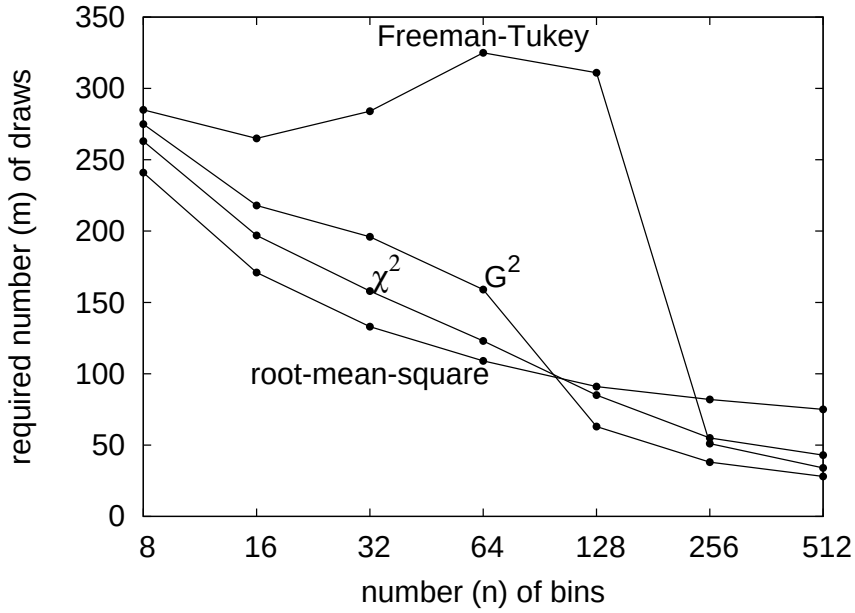


Fig. 6: Third example; see Subsection 6.3.

6.3. Third example

Let us again specify the model distribution to be

$$p_k = \frac{C_1}{k} \tag{32}$$

for $k = 1, 2, \ldots, n-1, n$, where

$$C_1 = \frac{1}{\sum_{k=1}^{n} 1/k}. \tag{33}$$

We now consider m i.i.d. draws from the distribution

$$\tilde{p}_k = \frac{C_{1/2}}{\sqrt{k}} \tag{34}$$

for $k = 1, 2, \ldots, n-1, n$, where

$$C_{1/2} = \frac{1}{\sum_{k=1}^{n} 1/\sqrt{k}}. \tag{35}$$

Figure 6 plots the number m of draws required to distinguish the actual distribution defined in (34) and (35) from the model distribution defined in (32) and (33). Remark 6.1 above specifies what we mean by "distinguish." The root-mean-square is not uniformly more powerful than the other statistics in this example; see Remark 6.2 below.

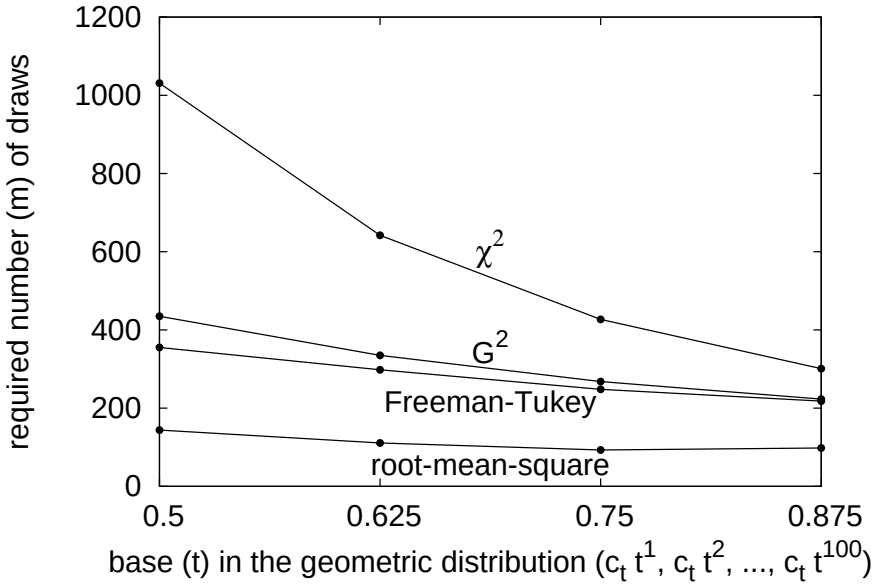


Fig. 7: Fourth example; see Subsection 6.4.

