Appl. Comput. Harmon. Anal. 36 (2014) 361–386

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Some deficiencies of χ^2 and classical exact tests of significance ☆

Will Perkins [a], Mark Tygert [b,*], Rachel Ward [c]

[a] Georgia Institute of Technology, 686 Cherry St., Atlanta, GA 30332-0160, United States
[b] Courant Institute, NYU, 251 Mercer St., New York, NY 10012, United States
[c] University of Texas, 1 University Station, C1200, Austin, TX 78712, United States

ARTICLE INFO

Article history:
Received 6 February 2012
Received in revised form 22 May 2013
Accepted 10 June 2013
Available online 14 June 2013
Communicated by Vladimir Rokhlin

Keywords:
Chi-square
Fisher's exact
Freeman–Tukey
Likelihood ratio
Power divergence
Root-mean-square

ABSTRACT

Goodness-of-fit tests based on the Euclidean distance often outperform χ^2 and other classical tests (including the standard exact tests) by at least an order of magnitude when the model being tested for goodness-of-fit is a discrete probability distribution that is not close to uniform. The present article discusses numerous examples of this. Goodness-of-fit tests based on the Euclidean metric are now practical and convenient: although the actual values taken by the Euclidean distance and similar goodness-of-fit statistics are seldom interpretable without the aid of a computer, black-box software can rapidly calculate their precise significance.

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 given "model", where the model may consist of either a single fully specified probability distribution or a parameterized family of probability distributions. The present paper concerns the case in which the draws are discrete random variables, taking values in a finite or countable 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 model 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 model distribution (see, for example, [23]; [28, p. 123] or Section 2 below). If the draws do in fact arise from the 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 were not taken i.i.d. from the model.

To quantify "large" and "confident", let us denote by x the value of the root-mean-square for the given i.i.d. draws; let us denote by X the root-mean-square statistic constructed for different i.i.d. draws that definitely do in fact come from the model (if the model is parameterized, then we draw from the distribution corresponding to the parameter given by a

☆ An extended version of the present article is available at http://arxiv.org/abs/1108.4126.

* Corresponding author.
E-mail addresses: perkins@math.gatech.edu (W. Perkins), tygert@courant.nyu.edu (M. Tygert), rward@math.utexas.edu (R. Ward).

1063-5203/$ – see front matter
http://dx.doi.org/10.1016/j.acha.2013.06.002

maximum-likelihood estimate for the experimental data). The "P-value" P is then defined to be the probability that $X \geqslant x$ (viewing X — but not x — as a random variable). Given the P-value P, we can have $100(1-P)\%$ confidence that the draws were not taken i.i.d. from the model.

Now, the P-values for the simple root-mean-square statistic can be different functions of x for different model probability distributions. To avoid this seeming inconvenience asymptotically (in the limit of large numbers of draws), K. Pearson replaced the uniformly weighted mean in the root-mean-square with a weighted average; the weights are the reciprocals of the model probabilities associated with the various bins. This produces the classic χ^2 statistic introduced by Pearson [17] — see, for example, formula (2) below. However, when model probabilities can be small (relative to others in the same model distribution), this weighted average can involve division by nearly zero. As demonstrated below, dividing by nearly zero severely restricts the statistical power of χ^2 — even in the absence of round-off errors — especially when dividing by nearly zero for each of many bins. The problem arises whether or not every bin contains several draws (see Remark 1.1). Press [22] tackled similar issues.

The main thesis of the present article is that using only the classic χ^2 statistic is no longer appropriate, that certain alternatives are far superior now that computers are widely available. As illustrated below, the simple root-mean-square, used in conjunction with the log-likelihood-ratio "G^2" goodness-of-fit statistic, is generally preferable to the classic χ^2 statistic. (The log-likelihood-ratio also involves division by nearly zero, but tempers this somewhat via the logarithm.) We do not claim that this is always the best alternative. In fact, the discrete Kolmogorov–Smirnov and related statistics used by Clauset et al. [3] and D'Agostino and Stephens [6] can be more powerful than the root-mean-square when there is a natural data-independent ordering (or partial order) for the bins; in any case, the discrete Kolmogorov–Smirnov statistic and the root-mean-square are similar in many ways, and complementary in others. We focus on the root-mean-square because it is simple and easy to understand; for example, computing the P-values of the root-mean-square in the limit of large numbers of draws is trivial, even when estimating continuous parameters via maximum-likelihood methods, as discussed by Perkins et al. [19,20]. Moreover, the classic χ^2 statistic is just a weighted version of the root-mean-square, facilitating their comparison. Finally, χ^2 and the root-mean-square coincide when the model distribution is uniform.

Please note that all statistical tests reported in the present paper (including those involving the χ^2 statistic) are exact; we compute P-values via Monte Carlo simulations providing guaranteed error bounds (see Section 3 below). In all numerical results reported below, we generated random numbers via the C programming language procedure given on page 9 of Marsaglia [15], implementing the recommended complementary multiply with carry.

To be sure, the problem with χ^2 is neither subtle nor esoteric. For a particularly revealing example, see Section 4.5 below.

Appropriate rebinning to uniformize the probabilities associated with the bins can mitigate much of the problem with χ^2. Yet rebinning is a black art that is liable to improperly influence the result of a goodness-of-fit test. Moreover, rebinning requires careful extra work, making χ^2 less easy-to-use. A principal advantage of the root-mean-square is that it does not require any rebinning; indeed, the root-mean-square is most powerful without any rebinning.

Remark 1.1. In many of our examples, there is a bin for which the expected number of draws is very small under the model. Please note that, although it is natural for the expected numbers of draws for some bins to be very small, especially when the model has many bins, the advantage of the root-mean-square over χ^2 is substantial even when the expected number of draws is at least five for every bin; see, for example, Section 5.1.1 or Section 5.2.4.

Remark 1.2. Goodness-of-fit tests are probably most useful in practice not for ascertaining whether a model is correct or not, but for determining whether the discrepancy between the model and experiment is larger than expected random fluctuations. While models outside the physical sciences typically are not exactly correct, testing the validity of using a model for virtually any purpose requires knowing whether observed discrepancies are due to inaccuracies or inadequacies in the models or (on the contrary) could be due to chance arising from necessarily finite sample sizes. Thus, goodness-of-fit tests are critical even when the models are not supposed to be exactly correct, in order to gauge the size of the unavoidable random fluctuations. For further clarification, see the remarkably extensive title used by Pearson [17] introducing χ^2; see also the modern treatments by Gelman [11] and Cox [5].

Remark 1.3. The assumption that the given observations are i.i.d. draws is not necessary. This paper focuses on the i.i.d. case for simplicity. A more general treatment has been given by Perkins et al. [21].

2. Definitions of the test statistics

In this section, we review the definitions of four goodness-of-fit statistics — the root-mean-square, χ^2, the log-likelihood-ratio or G^2, and the Freeman–Tukey or Hellinger distance. The latter three statistics are the best-known members of the standard Cressie–Read power-divergence family, as discussed by Rao [23]. We use $p_0^{(1)}, p_0^{(2)}, \ldots, p_0^{(m)}$ to denote the modeled fractions of n i.i.d. draws falling in m bins, numbered $1, 2, \ldots, m$, respectively, and we use $\hat{p}^{(1)}, \hat{p}^{(2)}, \ldots, \hat{p}^{(m)}$ to denote the observed fractions of the n draws falling in the respective bins. That is, $p_0^{(1)}, p_0^{(2)}, \ldots, p_0^{(m)}$ are the probabilities associated with the respective bins in the *model* distribution, whereas $\hat{p}^{(1)}, \hat{p}^{(2)}, \ldots, \hat{p}^{(m)}$ are the fractions of the n draws falling in

the respective bins when we take the draws from a distribution that may differ from the model — their *actual* distribution. Specifically, if $i_1, i_2, \ldots, i_n$ are the observed i.i.d. draws, then $\hat{p}^{(j)}$ is $\frac{1}{n}$ times the number of $i_1, i_2, \ldots, i_n$ falling in bin j, for $j = 1, 2, \ldots, m$. If the model is parameterized by a parameter θ, then the probabilities $p_0^{(1)}, p_0^{(2)}, \ldots, p_0^{(m)}$ are functions of θ; if the model is fully specified, then we can view the probabilities $p_0^{(1)}, p_0^{(2)}, \ldots, p_0^{(m)}$ as constant as functions of θ. We use $\hat{\theta}$ to denote a maximum-likelihood estimate of θ obtained from $\hat{p}^{(1)}, \hat{p}^{(2)}, \ldots, \hat{p}^{(m)}$.

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

$$x = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\bigl(\hat{p}^{(j)} - p_0^{(j)}(\hat{\theta})\bigr)^2}. \tag{1}$$

We use the designation "root-mean-square" to refer to x.

The classical Pearson χ^2 statistic is

$$\chi^2 = n\sum_{j=1}^{m}\frac{(\hat{p}^{(j)} - p_0^{(j)}(\hat{\theta}))^2}{p_0^{(j)}(\hat{\theta})}, \tag{2}$$

under the convention that $(\hat{p}^{(j)} - p_0^{(j)}(\hat{\theta}))^2/p_0^{(j)}(\hat{\theta}) = 0$ if $p_0^{(j)}(\hat{\theta}) = 0 = \hat{p}^{(j)}$. We use the standard designation "χ^2" to refer to χ^2.

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

$$g^2 = 2n\sum_{j=1}^{m}\hat{p}^{(j)}\ln\biggl(\frac{\hat{p}^{(j)}}{p_0^{(j)}(\hat{\theta})}\biggr), \tag{3}$$

under the convention that $\hat{p}^{(j)}\ln(\hat{p}^{(j)}/p_0^{(j)}(\hat{\theta})) = 0$ if $\hat{p}^{(j)} = 0$. We use the common designation "G^2" to refer to g^2.

The Freeman–Tukey or Hellinger-distance statistic is

$$h^2 = 4n\sum_{j=1}^{m}\Bigl(\sqrt{\hat{p}^{(j)}} - \sqrt{p_0^{(j)}(\hat{\theta})}\Bigr)^2 = 4n\sum_{j=1}^{m}\Bigl[\bigl(\hat{p}^{(j)} - p_0^{(j)}(\hat{\theta})\bigr)^2\Big/\Bigl(\sqrt{\hat{p}^{(j)}} + \sqrt{p_0^{(j)}(\hat{\theta})}\Bigr)^2\Bigr]. \tag{4}$$

We use the well-known designation "Freeman–Tukey" to refer to h^2.

In the limit that the number n of draws is large, the distributions of χ^2 defined in (2), g^2 defined in (3), and h^2 defined in (4) are all the same when the actual underlying distribution of the draws comes from the model, as discussed, for example, by Rao [23]. However, when the number n of draws is not large, then their distributions can differ substantially. In all our data and power analyses, we compute P-values via Monte Carlo simulations, without relying on the number n of draws to be large.

3. Hypothesis tests with parameter estimation

In this section, we discuss the testing of hypotheses involving parameterized models: Given a family $p_0(\theta)$ of probability distributions parameterized by θ, and given observed i.i.d. draws from some actual underlying (unknown) distribution p, we would like to test the hypothesis

$$H_0'\text{: for some } \theta, \quad p = p_0(\theta), \tag{5}$$

against the alternative

$$H_1'\text{: for all } \theta, \quad p \neq p_0(\theta). \tag{6}$$

Given only finitely many draws, the P-value for such a test would have to be independent of the parameter θ, since the proper value for θ is unknown (θ is known as a "nuisance" parameter). Unfortunately, it is not clear how to devise such a test when the probability distributions are discrete. None of the standard methods (including χ^2, the log-likelihood-ratio, the Freeman–Tukey/Hellinger distance, and other Cressie–Read power-divergence statistics) produce P-values that are independent of the parameter θ. Some methods do produce P-values that are independent of θ in the limit of large numbers of draws, but this is not especially useful, since in the limit of large numbers of draws any actual parameter θ would be almost surely known exactly anyway; further elaboration is available in Appendix B of Perkins et al. [18].

In the present paper, we test the significance of assuming

$$H_0 : p = p_0(\hat{\theta}) \quad \text{for the particular observed value of } \hat{\theta}, \tag{7}$$

where $\hat{\theta}$ is a maximum-likelihood estimate of θ; i.e., H_0 is the hypothesis that $p = p_0(\hat{\theta})$ for the value of $\hat{\theta}$ associated with the single realization of the experiment that was measured (subsequent repetitions of the experiment, including those

considered when calculating the P-value as in Remark 3.3, can yield different estimates of the parameter, even though the repetitions' actual distribution p is the same). Of course, the accuracy of the estimate $\hat{\theta}$ generally improves as the number of draws increases; (5) and (7) are asymptotically equivalent, in the limit of large numbers of draws, under the conditions of Romano [25] or Henze [13].

As testing exactly H_0' defined in (5) does not seem to be feasible in general when the probability distributions are discrete and there are more than just a few bins, we focus on testing the closely related assumption H_0 defined in (7). The latter is more relevant for many applications, anyways — plots typically display the particular fitted distribution in (7); interpreting such plots naturally involves (7). All tests of the present paper concern the significance of assuming H_0 defined in (7) (if the model is fully specified, then the probability distribution $p_0(\theta)$ is the same for all θ). Please be sure to bear in mind Remark 1.2 of Section 1. A significance test simply gauges the consistency of the observed data with our assumption; we are not trying to decide whether the assumption is likely to be true (or false), nor are we trying to decide whether some alternative assumption is likely to be true (or false).

Remark 3.1. Another means of handling nuisance parameters is to test the hypothesis

$$H_0'' : p = p_0(\hat{\theta}) \quad \text{for all possible realizations of the experiment,} \tag{8}$$

that is, H_0'' is the hypothesis that $p = p_0(\hat{\theta})$ and that $p_0(\hat{\theta})$ always takes exactly the same value during repetitions of the experiment. The assumption that (8) is true seems to be more extreme, a more substantial departure from (5), than (7). Nevertheless, testing (8) is standard; see, for example, Section 6 of Cochran [4]. Assuming (8) amounts to conditioning (5) on a statistic that is minimally sufficient for estimating θ; computing the associated P-values is not always trivial. Testing the significance of assuming (7) would seem to be more apropos in practice for applications in which the experimental design does not enforce that repeated experiments always yield the same value for $p_0(\hat{\theta})$.

Remark 3.2. The parameter θ can be integer-valued, real-valued, complex-valued, vector-valued, matrix-valued, or any combination of the many possibilities. For instance, when we do not know the proper ordering of the bins a priori, we must include a parameter that contains a permutation (or permutation matrix) specifying the order of the bins; maximum-likelihood estimation then entails sorting the model and all empirical frequencies (whether experimental or simulated) — see Section 4.2 for details. With Remark 3.3, we need not contemplate how many degrees of freedom are in a permutation.

Remark 3.3. To compute the P-value assessing the consistency of the experimental data with assuming (7), we can use Monte Carlo simulations (very similar to those used by Clauset et al. [3]). First, we estimate the parameter θ from the n given experimental draws, obtaining $\hat{\theta}$, and calculate the statistic (χ^2, G^2, Freeman–Tukey, or the root-mean-square), using the given data and taking the model distribution to be $p_0(\hat{\theta})$. We then run many simulations. To conduct a single simulation, we perform the following three-step procedure:

1. we generate n i.i.d. draws according to the model distribution $p_0(\hat{\theta})$, where $\hat{\theta}$ is the estimate calculated from the experimental data,
2. we estimate the parameter θ from the data generated in Step 1, obtaining a new estimate $\tilde{\theta}$, and
3. we calculate the statistic under consideration (χ^2, G^2, Freeman–Tukey, or the root-mean-square), using the data generated in Step 1 and taking the model distribution to be $p_0(\tilde{\theta})$, where $\tilde{\theta}$ is the estimate calculated in Step 2 from the data generated in Step 1.

After conducting many such simulations, we may estimate the P-value for assuming (7) as the fraction of the statistics calculated in Step 3 that are greater than or equal to the statistic calculated from the empirical data. The accuracy of the estimated P-value is inversely proportional to the square root of the number of simulations conducted; for details, see Remark 3.4 below. This procedure works since, by definition, the P-value is the probability that

$$d\left[\begin{pmatrix}\hat{p}^{(1)}\\ \hat{p}^{(2)}\\ \vdots\\ \hat{p}^{(m)}\end{pmatrix}, \begin{pmatrix}p_0^{(1)}(\hat{\Theta})\\ p_0^{(2)}(\hat{\Theta})\\ \vdots\\ p_0^{(m)}(\hat{\Theta})\end{pmatrix}\right] \geqslant d\left[\begin{pmatrix}\hat{p}^{(1)}\\ \hat{p}^{(2)}\\ \vdots\\ \hat{p}^{(m)}\end{pmatrix}, \begin{pmatrix}p_0^{(1)}(\hat{\theta})\\ p_0^{(2)}(\hat{\theta})\\ \vdots\\ p_0^{(m)}(\hat{\theta})\end{pmatrix}\right], \tag{9}$$

where

- m is the number of all possible values that the draws can take,
- d is the measure of the discrepancy between two probability distributions over m bins (i.e., between two vectors each with m entries) that is associated with the statistic under consideration (d is the Euclidean distance for the root-mean-square, a weighted Euclidean distance for χ^2, the Hellinger distance for the Freeman–Tukey statistic, and the relative entropy — the Kullback–Leibler divergence — for the log-likelihood-ratio),

- $\hat{p}^{(1)}, \hat{p}^{(2)}, \ldots, \hat{p}^{(m)}$ are the fractions of the n given experimental draws falling in the respective bins,
- $\hat{\theta}$ is the estimate of θ obtained from $\hat{p}^{(1)}, \hat{p}^{(2)}, \ldots, \hat{p}^{(m)}$,
- $\hat{P}^{(1)}, \hat{P}^{(2)}, \ldots, \hat{P}^{(m)}$ are the fractions of n i.i.d. draws falling in the respective bins when taking the draws from the distribution $p_0(\hat{\theta})$ assumed in (7), and
- $\hat{\Theta}$ is the estimate of the parameter θ obtained from $\hat{P}^{(1)}, \hat{P}^{(2)}, \ldots, \hat{P}^{(m)}$ (note that $\hat{\Theta}$ is not necessarily equal to $\hat{\theta}$: even under the null hypothesis, repetitions of the experiment could yield different estimates of the parameter; see also Remark 3.5).

When taking the probability that (9) occurs, only the left-hand side is random — we regard the left-hand side of (9) as a random variable and the right-hand side as a fixed number determined via the experimental data. As with any probability, to compute the probability that (9) occurs, we can calculate many independent realizations of the random variable and observe that the fraction which satisfy (9) is a good approximation to the probability when the number of realizations is large; Remark 3.4 details the accuracy of the approximation. (The procedure in the present remark follows this prescription to estimate P-values.)

Remark 3.4. The standard error of the estimate from Remark 3.3 for an exact P-value P is $\sqrt{P(1-P)/\ell}$, where ℓ is the number of simulations conducted to produce the estimate. Indeed, each simulation has probability P of producing a statistic that is greater than or equal to the statistic corresponding to an exact P-value of P. Since the simulations are all independent, the number of the ℓ simulations that produce statistics greater than or equal to that corresponding to P-value P follows the binomial distribution with ℓ trials and probability P of success in each trial. The standard deviation of the number of simulations whose statistics are at least that corresponding to P-value P is therefore $\sqrt{\ell P(1-P)}$, and so the standard deviation of the *fraction* producing such statistics is $\sqrt{P(1-P)/\ell}$. Of course, the fraction itself is the Monte Carlo estimate of the exact P-value (we use this estimate in place of the unknown P when calculating the standard error $\sqrt{P(1-P)/\ell}$).

Remark 3.5. For any family $p_0(\theta)$ of discrete probability distributions parameterized by a permutation θ that specifies the order of the bins (meaning that there exists a discrete probability distribution q such that $p_0^{(j)}(\theta) = q^{(\theta(j))}$ for all j), and for any number n of draws, the P-values defined in Remark 3.3 have the following highly desirable property: Suppose that the actual underlying distribution p of the experimental draws is equal to $p_0(\theta)$ for some (unknown) θ. Suppose further that P is the P-value for assuming (7), calculated for a particular realization of the experiment. Consider repeating the same experiment over and over, and calculating the P-value for each realization, each time using that realization's particular maximum-likelihood estimate of the parameter in the hypothesis (7). Then, the fraction of the P-values that are greater than or equal to P is equal to P in the limit of many repetitions of the experiment. This property is a compelling reason to use $d(\hat{P}, p_0(\hat{\Theta}))$ rather than $d(\hat{P}, p_0(\hat{\theta}))$ in the left-hand side of (9). This property also holds when there is no parameter in the model. Furthermore, the procedure of Remark 3.3 can be viewed as a parametric bootstrap approximation, as discussed, for example, by Romano [25], Efron and Tibshirani [7], Henze [13], and Bickel et al. [2]:

For any family $p_0(\theta)$ of discrete probability distributions, the P-values defined in Remark 3.3 have the following additional highly desirable property: Suppose that the actual underlying distribution p of the experimental draws is equal to $p_0(\theta)$ for some (unknown) θ. Consider repeating the experiment over and over, and calculating the P-value for each realization, each time using that realization's particular maximum-likelihood estimate of the parameter in the hypothesis (7). Then, the resulting P-values converge in distribution to the uniform distribution over $(0, 1)$, in the limit of large numbers of draws.

It may be somewhat fortuitous that the scheme in Remark 3.3 has so many favorable properties; indeed, Bayarri and Berger [1] and Robins et al. [24] (among others) have pointed out problems with certain generalizations.

4. Data analysis

In this section, we use several data sets to investigate the performance of goodness-of-fit statistics. Here, the root-mean-square generally performs much better than the classical statistics. We take the position that a user of statistics should not have to worry about rebinning; we discuss rebinning only briefly. We compute all P-values via Monte Carlo as in Remark 3.3; Remark 3.4 details the guaranteed accuracy of the computed P-values.