6.4. Fourth example

We turn now to models involving parameter estimation (for details, see [8]). Let us specify the model distribution to be the Zipf distribution

$$p_k(\theta) = \frac{C_\theta}{k^\theta} \tag{36}$$

for $k = 1, 2, \ldots, 99, 100$, where

$$C_\theta = \frac{1}{\sum_{k=1}^{100} 1/k^\theta}; \tag{37}$$

we estimate the parameter θ via maximum-likelihood methods (see [8]). We consider m i.i.d. draws from the (truncated) geometric distribution

$$\tilde{p}_k = c_t\, t^k \tag{38}$$

for $k = 1, 2, \ldots, 99, 100$, where

$$c_t = \frac{1}{\sum_{k=1}^{100} t^k}; \tag{39}$$

Figure 7 considers several values for t.

Figure 7 plots the number m of draws required to distinguish the actual distribution defined in (38) and (39) from the model distribution defined in (36) and (37), estimating the parameter θ in (36) and (37) via maximum-likelihood methods. Remark 6.1 above specifies what we mean by "distinguish."

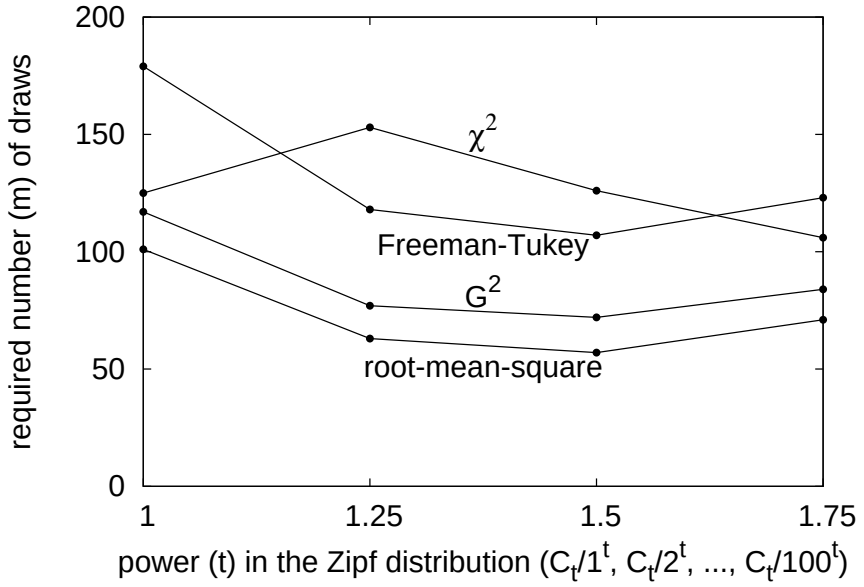


Fig. 8: Fifth example; see Subsection 6.5.

6.5. Fifth example

The model for our final example involves parameter estimation, too (for details, see [8]). Let us specify the model distribution to be

$$p_k(\theta) = \theta^{k-1}(1-\theta) \tag{40}$$

for $k = 1, 2, \ldots, 98, 99$, and

$$p_{100}(\theta) = \theta^{99}; \tag{41}$$

we estimate the parameter θ via maximum-likelihood methods (see [8]). We consider m i.i.d. draws from the Zipf distribution

$$\tilde{p}_k = \frac{C_t}{k^t} \tag{42}$$

for $k = 1, 2, \ldots, 99, 100$, where

$$C_t = \frac{1}{\sum_{k=1}^{100} 1/k^t}; \tag{43}$$

Figure 8 considers several values for t.

Figure 8 plots the number m of draws required to distinguish the actual distribution defined in (42) and (43) from the model distribution defined in (40) and (41), estimating the parameter θ in (40) and (41) via maximum-likelihood methods. Remark 6.1 above specifies what we mean by "distinguish."