4.1. Synthetic examples

To better explicate the performance of the goodness-of-fit statistics, we first analyze some toy examples. We consider the model distribution

$$p_0^{(1)} = \frac{1}{4}, \tag{10}$$

$$p_0^{(2)} = \frac{1}{4}, \tag{11}$$

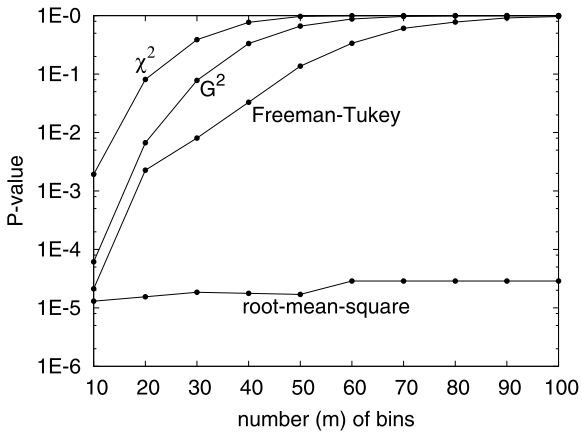

Fig. 1. P-values for the hypothesis that the model (10)–(12) agrees with the data of 15 draws in the first bin, 5 draws in the second bin, and no draw in any other bin.

and

$$p_0^{(j)} = \frac{1}{2m-4} \tag{12}$$

for $j = 3, 4, \ldots, m$. For the empirical distribution, we first use $n = 20$ draws, with 15 in the first bin, 5 in the second bin, and no draw in any other bin. This data is clearly unlikely to arise from the model specified in (10)–(12), but we would like to see exactly how well the various goodness-of-fit statistics detect the obvious discrepancy.

Fig. 1 plots the P-values for testing whether the empirical data arises from the model specified in (10)–(12). We computed the P-values via 4 000 000 Monte Carlo simulations (i.e., 4 000 000 per empirical P-value being evaluated), with each simulation taking $n = 20$ draws from the model. The root-mean-square consistently and with extremely high confidence rejects the hypothesis that the data arises from the model, whereas the classical statistics find less and less evidence for rejecting the hypothesis as the number m of bins increases; in fact, the P-values for the classical statistics get very close to 1 as m increases — the discrepancy of (12) from 0 is usually less than the discrepancy of (12) from a typical realization drawn from the model, since under the model the sum of the expected numbers of draws in bins $3, 4, \ldots, m$ is $n/2$ for any m.

Fig. 1 demonstrates that the root-mean-square can be much more powerful than the classical statistics, rejecting with nearly 100% confidence while the classical statistics report nearly 0% confidence for rejection. Moreover, the classical statistics can report P-values very close to 1 even when the data manifestly does not arise from the model. (Incidentally, the model for smaller m can be viewed as a rebinning of the model for larger m. The classical statistics do reject the model for smaller m, while asserting for larger m that there is no evidence for rejecting the model.) The performance of the classical statistics displays a dramatic dependence on the number $(m-2)$ of unlikely bins in the model, even though the data are the same for all m. This suggests a sure-fire scheme for supporting any model (no matter how invalid) with arbitrarily high P-values: just append enough irrelevant, more or less uniformly improbable bins to the model, and then report the P-values for the classical goodness-of-fit statistics. In contrast, the root-mean-square robustly and reliably rejects the invalid model, independently of the size of the model.

We will see in the following section that the classic Zipf power law behaves similarly.

For another example, we again consider the model specified in (10)–(12). For the empirical distribution, we now use $n = 96$ draws, with 36 in the first bin, 12 in the second bin, 1 each for bins $3, 4, \ldots, 50$, and no draw in any other bin. As before, this data clearly is unlikely to arise from the model specified in (10)–(12), but we would like to see exactly how well the various goodness-of-fit statistics detect the obvious discrepancy.

Fig. 2 plots the P-values for testing whether the empirical data arises from the model specified in (10)–(12). We computed the P-values via 160 000 Monte Carlo simulations (that is, 160 000 per empirical P-value being evaluated), with each simulation taking $n = 96$ draws from the model. Again the root-mean-square consistently and confidently rejects the hypothesis that the data arises from the model, whereas the classical statistics find little evidence for rejecting the manifestly invalid model.

4.2. Zipf's power law of word frequencies

Zipf popularized his eponymous law by analyzing four "chief sources of statistical data referred to in the main text" (this is a quotation from the "Notes and References" section — page 311 — of Zipf [30]); the chief source for the English language is Eldridge [8]. We revisit the data of [8] in the present subsection to assess the performance of the goodness-of-fit statistics.

We first analyze List 1 of Eldridge [8], which consists of 2 890 different English words, such that there are 13 825 words in total counting repetitions; the words come from the Buffalo Sunday News of August 8, 1909. We randomly choose

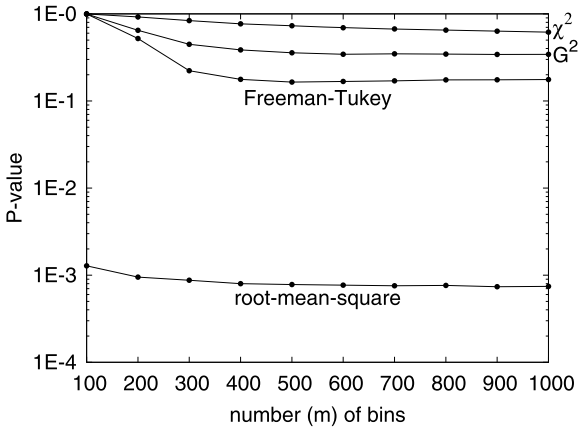

Fig. 2. P-values for the hypothesis that the model (10)–(12) agrees with the data of 36 draws in the first bin, 12 draws in the second bin, 1 draw each in bins $3, 4, \ldots, 50$, and no draw in any other bin.

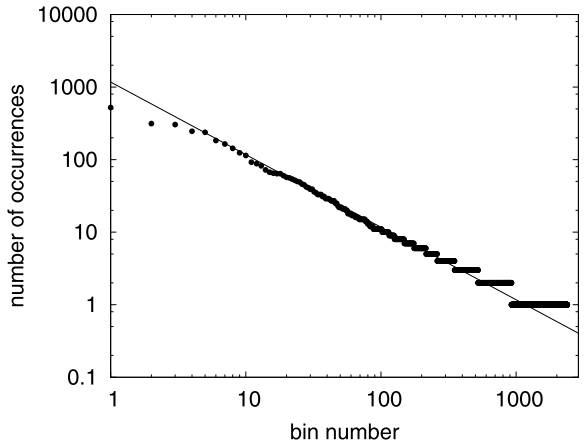

Fig. 3. Numbers of occurrences of the various words (one bin for each distinct word) in a corpus of 10 000 random draws from List 1 of Eldridge [8].

$n = 10\,000$ of the 13 825 words to obtain a corpus of $n = 10\,000$ draws over 2 890 bins. Fig. 3 plots the frequencies of the different words when sorted in rank order (so that the frequencies are nonincreasing). Using goodness-of-fit statistics we test the significance of the (null) hypothesis that the empirical draws actually arise from the Zipf distribution

$$p_0^{(j)}(\theta) = \frac{C_1}{\theta(j)} \tag{13}$$

for $j = 1, 2, \ldots, m$, where θ is a permutation of the integers $1, 2, \ldots, m$, and

$$C_1 = \frac{1}{\sum_{j=1}^{m} 1/j}; \tag{14}$$

we estimate the permutation θ via maximum-likelihood methods, that is, by sorting the frequencies: first we choose j_1 to be the number of a bin containing the greatest number of draws among all m bins, then we choose j_2 to be the number of a bin containing the greatest number of draws among the remaining $m - 1$ bins, then we choose j_3 to be the number of a bin containing the greatest among the remaining $m - 2$ bins, and so on, and finally we find θ such that $\theta(j_1) = 1, \theta(j_2) = 2, \ldots, \theta(j_m) = m$. We have to obtain the ordering θ from the data via such sorting since we do not know the proper ordering a priori.

Similarly, we do not know the proper value of the number m of bins, so in Fig. 4 we plot P-values (each computed via 40 000 Monte Carlo simulations) for varying values of m; although List 1 of Eldridge [8] involves only 2 890 distinct words, we must also include bins for words that did not appear in the original list, words whose frequencies are zeros for List 1 of Eldridge [8]. Note that Fig. 4 displays the P-values with $m = 2\,890$ for reference, even though m must be independent of the data, and so m must be substantially larger than 2 890 in order for the assumptions of goodness-of-fit testing to hold.

With respect to testing goodness-of-fit, the number m of bins is the number of words in the dictionary from which List 1 of Eldridge [8] was drawn. It is not clear a priori which dictionary is appropriate. Fortunately, the P-values for

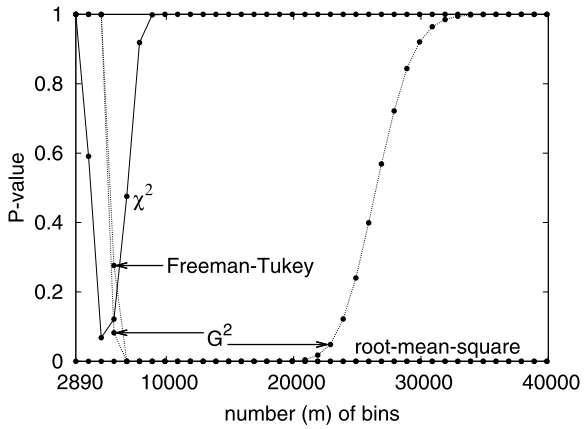

Fig. 4. P-values for the data plotted in Fig. 3 to follow the Zipf distribution.

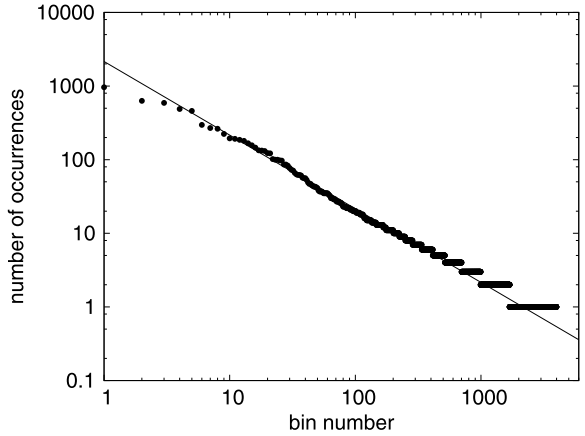

Fig. 5. Numbers of occurrences of the various words (one bin for each distinct word) in a corpus of 20 000 random draws from List 5 of Eldridge [8].

the root-mean-square are always 0 to several digits of accuracy, independent of the value of m — the root-mean-square determines that List 1 does not follow the classic Zipf distribution (defined in (13) and (14)) for any m. In contrast, the P-values for the classical statistics vary wildly depending on the value of m. In fact, for any of the classical statistics, and for any prescribed number P between 0.05 and 0.95, there is at least one value of m between 4 000 and 40 000 such that the P-value is P. Thus, without knowing the proper size of the dictionary a priori, the classical statistics are meaningless.

Unsurprisingly, analyzing List 5 of Eldridge [8] produces results analogous to those reported above for List 1. List 5 consists of 6 002 different English words, such that there are 43 989 words in total counting repetitions; the words come from amalgamating Lists 1–4 of Eldridge [8]. We randomly choose $n = 20\,000$ of the 43 989 words to obtain a corpus of $n = 20\,000$ draws over 6 002 bins. Fig. 5 plots the frequencies of the different words when sorted in rank order (so that the frequencies are nonincreasing).

Again we do not know the proper value of the number m of bins, so in Fig. 6 we plot P-values (each computed via 40 000 Monte Carlo simulations) for varying values of m; although List 5 of Eldridge [8] involves only 6 002 distinct words, we must also include bins for words that did not appear in the original list, words whose frequencies are zeros for List 5 of Eldridge [8]. Please note that Fig. 6 displays the P-values with $m = 6\,002$ for reference, even though m must be independent of the data, and so m must be substantially larger than 6 002 in order for the assumptions of goodness-of-fit testing to hold. Comparing Figs. 4 and 6 shows that the above remarks about List 1 pertain to the analysis of the larger List 5, too. Once again, without knowing the proper size of the dictionary a priori, the classical statistics are meaningless, whereas the root-mean-square is very powerful.

Interestingly, by introducing parameters θ_1, θ_2, and θ_3 to fit perfectly the bins containing the three greatest numbers of draws, a truncated power-law becomes a good fit for the corpus of 20 000 words drawn randomly from List 5 of Eldridge [8], with the number m of bins set to 7 500. Indeed, let us consider the model

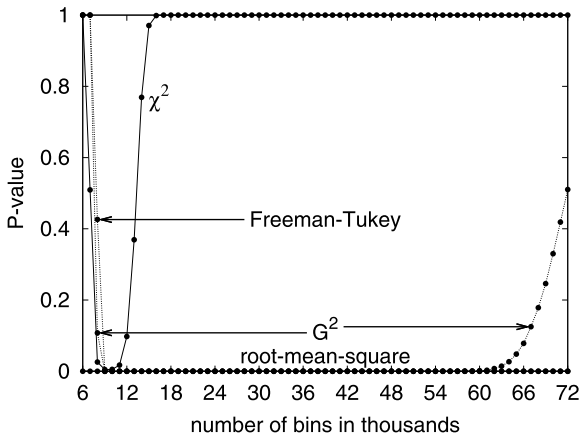

Fig. 6. P-values for the data plotted in Fig. 5 to follow the Zipf distribution.

$$p_0^{(j)}(\theta_0,\theta_1,\theta_2,\theta_3,\theta_4) = \begin{cases} \theta_1, & \theta_0(j)=1, \\ \theta_2, & \theta_0(j)=2, \\ \theta_3, & \theta_0(j)=3, \\ C/(\theta_0(j))^{\theta_4}, & \theta_0(j)=4,5,\ldots,7500, \end{cases} \tag{15}$$

where

$$C = C_{\theta_1,\theta_2,\theta_3,\theta_4} = \frac{1-\theta_1-\theta_2-\theta_3}{\sum_{j=4}^{7500} 1/j^{\theta_4}}, \tag{16}$$

with θ_0 being a permutation of the integers $1,2,\ldots,7500$, and $\theta_1,\theta_2,\theta_3,\theta_4$ being nonnegative real numbers; we estimate $\theta_0,\theta_1,\theta_2,\theta_3,\theta_4$ via maximum-likelihood methods, determining θ_0 by sorting as discussed above, and setting θ_1, θ_2, and θ_3 to be the three greatest relative frequencies. This model fits the empirical data exactly in the bins whose probabilities under the model are θ_1,θ_2, and θ_3 — there will be no discrepancy between the data and the model in those bins — so that these bins do not contribute to any goodness-of-fit statistic, aside from altering the number of draws in the remaining bins. Of the 20 000 total draws in the given experimental data, 16 486 do not fall in the bins associated with the three most frequently occurring words. The maximum-likelihood estimate of the power-law exponent θ_4 for the experimental data turns out to be about 1.0484.

For the model defined in (15) and (16), the P-values calculated via 4 000 000 Monte Carlo simulations are

- χ^2: 0.510;
- G^2: 0.998;
- Freeman–Tukey: 1.000;
- root-mean-square: 0.587.

Thus, all four statistics indicate that the truncated power-law model defined in (15) and (16) is a good fit. This is in accord with Fig. 5, in which all but the three greatest frequencies appear to follow a truncated power-law.

4.3. A Poisson law for radioactive decays

Table 1 summarizes the classic example of a Poisson-distributed experiment in radioactive decay of Rutherford et al. [26]; Fig. 7 plots the data, along with the Poisson distribution whose mean is the same as the data's. Fig. 8 reports the P-values for testing whether the data, while retaining only bins $1,2,\ldots,m$, are distributed according to a Poisson distribution (the model Poisson distribution is also truncated to the first m bins, with the mean estimated from the data). Since the total number n of draws depends little on the numbers in bins $13,14,15,\ldots$, the truncation amounts to ignoring draws in bins $m+1, m+2, m+3,\ldots$ when $m \geqslant 12$, and demonstrates that the scant experimental draws in bins 13–15 strongly influence the P-values of the classical statistics. We computed the P-values via 40 000 Monte Carlo simulations (for each number m of bins and each of the four statistics), estimating the mean of the model Poisson distribution for each simulated data set. All four goodness-of-fit statistics indicate reasonably good agreement between the data and a Poisson distribution; the classical statistics are very sensitive in the tail to discrepancies between the data and the model distribution, whereas the root-mean-square is relatively insensitive to the truncation after 12 or more bins.

Table 1
Numbers of α-particles emitted by a film of polonium in 2608 intervals of 7.5 seconds.

Bin number	Number of particles observed in an interval of 7.5 seconds	Number of such intervals
1	0	57
2	1	203
3	2	383
4	3	525
5	4	532
6	5	408
7	6	273
8	7	139
9	8	45
10	9	27
11	10	10
12	11	4
13	12	0
14	13	1
15	14	1
16, 17, 18, ...	15, 16, 17, ...	0
1, 2, 3, 4, 5, ...	0, 1, 2, 3, 4, ...	2608

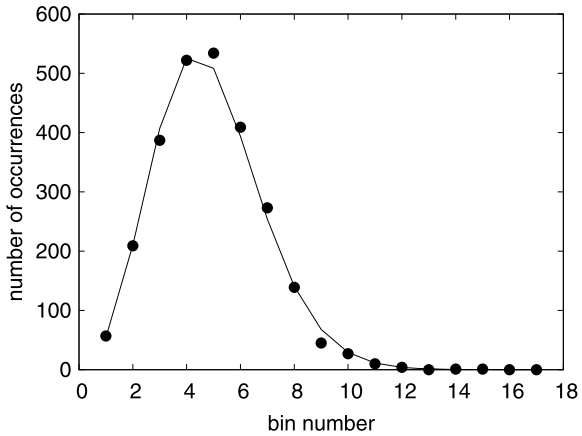

Fig. 7. The data in Table 1 (the dots) and the best-fit Poisson distribution (the lines).

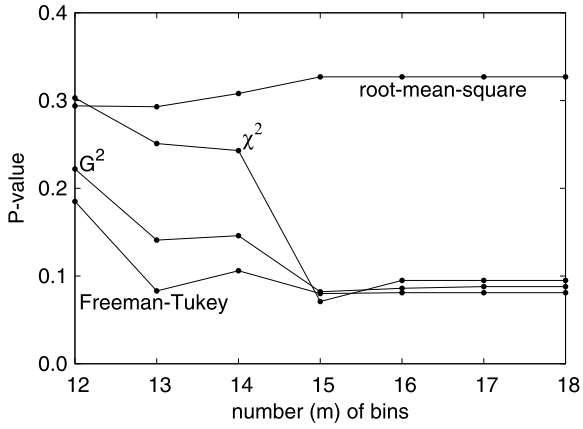

Fig. 8. P-values for the distribution of Table 1 to be Poisson.

Table 2
Numbers of yeast cells in 400 squares of a hæmacytometer.

Bin number	Number of yeast in a square	Number of such squares
1	0	0
2	1	20
3	2	43
4	3	53
5	4	86
6	5	70
7	6	54
8	7	37
9	8	18
10	9	10
11	10	5
12	11	2
13	12	2
14, 15, 16, ...	13, 14, 15, ...	0
1, 2, 3, 4, 5, ...	0, 1, 2, 3, 4, ...	400

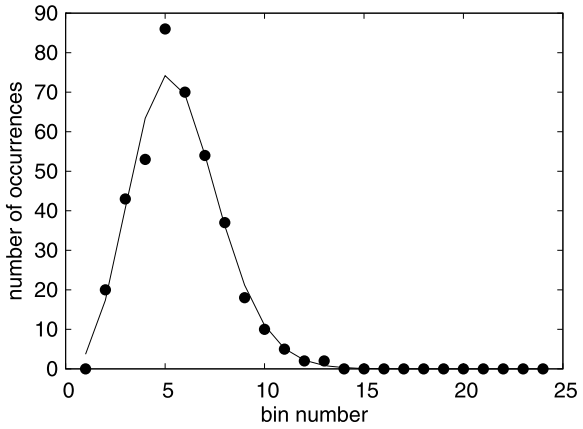

Fig. 9. The data in Table 2 (the dots) and the best-fit Poisson distribution (the lines).

4.4. A Poisson law for counting with a hæmacytometer

Page 357 of Student [27] reports on the number of yeast cells observed in each of 400 squares in a hæmacytometer microscope slide. Table 2 displays the counts; Fig. 9 plots them, along with the Poisson distribution whose mean matches the data's. The P-values for the data to arise from a Poisson distribution (with the mean estimated from the data) are

- χ^2: 0.627;
- G^2: 0.365;
- Freeman–Tukey: 0.111;
- root-mean-square: 0.490.

We calculated the P-values via 4 000 000 Monte Carlo simulations, estimating the mean of the model Poisson distribution for each simulated data set. Evidently, all four statistics report that a Poisson distribution is a reasonably good model for the experimental data.

4.5. A Hardy–Weinberg law for Rhesus blood groups

In a population with suitably random mating, the proportions of pairs of Rhesus haplotypes in members of the population (each member has one pair) can be expected to follow the Hardy–Weinberg law discussed by Guo and Thompson [12], namely to arise via random sampling from the model

$$p_0^{(j,k)}(\theta_1, \theta_2, \ldots, \theta_9) = \begin{cases} 2 \cdot \theta_j \cdot \theta_k, & j > k, \\ (\theta_k)^2, & j = k, \end{cases} \tag{17}$$

for $j, k = 1, 2, \ldots, 9$ with $j \geqslant k$, under the constraint that

Table 3
Frequencies of pairs of Rhesus haplotypes.

j	k								
	1	2	3	4	5	6	7	8	9
1	1236								
2	120	3							
3	18	0	0						
4	982	55	7	249					
5	32	1	0	12	0				
6	2582	132	20	1162	29	1312			
7	6	0	0	4	0	4	0		
8	2	0	0	0	0	0	0	0	
9	115	5	2	53	1	149	0	0	4

$$\sum_{j=1}^{9} \theta_j = 1, \tag{18}$$

where the parameters $\theta_1, \theta_2, \ldots, \theta_9$ are the proportions of the nine Rhesus haplotypes in the population (naturally, their maximum-likelihood estimates are the proportions of the haplotypes in the given data). For $j, k = 1, 2, \ldots, 9$ with $j \geqslant k$, therefore, $p_0^{(j,k)}$ is the expected probability that the pair of haplotypes in the genome of an individual is the pair j and k, given the parameters $\theta_1, \theta_2, \ldots, \theta_9$.

In this formulation, the hypothesis of suitably random mating entails that the members of the sample population are i.i.d. draws from the model specified in (17); if a goodness-of-fit statistic rejects the model with high confidence, then we can be confident that mating has not been suitably random. Table 3 provides data on $n = 8297$ individuals; we duplicated Fig. 3 of Guo and Thompson [12] to obtain Table 3.

The P-values calculated via 4 000 000 Monte Carlo simulations are

- χ^2: 0.693;
- G^2: 0.600;
- Freeman–Tukey: 0.562;
- negative log-likelihood (see Remark 4.2 below): 0.649;
- root-mean-square: 0.039.