Remark 6.2. The root-mean-square statistic is not very sensitive to relative discrepancies between the model and actual distributions in bins whose associated model probabilities are small. When sensitivity in these bins is desirable, we recommend using both the root-mean-square statistic defined in (2) and an asymptotically equivalent variation of χ^2 defined in (3), such as the (log)likelihood-ratio or "G^2" test; see, for example, [1].

7. Conclusions and generalizations

This paper provides efficient black-box algorithms for computing the confidence levels for one of the most natural goodness-of-fit statistics, in the limit of large numbers of draws. As mentioned briefly above (in Remark 2.3), our methods can handle model distributions specified via the multinomial maximum-likelihood estimation of parameters from the data; for details, see [8]. Moreover, we can handle model distributions with infinitely many bins; for details, see Observation 1 in Section 4 of [8]. Furthermore, we can handle arbitrarily weighted means in the root-mean-square, in addition to the usual, uniformly weighted average considered above. Finally, combining our methods and the statistical bootstrap should produce a test for whether two separate sets of draws arise from the same or from different distributions, when each set is taken i.i.d. from some (unspecified) distribution associated with the set (see, for example, [13]).

The natural statistic has many advantages over more standard χ^2 tests, as forthcoming papers will demonstrate. The classic χ^2 statistic for goodness-of-fit, and especially variations such as the (log)likelihood-ratio, "G^2," and power-divergence statistics (see [1]), can be sensible supplements, but are not good alternatives when used alone. With the now widespread availability of computers, calculating significance levels via Monte Carlo simulations for the more natural statistic of the present article can be feasible; the algorithms of the present paper can also be suitable, and are efficient and easy-to-use.

Acknowledgements

We would like to thank Tony Cai, Jianqing Fan, Peter W. Jones, Ron Peled, and Vladimir Rokhlin for many helpful discussions. We would like to thank the anonymous referees for their helpful suggestions. William Perkins was supported in part by NSF Grant OISE-0730136. Mark Tygert was supported in part by an Alfred P. Sloan Research Fellowship. Rachel Ward was supported in part by an NSF Postdoctoral Research Fellowship.

References

[1] C. R. Rao, Karl Pearson chi-square test: The dawn of statistical inference, in: C. Huber-Carol, N. Balakrishnan, M. S. Nikulin, M. Mesbah (Eds.), Goodness-of-Fit Tests and Model Validity, Birkhäuser, Boston, 2002, pp. 9–24.

[2] S. R. S. Varadhan, M. Levandowsky, N. Rubin, Mathematical Statistics, Lecture Notes Series, Courant Institute of Mathematical Sciences, NYU, New York, 1974.

[3] 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, Philosophical Magazine, Series 5, 50 (1900) 157–175.

[4] D. S. Moore, M. C. Spruill, Unified large-sample theory of general chi-squared statistics for tests of fit, Ann. Statist. 3 (1975) 599–616.

[5] M. G. Kendall, A. Stuart, K. Ord, S. Arnold, Kendall's Advanced Theory of Statistics, volume 1 and 2A, Wiley, 6th edition, 2009.

[6] M. Gu, S. C. Eisenstat, A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem, SIAM J. Matrix Anal. Appl. 15 (1994) 1266–1276.

[7] M. Gu, S. C. Eisenstat, A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem, SIAM J. Matrix Anal. Appl. 16 (1995) 172–191.

[8] W. Perkins, M. Tygert, R. Ward, Computing the confidence levels for a root-mean-square test of goodness-of-fit, II, Technical Report 1009.2260, arXiv, 2010.

[9] W. Press, S. Teukolsky, W. Vetterling, B. Flannery, Numerical Recipes, Cambridge University Press, Cambridge, UK, 3rd edition, 2007.

[10] S. O. Rice, Distribution of quadratic forms in normal random variables — Evaluation by numerical integration, SIAM J. Sci. Stat. Comput. 1 (1980) 438–448.

[11] P. Duchesne, P. L. de Micheaux, Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods, Comput. Statist. Data Anal. 54 (2010) 858–862.

[12] G. H. Golub, C. F. Van Loan, Matrix Computations, Johns Hopkins University Press, Baltimore, Maryland, 3rd edition, 1996.

[13] B. Efron, R. Tibshirani, An introduction to the bootstrap, Chapman & Hall/CRC Press, Boca Raton, Florida, 1993.