Unlike the root-mean-square, the classical statistics are blind to the significant discrepancy between the data and the Hardy–Weinberg model. The P-values associated with the classical statistics are over an order of magnitude larger than the P-value associated with the root-mean-square.

Remark 4.1. For the example of the present subsection, rejecting the null hypothesis (5) from Section 3 might seem in principle to be more interesting than rejecting the assumption (7). Fortunately, the difference between (5) and (7) is essentially irrelevant for the root-mean-square in this example. Indeed, the root-mean-square is not very sensitive to bins associated with the parameters whose estimated values are potentially inaccurate — the potentially inaccurate estimates are all small, and the root-mean-square is not very sensitive to bins whose probabilities under the model are small relative to others.

Remark 4.2. The term "negative log-likelihood" used in the present section refers to the statistic that is simply the negative of the logarithm of the likelihood. The negative log-likelihood is the same statistic used in the generalization of Fisher's exact test discussed by Guo and Thompson [12]; unlike G^2, this statistic involves only one likelihood, not the ratio of two. We mention the negative log-likelihood just to facilitate comparisons; we are not asserting that the likelihood on its own (rather than in a ratio) is a good gauge of the relative sizes of deviations from a model.

Remark 4.3. Table 4 provides data on $n = 45$ individuals from the other set of real-world measurements given by Guo and Thompson [12]; we duplicated Fig. 2 of Guo and Thompson [12] to obtain Table 4. The associated Hardy–Weinberg model is then the same as (17), but with only four parameters, $\theta_1, \theta_2, \theta_3, \theta_4$, such that $\sum_{j=1}^{4} \theta_j = 1$. The P-values calculated via 4 000 000 Monte Carlo simulations are

- χ^2: 0.021;
- G^2: 0.013;
- Freeman–Tukey: 0.027;
- negative log-likelihood (see Remark 4.2 above): 0.016;
- root-mean-square: 0.0019.

The root-mean-square is more powerful than the classical statistics on this data set, too.

Table 4
Frequencies of genotypes.

j	k			
	1	2	3	4
1	0			
2	3	1		
3	5	18	1	
4	3	7	5	2

Table 5
Self-reported physical health for matched pairs of Asian Americans.

	Foreign-born					
		Excellent	Very good	Good	Fair	Poor
US-born	Excellent	10	21	22	5	0
	Very good	24	53	43	15	3
	Good	21	43	34	11	0
	Fair	3	11	8	4	1
	Poor	1	1	1	0	0

Table 6
A variation on Table 5.

	Foreign-born					
		Excellent	Very good	Good	Fair	Poor
US-born	Excellent	10	21	22	5	0
	Very good	24	53	*56*	15	3
	Good	21	*30*	34	11	0
	Fair	3	11	8	4	1
	Poor	1	1	1	0	0

4.6. Symmetry between the self-reported health assessments of foreign- and US-born Asian Americans

Using propensity scores, Erosheva et al. [9] matched each of 335 surveyed foreign-born Asian Americans to a similar surveyed US-born Asian American. Table 5 duplicates Table 4 of Erosheva et al. [9], tabulating the numbers of matched pairs reporting various combinations of physical health; the propensity scores were generated without reference to the health ratings. Table 5 does not reveal any significant difference between foreign-born Asian Americans' ratings of their health and US-born Asian Americans'. Indeed, the P-values calculated via 4 000 000 Monte Carlo simulations for testing the symmetry of Table 5 are

- χ^2: 0.784;
- G^2: 0.739;
- Freeman–Tukey: 0.642;
- root-mean-square: 0.973.

After noting that χ^2 does not reveal any statistically significant asymmetry in Table 5, Erosheva et al. [9] reported that, "to address the issue of power of this test, we investigated what is the smallest departure from symmetry that our test could detect…". Such an investigation requires considering modifications to Table 5. Table 6 provides one possible modification. The P-values calculated via 4 000 000 Monte Carlo simulations for testing the symmetry of Table 6 are

- χ^2: 0.109;
- G^2: 0.123;
- Freeman–Tukey: 0.155;
- root-mean-square: 0.014.

Evidently, the root-mean-square is more powerful for detecting the asymmetry of Table 6.

Table 7 provides another hypothetical cross-tabulation. The P-values calculated via 64 000 000 Monte Carlo simulations for testing the symmetry of Table 7 are

- χ^2: 0.0015;
- G^2: 0.00016;
- Freeman–Tukey: 0.000006, i.e., 6E–6;
- root-mean-square: 0.131.

Table 7
Another variation on Table 5.

	Foreign-born					
		Excellent	Very good	Good	Fair	Poor
US-born	Excellent	10	21	22	5	0
	Very good	24	53	43	15	3
	Good	21	43	34	*19*	0
	Fair	3	11	*0*	4	1
	Poor	1	1	1	0	0

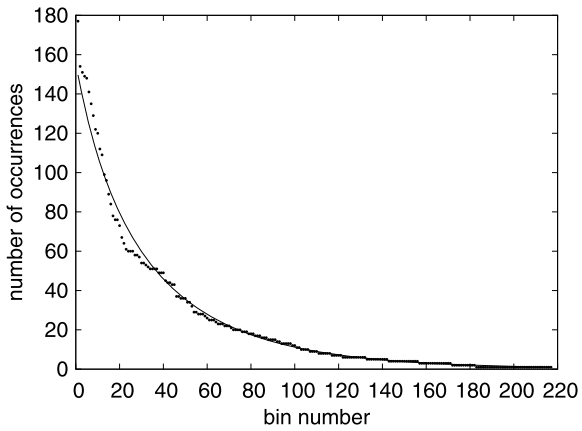

Fig. 10. Numbers of specimens (the dots) from 217 species of butterflies (one bin per species), and the best-fit distribution (the lines).

The classical statistics are much more powerful for detecting the asymmetry of Table 7, contrasting how the root-mean-square is more powerful for detecting the asymmetry of Table 6. Indeed, 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 and an asymptotically equivalent variation of χ^2 such as the log-likelihood-ratio G^2.

4.7. A modified geometric law for the species of butterflies

Fisher et al. [10] reported on 5300 butterflies from 217 readily identified species (these exclude the 23 most common readily identified species) that they collected via random sampling at the Rothamsted Experimental Station in England. Fig. 10 plots the numbers of individual butterflies collected from the 217 species when sorted in rank order (so that the numbers are nonincreasing).

To build a model appropriate for Fig. 10, we must include a permutation of the bins as a parameter, since we have sorted the data (see Section 4.2 for further discussion of sorting and permutations). We take the model to be

$$p_0^{(j)}(\theta_0, \theta_1) = A_{\theta_1} \frac{(\theta_1)^{\theta_0(j)}}{\sqrt{\theta_0(j) + 23}} \tag{19}$$

for $j = 1, 2, \ldots, 217$, where θ_0 is a permutation of the integers $1, 2, \ldots, 217$, the parameter θ_1 is a positive real number less than 1, and

$$A_{\theta_1} = \frac{1}{\sum_{j=1}^{217} (\theta_1)^j / \sqrt{j + 23}}; \tag{20}$$

we estimate θ_0 and θ_1 via maximum-likelihood methods (thus obtaining θ_0 by sorting the frequencies into nonincreasing order). Please note that this model is not very carefully chosen — the model is just a truncated geometric distribution weighted by the nonsingular function $1/\sqrt{\theta_0(j) + 23}$, with 23 being the number of common species omitted from the collection. More complicated models may fit better.

The P-values calculated via 4 000 000 Monte Carlo simulations are

- χ^2: 0.0050;
- G^2: 0.349;
- Freeman–Tukey: 0.951;
- root-mean-square: 0.00002, i.e., 2E–5.

As Fig. 10 indicates, the discrepancy between the empirical data and the model is substantial, and, given the large number of draws (5300), cannot be due solely to random fluctuations. The log-likelihood-ratio (G^2) and Freeman–Tukey statistics are unable to detect this discrepancy, while the root-mean-square easily determines that the discrepancy is very highly significant.

5. The power and efficiency of the root-mean-square

In this section, we consider many numerical experiments and models, plotting the numbers of draws required for goodness-of-fit statistics to detect divergence from the models. We consider both fully specified models and parameterized models. To quantify a statistic's success at detecting discrepancies from the models, we use the formulation of the following remark.

Remark 5.1. 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 P-value is at most *1%* for *99%* of 40 000 simulations, with each simulation generating n i.i.d. draws according to the actual distribution. We computed the P-values by conducting another 40 000 simulations, with each simulation generating n i.i.d. draws according to the model distribution. Appendix A of Perkins et al. [18] uses a weaker notion of "distinguish" — in Appendix A 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 P-value is at most *5%* for *95%* of 40 000 simulations, while running simulations and computing P-values exactly as for the plots in the present section.

Remark 5.2. To compute the P-values for each example in Section 5.2, we should in principle calculate the maximum-likelihood estimate $\hat{\theta}$ for each of 40 000 simulations and (for each goodness-of-fit statistic) use these estimates to perform $(40\,000)^2$ times the three-step procedure described in Remark 3.3. The computational costs for generating the plots in Section 5.2 would then be excessive. Instead, when computing the P-values as a function of the value of the statistic under consideration, we calculated $\hat{\theta}$ only once, using as the empirical data 1 000 000 draws from the underlying distribution, and (for each goodness-of-fit statistic) performed 40 000 times the three-step procedure described in Remark 3.3, using the single value of $\hat{\theta}$ (but many values of $\tilde{\theta}$ from Remark 3.3). The parameter estimates did not vary much over the 40 000 simulations, so approximating the P-values thus is accurate. Furthermore, when the parameter is just a permutation, as in Section 5.2.6, the "approximation" described in the present remark is exactly equivalent to recomputing the P-values 40 000 times — we are not making any approximation at all. Please note that we did recalculate the maximum-likelihood estimate $\hat{\theta}$ (and $\tilde{\theta}$ from Remark 3.3) for each of 40 000 simulations when computing the values of the statistics for the simulation; however, when calculating the P-values as a function of the values of the statistics, we always drew from the model distribution associated with the same value of the parameter.

Remark 5.3. 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 and an asymptotically equivalent variation of χ^2, such as the log-likelihood-ratio or "G^2" statistic.

5.1. Examples without parameter estimation

5.1.1. A simple, illustrative example

Let us first specify the model distribution to be

$$p_0^{(1)} = \frac{1}{4}, \tag{21}$$

$$p_0^{(2)} = \frac{1}{4}, \tag{22}$$

and

$$p_0^{(j)} = \frac{1}{2m-4} \tag{23}$$

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

$$p^{(1)} = \frac{3}{8}, \tag{24}$$

$$p^{(2)} = \frac{1}{8}, \tag{25}$$

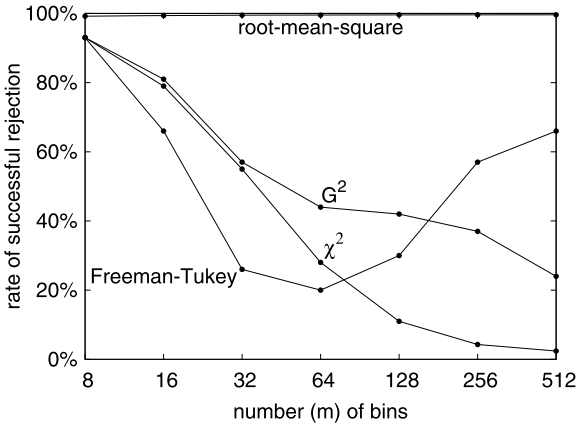

Fig. 11. First example, with $n = 200$ draws; see Section 5.1.1.

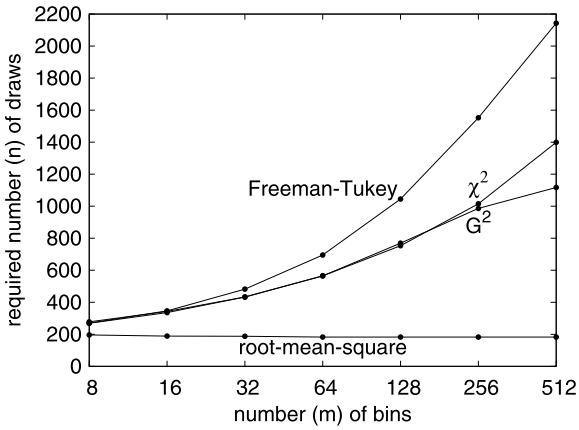

Fig. 12. First example (statistical "efficiency"); see Section 5.1.1.

and

$$p^{(j)} = p_0^{(j)} \tag{26}$$

for $j = 3, 4, \ldots, m$, where $p_0^{(3)}, p_0^{(4)}, \ldots, p_0^{(m)}$ are the same as in (23).

Fig. 11 plots the percentage of 40 000 simulations, each generating 200 i.i.d. draws according to the actual distribution defined in (24)–(26), that are successfully detected as not arising from the model distribution at the 1% significance level. We computed the P-values by conducting 40 000 simulations, each generating 200 i.i.d. draws according to the model distribution defined in (21)–(23). Fig. 11 shows that the root-mean-square is successful in at least 99% of the simulations, while the classical χ^2 statistic fails often, succeeding in less than 80% of the simulations for $m = 16$, and less than 5% for $m \geqslant 256$.

Fig. 12 plots the number n of draws required to distinguish the actual distribution defined in (24)–(26) from the model distribution defined in (21)–(23). Remark 5.1 above specifies what we mean by "distinguish". Fig. 12 shows that the root-mean-square requires only about $n = 185$ draws for any number m of bins, while the classical χ^2 statistic requires 90% more draws for $m = 16$, and greater than 300% more for $m \geqslant 128$. Furthermore, the classical χ^2 statistic requires increasingly many draws as the number m of bins increases, unlike the root-mean-square.

5.1.2. Truncated power-laws

Next, let us specify the model distribution to be

$$p_0^{(j)} = \frac{C_1}{j} \tag{27}$$

for $j = 1, 2, \ldots, m$, where

$$C_1 = \frac{1}{\sum_{j=1}^{m} 1/j}. \tag{28}$$

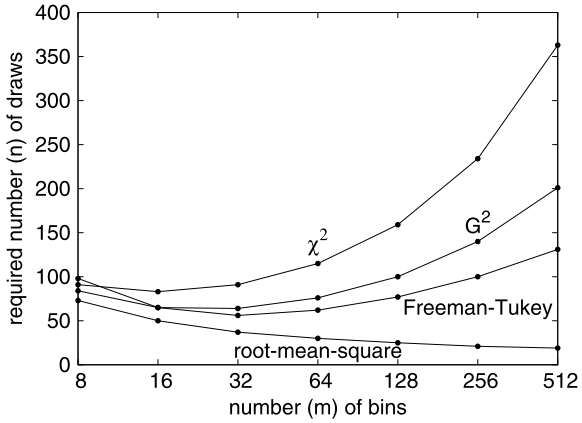

Fig. 13. Second example; see Section 5.1.2.

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

$$p^{(j)} = \frac{C_2}{j^2} \tag{29}$$

for $j = 1, 2, \ldots, m$, where

$$C_2 = \frac{1}{\sum_{j=1}^{m} 1/j^2}. \tag{30}$$

Fig. 13 plots the number n of draws required to distinguish the actual distribution defined in (29) and (30) from the model distribution defined in (27) and (28). Remark 5.1 above specifies what we mean by "distinguish". Fig. 13 shows that the classical χ^2 statistic requires increasingly many draws as the number m of bins increases, while the root-mean-square exhibits the opposite behavior.

5.1.3. Additional truncated power-laws

Let us again specify the model distribution to be

$$p_0^{(j)} = \frac{C_1}{j} \tag{31}$$

for $j = 1, 2, \ldots, m$, where

$$C_1 = \frac{1}{\sum_{j=1}^{m} 1/j}. \tag{32}$$

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

$$p^{(j)} = \frac{C_{1/2}}{\sqrt{j}} \tag{33}$$

for $j = 1, 2, \ldots, m$, where

$$C_{1/2} = \frac{1}{\sum_{j=1}^{m} 1/\sqrt{j}}. \tag{34}$$

Fig. 14 plots the number n of draws required to distinguish the actual distribution defined in (33) and (34) from the model distribution defined in (31) and (32). Remark 5.1 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 5.3 at the beginning of the present section.

5.1.4. Additional truncated power-laws, reversed

Let us next specify the model distribution to be

$$p_0^{(j)} = \frac{C_{1/2}}{\sqrt{j}} \tag{35}$$

for $j = 1, 2, \ldots, m$, where

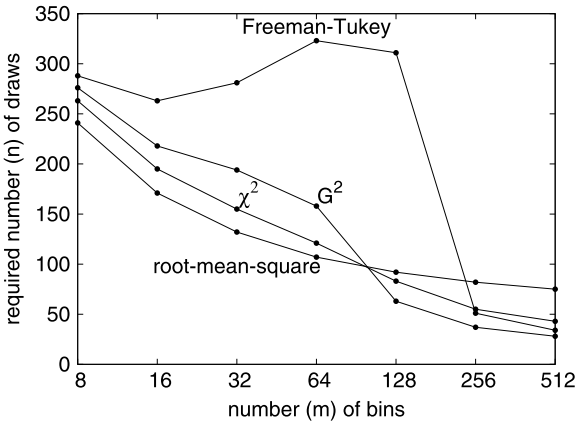

Fig. 14. Third example; see Section 5.1.3.

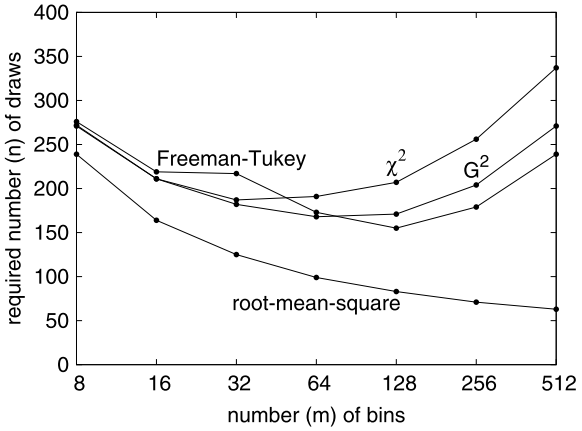

Fig. 15. Fourth example; see Section 5.1.4.

$$C_{1/2} = \frac{1}{\sum_{j=1}^{m} 1/\sqrt{j}}. \tag{36}$$

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

$$p^{(j)} = \frac{C_1}{j} \tag{37}$$

for $j = 1, 2, \ldots, m$, where

$$C_1 = \frac{1}{\sum_{j=1}^{m} 1/j}. \tag{38}$$

Fig. 15 plots the number n of draws required to distinguish the actual distribution defined in (37) and (38) from the model distribution defined in (35) and (36). Remark 5.1 above specifies what we mean by "distinguish". Fig. 15 shows that the classical χ^2 statistic requires many times more draws than the root-mean-square, as the number m of bins increases.

5.1.5. A final example with fully specified truncated power-laws

Let us next specify the model distribution to be

$$p_0^{(j)} = \frac{C_2}{j^2} \tag{39}$$

for $j = 1, 2, \ldots, m$, where

$$C_2 = \frac{1}{\sum_{j=1}^{m} 1/j^2}. \tag{40}$$

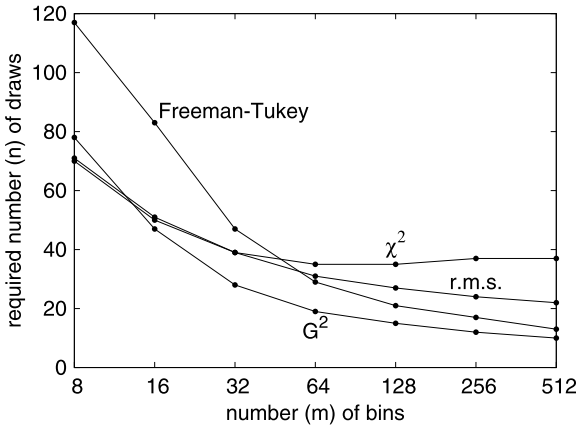

Fig. 16. Fifth example; see Section 5.1.5.

We again consider n i.i.d. draws from the distribution

$$p^{(j)} = \frac{C_1}{j} \tag{41}$$

for $j = 1, 2, \ldots, m$, where

$$C_1 = \frac{1}{\sum_{j=1}^{m} 1/j}. \tag{42}$$

Fig. 16 plots the number n of draws required to distinguish the actual distribution defined in (41) and (42) from the model distribution defined in (39) and (40). Remark 5.1 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 5.3 at the beginning of the present section.

5.1.6. Modified Poisson distributions

Let us specify the model distribution to be the (truncated) Poisson distribution

$$p_0^{(j)} = \frac{B_{3m/8}(\frac{3m}{8})^{j-1}}{(j-1)!} \tag{43}$$

for $j = 1, 2, \ldots, m$, where

$$B_{3m/8} = \frac{1}{\sum_{j=1}^{m}(\frac{3m}{8})^{j-1}/(j-1)!}. \tag{44}$$

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

$$p^{((3m/8)-1)} = S/10, \tag{45}$$

$$p^{(3m/8)} = 4S/5, \tag{46}$$

$$p^{((3m/8)+1)} = S/10, \tag{47}$$

$$S = p_0^{((3m/8)-1)} + p_0^{(3m/8)} + p_0^{((3m/8)+1)}, \tag{48}$$

$$p^{(j)} = p_0^{(j)} \tag{49}$$

for the remaining values of j (for $j = 1, 2, \ldots, \frac{3m}{8} - 2$ and $j = \frac{3m}{8} + 2, \frac{3m}{8} + 3, \ldots, m$).

Fig. 17 plots the number n of draws required to distinguish the actual distribution defined in (45)–(49) from the model distribution defined in (43) and (44). Remark 5.1 above specifies what we mean by "distinguish".

5.1.7. A truncated power-law and a truncated geometric distribution

Let us finally specify the model distribution to be

$$p_0^{(j)} = \frac{C_1}{j} \tag{50}$$

for $j = 1, 2, \ldots, 100$, where

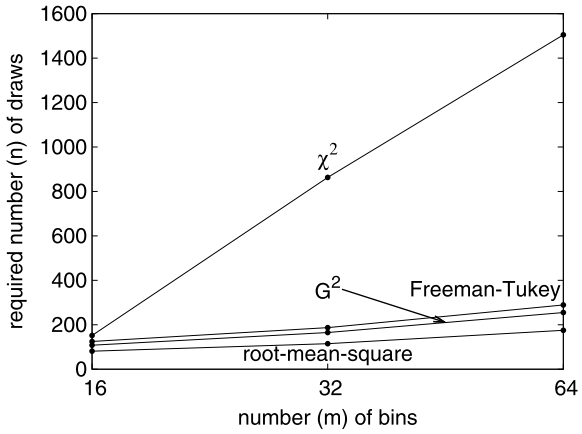

Fig. 17. Sixth example; see Section 5.1.6.

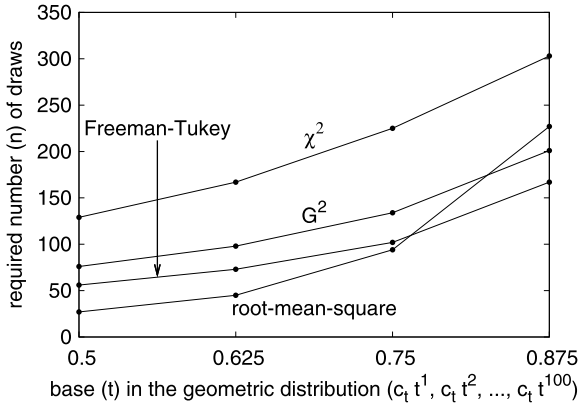

Fig. 18. Seventh example; see Section 5.1.7.

$$C_1 = \frac{1}{\sum_{j=1}^{100} 1/j}. \tag{51}$$

We consider n i.i.d. draws from the (truncated) geometric distribution

$$p^{(j)} = c_t t^j \tag{52}$$

for $j = 1, 2, \ldots, 100$, where

$$c_t = \frac{1}{\sum_{j=1}^{100} t^j}; \tag{53}$$

Fig. 18 considers several values for t.

Fig. 18 plots the number n of draws required to distinguish the actual distribution defined in (52) and (53) from the model distribution defined in (50) and (51). Remark 5.1 above specifies what we mean by "distinguish". See the next section, Section 5.2.1, for a similar example, this time involving parameter estimation.

5.2. Examples with parameter estimation

5.2.1. A truncated power-law and a truncated geometric distribution

We turn now to models involving parameter estimation, as detailed by Perkins et al. [20]. Let us specify the model distribution to be the Zipf distribution

$$p_0^{(j)}(\theta) = \frac{C_\theta}{j^\theta} \tag{54}$$

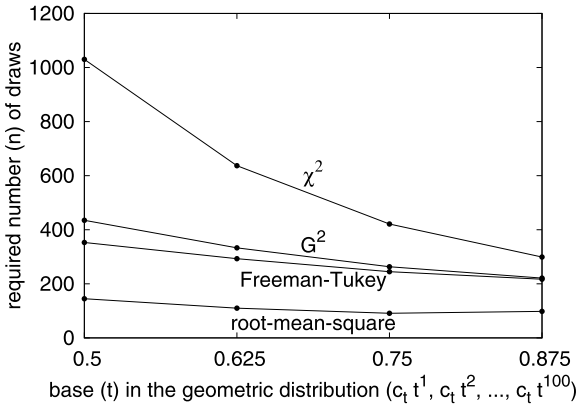

Fig. 19. First example; see Section 5.2.1.

for $j = 1, 2, \ldots, 100$, where

$$C_\theta = \frac{1}{\sum_{j=1}^{100} 1/j^\theta}; \tag{55}$$

we estimate the parameter θ via maximum-likelihood methods. We consider n i.i.d. draws from the (truncated) geometric distribution

$$p^{(j)} = c_t t^j \tag{56}$$

for $j = 1, 2, \ldots, 100$, where

$$c_t = \frac{1}{\sum_{j=1}^{100} t^j}; \tag{57}$$

Fig. 19 considers several values for t.

Fig. 19 plots the number n of draws required to distinguish the actual distribution defined in (56) and (57) from the model distribution defined in (54) and (55), estimating the parameter θ in (54) and (55) via maximum-likelihood methods. Remark 5.1 above specifies what we mean by "distinguish".

5.2.2. A rebinned geometric distribution and a truncated power-law

Let us specify the model distribution to be

$$p_0^{(j)}(\theta) = \theta^{j-1}(1-\theta) \tag{58}$$

for $j = 1, 2, \ldots, 99$, and

$$p_0^{(100)}(\theta) = \theta^{99}; \tag{59}$$

we estimate the parameter θ via maximum-likelihood methods. We consider n i.i.d. draws from the Zipf distribution

$$p^{(j)} = \frac{C_t}{j^t} \tag{60}$$

for $j = 1, 2, \ldots, 100$, where

$$C_t = \frac{1}{\sum_{j=1}^{100} 1/j^t}; \tag{61}$$

Fig. 20 considers several values for t.

Fig. 20 plots the number n of draws required to distinguish the actual distribution defined in (60) and (61) from the model distribution defined in (58) and (59), estimating the parameter θ in (58) and (59) via maximum-likelihood methods. Remark 5.1 above specifies what we mean by "distinguish".

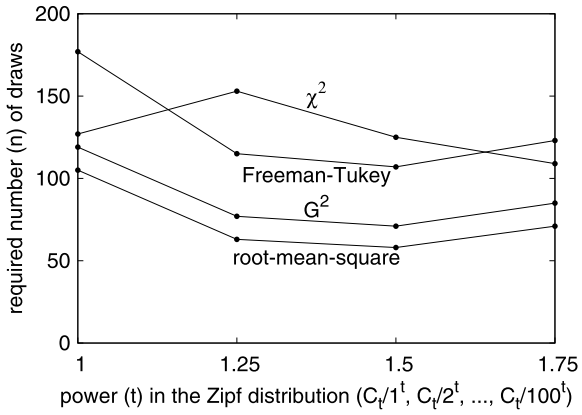

Fig. 20. Second example; see Section 5.2.2.

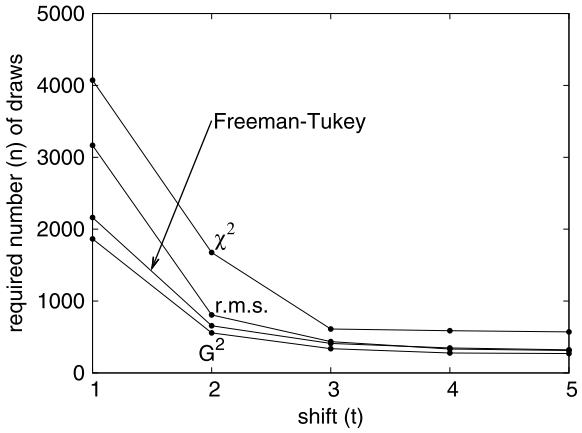

Fig. 21. Third example; see Section 5.2.3.

5.2.3. Truncated shifted Poisson distributions

Let us specify the model distribution to be the (truncated) Poisson distribution

$$p_0^{(j)}(\theta) = \frac{B_\theta\, \theta^{j-1}}{(j-1)!} \tag{62}$$

for $j = 1, 2, \ldots, 21$, where

$$B_\theta = \frac{1}{\sum_{j=1}^{21} \theta^{j-1}/(j-1)!}; \tag{63}$$

we estimate the parameter θ via maximum-likelihood methods. We consider n i.i.d. draws from the distribution

$$p^{(j)} = \frac{\tilde{B}_t\, 5^{j-1+t}}{(j-1+t)!} \tag{64}$$

for $j = 1, 2, \ldots, 21$, where

$$\tilde{B}_t = \frac{1}{\sum_{j=1}^{21} 5^{j-1+t}/(j-1+t)!}; \tag{65}$$

Fig. 21 considers several values for t. Clearly, $p^{(j)} = p_0^{(j)}(5)$ for $j = 1, 2, \ldots, 21$, if $t = 0$.

Fig. 21 plots the number n of draws required to distinguish the actual distribution defined in (64) and (65) from the model distribution defined in (62) and (63), estimating the parameter θ in (62) and (63) via maximum-likelihood methods. Remark 5.1 above specifies what we mean by "distinguish".

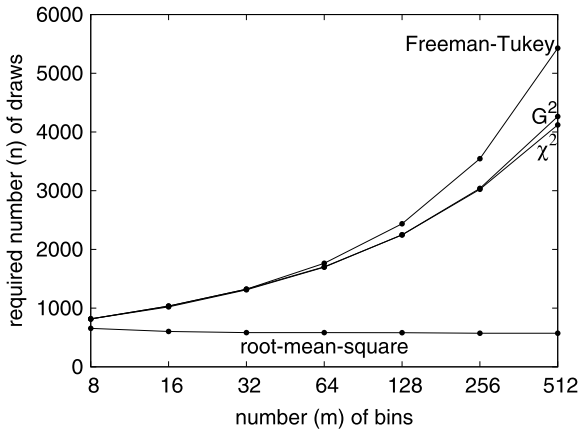

Fig. 22. Fourth example; see Section 5.2.4.

5.2.4. An example with a uniform tail

Let us specify the model distribution to be

$$p_0^{(1)}(\theta) = \theta, \tag{66}$$

$$p_0^{(2)}(\theta) = \theta, \tag{67}$$

$$p_0^{(3)}(\theta) = \frac{1}{2} - 2\theta, \tag{68}$$

$$p_0^{(j)}(\theta) = \frac{1}{2m-6} \tag{69}$$

for $j = 4, 5, \ldots, m$; we estimate the parameter θ via maximum-likelihood methods. We consider n i.i.d. draws from the distribution

$$p^{(1)} = \frac{1}{4}, \tag{70}$$

$$p^{(2)} = \frac{1}{8}, \tag{71}$$

$$p^{(3)} = \frac{1}{8}, \tag{72}$$

$$p^{(j)} = \frac{1}{2m-6} \tag{73}$$

for $j = 4, 5, \ldots, m$.

Fig. 22 plots the number n of draws required to distinguish the actual distribution defined in (70)–(73) from the model distribution defined in (66)–(69), estimating the parameter θ in (66)–(69) via maximum-likelihood methods. Remark 5.1 above specifies what we mean by "distinguish".

5.2.5. A model with an integer-valued parameter

Let us specify the model distribution to be

$$p_0^{(j)}(\theta) = \frac{1}{2\theta} \tag{74}$$

for $j = 1, 2, \ldots, \theta$, and

$$p_0^{(j)}(\theta) = \frac{1}{2(m-\theta)} \tag{75}$$

for $j = \theta + 1, \theta + 2, \ldots, m$; we estimate the parameter θ via maximum-likelihood methods. We consider n i.i.d. draws from the distribution

$$p^{(1)} = \frac{1}{4}, \tag{76}$$

$$p^{(2)} = \frac{1}{4}, \tag{77}$$

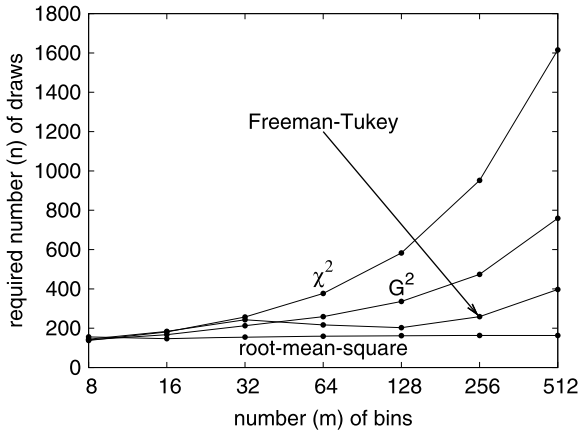

Fig. 23. Fifth example; see Section 5.2.5.

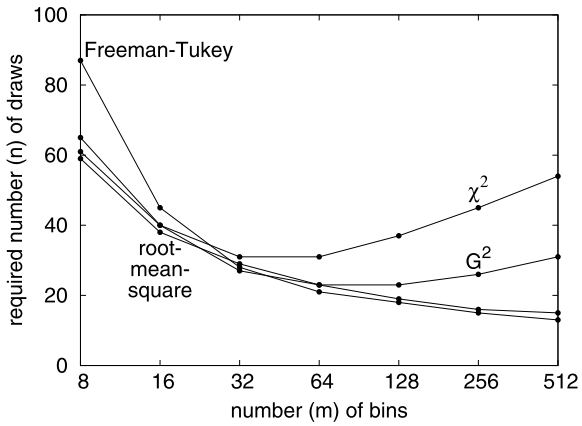

Fig. 24. Sixth example; see Section 5.2.6.

$$p^{(3)} = \frac{1}{4}, \tag{78}$$

and

$$p^{(j)} = \frac{1}{4m - 12} \tag{79}$$

for $j = 4, 5, \ldots, m$.

Fig. 23 plots the number n of draws required to distinguish the actual distribution defined in (76)–(79) from the model distribution defined in (74) and (75), estimating the parameter θ in (74) and (75) via maximum-likelihood methods. Remark 5.1 above specifies what we mean by "distinguish".

5.2.6. Truncated power-laws parameterized with a permutation

Let us specify the model to be the Zipf distribution

$$p_0^{(j)}(\theta) = \frac{C_1}{\theta(j)} \tag{80}$$

for $j = 1, 2, \ldots, m$, where θ is a permutation of the integers $1, 2, \ldots, m$, and

$$C_1 = \frac{1}{\sum_{j=1}^{m} 1/j}; \tag{81}$$

we estimate the permutation θ via maximum-likelihood methods, that is, by sorting the frequencies: first we choose j_1 to be the number of a bin containing the greatest number of draws among all m bins, then we choose j_2 to be the number of a bin containing the greatest number of draws among the remaining $m - 1$ bins, then we choose j_3 to be the number of a bin containing the greatest among the remaining $m - 2$ bins, and so on, and finally we find θ such that $\theta(j_1) = 1, \theta(j_2) = 2, \ldots, \ \theta(j_m) = m$.

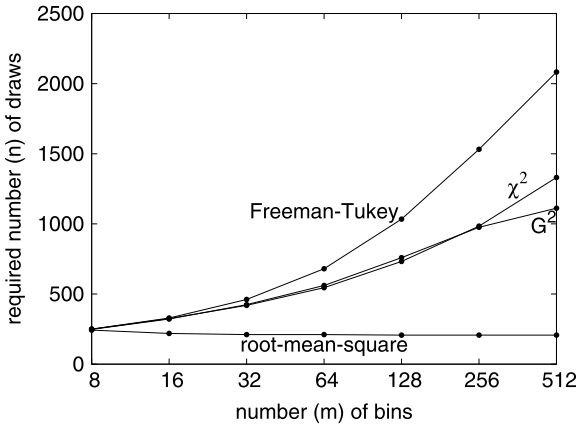

Fig. 25. Seventh example; see Section 5.2.7.

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

$$p^{(j)} = \frac{C_2}{j^2} \tag{82}$$

for $j = 1, 2, \ldots, m$, where

$$C_2 = \frac{1}{\sum_{j=1}^{m} 1/j^2}. \tag{83}$$

Fig. 24 plots the number n of draws required to distinguish the actual distribution defined in (82) and (83) from the model distribution defined in (80) and (81), estimating the parameter θ in (80) via maximum-likelihood methods (that is, by sorting). Remark 5.1 above specifies what we mean by "distinguish".

5.2.7. A model with two parameters

For the final example, let us specify the model distribution to be

$$p_0^{(1)}(\theta_1, \theta_2) = \theta_1, \tag{84}$$

$$p_0^{(2)}(\theta_1, \theta_2) = \theta_1, \tag{85}$$

$$p_0^{(3)}(\theta_1, \theta_2) = \theta_2, \tag{86}$$

$$p_0^{(4)}(\theta_1, \theta_2) = \theta_2, \tag{87}$$

and

$$p_0^{(j)}(\theta_1, \theta_2) = \frac{1 - 2\theta_1 - 2\theta_2}{m - 4} \tag{88}$$

for $j = 5, 6, \ldots, m$; we estimate the parameters θ_1 and θ_2 via maximum-likelihood methods. We consider n i.i.d. draws from the distribution

$$p^{(1)} = \frac{9}{32}, \tag{89}$$

$$p^{(2)} = \frac{3}{32}, \tag{90}$$

$$p^{(3)} = \frac{3}{32}, \tag{91}$$

$$p^{(4)} = \frac{1}{32}, \tag{92}$$

and

$$p^{(j)} = \frac{1}{2m - 8} \tag{93}$$

for $j = 5, 6, \ldots, m$.

Fig. 25 plots the number n of draws required to distinguish the actual distribution defined in (89)–(93) from the model distribution defined in (84)–(88), estimating the parameters θ_1 and θ_2 in (84)–(88) via maximum-likelihood methods. Remark 5.1 above specifies what we mean by "distinguish".

5.3. Summary

In many of the above examples, the root-mean-square outperforms the classic χ^2 statistic. Indeed, the root-mean-square often outperforms not only χ^2, but also G^2 and the Freeman–Tukey/Hellinger distance. The root-mean-square is uniformly the best in all but 4 of the 14 examples from the present section (these 4 are Sections 5.1.3, 5.1.5, 5.1.7, and 5.2.3); moreover, the root-mean-square is the best for some choices of m and t even in those 4 examples. Of course, such superb performance of the root-mean-square may not be representative of all possible applications; nevertheless, the root-mean-square often deserves consideration as a complement to the classical statistics.

Acknowledgments

We would like to thank Alex Barnett, Andrew Barron, Gérard Ben Arous, James Berger, Tony Cai, Raymond Carroll, Sourav Chatterjee, Ronald Raphy Coifman, Ingrid Daubechies, Jianqing Fan, Matan Gavish, Andrew Gelman, Leslie Greengard, Michael Hoffman, Peter W. Jones, Deborah Mayo, Peter McCullagh, Michael O'Neil, Abhinav Nellore, Ron Peled, William Press, Vladimir Rokhlin, Joseph Romano, Gary Simon, Amit Singer, Michael Stein, Stephen Stigler, John Strain, Joel Tropp, Larry Wasserman, Douglas A. Wolfe, and Bin Yu [14,16,29]. We would also like to thank Jiayang Gao for her many observations, which include pointing out the identity in (4), showing that the Freeman–Tukey/Hellinger-distance statistic is just a weighted version of the root-mean-square.

William Perkins was supported in part by NSF Grant OISE-0730136 and an NSF Postdoctoral Research Fellowship. Mark Tygert was supported in part by a Research Fellowship from the Alfred P. Sloan Foundation and a DARPA Young Faculty Award. Rachel Ward was supported in part by an NSF Postdoctoral Research Fellowship, a Donald D. Harrington Faculty Fellowship, a Research Fellowship from the Alfred P. Sloan Foundation, ONR Grant N00014-12-1-0743, and an NSF CAREER Award.

References

[1] M.J. Bayarri, J.O. Berger, P-values for composite null models, J. Amer. Statist. Assoc. 95 (452) (2000) 1127–1142.
[2] P.J. Bickel, Y. Ritov, T.M. Stoker, Tailor-made tests for goodness of fit to semiparametric hypotheses, Ann. Statist. 34 (2) (2006) 721–741.
[3] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703.
[4] W.G. Cochran, The χ^2 test of goodness of fit, Ann. Math. Statist. 23 (3) (1952) 315–345.
[5] D.R. Cox, Principles of Statistical Inference, Cambridge University Press, Cambridge, UK, 2006.
[6] R.B. D'Agostino, M.A. Stephens, Goodness-of-Fit Techniques, Marcel Dekker, New York, 1986.
[7] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC Press, Boca Raton, Florida, 1993.
[8] R.C. Eldridge, Six Thousand Common English Words: Their Comparative Frequency and What Can Be Done with Them, Nabu Press, Charleston, SC, 2010, reprint of the 1911 edition, available online at http://www.archive.org/details/sixthousandcomm00eldrgoog.
[9] E. Erosheva, E.C. Walton, D.T. Takeuchi, Self-rated health among foreign- and US-born Asian Americans: A test of comparability, Med. Care 45 (1) (2007) 80–87.
[10] R.A. Fisher, A.S. Corbet, C.B. Williams, The relation between the number of species and the number of individuals in a random sample of an animal population, J. Animal Ecology 12 (1) (1943) 42–58.
[11] A. Gelman, A Bayesian formulation of exploratory data analysis and goodness-of-fit testing, Internat. Stat. Rev. 71 (2) (2003) 369–382.
[12] S.W. Guo, E.A. Thompson, Performing the exact test of Hardy–Weinberg proportion for multiple alleles, Biometrics 48 (2) (1992) 361–372.
[13] N. Henze, Empirical-distribution-function goodness-of-fit tests for discrete models, Canad. J. Statist. 24 (1) (1996) 81–93.
[14] M. Hollander, D.A. Wolfe, Nonparametric Statistical Methods, 2nd edition, Wiley, 1999.
[15] G. Marsaglia, Random number generators, J. Modern Appl. Stat. Meth. 2 (1) (2003) 2–13.
[16] D.G. Mayo, Error and the Growth of Experimental Knowledge, University of Chicago Press, Chicago, 1996.
[17] 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. Mag. (Ser. 5) 50 (1900) 157–175.
[18] W. Perkins, M. Tygert, R. Ward, χ^2 and classical exact tests often wildly misreport significance, Technical report, arXiv:1108.4126, 2011.
[19] W. Perkins, M. Tygert, R. Ward, Computing the confidence levels for a root-mean-square test of goodness-of-fit, Appl. Math. Comput. 217 (22) (2011) 9072–9084.
[20] W. Perkins, M. Tygert, R. Ward, Computing the confidence levels for a root-mean-square test of goodness-of-fit, II, Tech. Rep. 2011, arXiv:1009.2260.
[21] W. Perkins, M. Tygert, R. Ward, Significance testing without truth, Tech. Rep. 2013, arXiv:1301.1208.
[22] W.H. Press, "What is better than chi-square?" and related koans, available at http://www.nr.com/whp/notes/betterthanchi.pdf, 2005.
[23] 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.
[24] J.M. Robins, A. van der Vaart, V. Ventura, Asymptotic distribution of P-values in composite null models, J. Amer. Statist. Assoc. 95 (452) (2000) 1143–1156.
[25] J.P. Romano, A bootstrap revival of some nonparametric distance tests, J. Amer. Statist. Assoc. 83 (403) (1988) 698–708.
[26] E. Rutherford, H. Geiger, H. Bateman, The probability variations in the distribution of α-particles, Philos. Mag. (Ser. 6) 20 (118) (1910) 698–707.
[27] Student, On the error of counting with a haemacytometer, Biometrika 5 (3) (1907) 351–360.
[28] S.R.S. Varadhan, M. Levandowsky, N. Rubin, Mathematical Statistics, Lect. Notes Ser., Courant Institute of Mathematical Sciences, NYU, New York, 1974.
[29] L. Wasserman, All of Statistics, Springer, 2003.
[30] G.K. Zipf, The Psycho-Biology of Language: An Introduction to Dynamic Philology, Houghton Mifflin, Boston, Massachusetts, 1935.