Computer-enabled Metrics of Statistical Significance for Discrete Data

Will Perkins, Mark Tygert, and Rachel Ward
with help from Raymond Carroll and Gary Simon

May 28, 2014

Brief Table of Contents

Table of Contents

Preface

The main thesis of this book is that using only classic χ^2 statistics is no longer appropriate, that the simple root-mean-square and cumulative statistics of Kolmogorov and Smirnov and others are often superior now that computers are widely available.

This book concerns the significance testing covered in introductory courses on statistics — including those at the high-school (advanced-placement) and undergraduate levels — and should be accessible to anyone familiar with elementary (discrete) probability theory. We have tried to make the individual chapters largely self-contained, so that the reader may focus on the topics of greatest interest for a given application. An expert can start directly with the relevant chapter. For readers without any particular application in mind, we recommend starting with either the appendix, which discusses nuisance parameters in general, or Chapter 1, which discusses goodness-of-fit testing (with or without nuisance parameters). Chapter 2 provides rules-of-thumb for choosing measures of discrepancy, based on whether the data is nominal or ordinal. Chapter 3 presents an application to elementary population genetics. Chapter 4 discusses tests for homogeneity in contingency-tables/cross-tabulations. Chapter 5 discusses goodness-of-fit testing for generalized linear models (including logistic and Poisson regressions). Chapter 6 discusses asymptotic distributions of statistics in the limit of large numbers of draws. Chapter 7 extends the methods of Chapter 6 to models with nuisance parameters. Chapter 8 discusses asymptotic power in the limit of large numbers of draws. The appendix discusses nuisance parameters in detail.

A premier application of significance testing is checking for proper randomization in statistical studies. For instance, while perfect randomization in medical trials or econometric investigations is not always feasible, best practices nevertheless require that the ethnicities in the case/treated and control groups be consistent with having arisen as independent and identically distributed (i.i.d.) draws from the pooled case-control sample (that is, that pooling the estimates of ethnic prevalences is permissible). Chapter 4 discusses such tests; Chapters 1 and 3 discuss simpler variants. For example, a study that detects a difference in outcomes between a largely Irish case/treated group and a largely Italian control group may not be detecting a difference between the treatment case and the placebo control. The classical χ^2, likelihood-ratio, G^2, and other tests from the Cressie-Read power-divergence family can be blind to such a discrepancy in ethnicities, unlike the root-mean-square. We recommend using both a classical statistic (such as χ^2) and either the root-mean-square or one of the cumulative variants discussed in Chapter 2.

Will Perkins
Mark Tygert
Rachel Ward

Chapter 1

Goodness of fit for distributional profile

Goodness-of-fit tests based on the root-mean-square 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 chapter discusses numerous examples of this. Goodness-of-fit tests based on the root-mean-square are now practical and convenient: although the actual values taken by the root-mean-square and similar goodness-of-fit statistics are seldom interpretable without the aid of a computer, black-box software can rapidly calculate their precise significance.

1.1 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 chapter 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, Rao, 2002; Varadhan et al., 1974, page 123; or Section 1.2 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," we denote by x the value of the root-mean-square for the given i.i.d. draws; we denote by X the root-mean-square 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 maximum-likelihood estimate for the experimental data; see the appendix for discussion of how this differs from other common choices). The "P-value" P is then defined to be the probability that $X \geq x$ (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 χ^2 statistic of Pearson (1900) — see, for example, formula (1.2) below. However, when some model probabilities can be small (relative to others in the same 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 1.1.1). Press (2005) tackled similar issues.

The main thesis of the present chapter is that using only the classic χ^2 statistic is no longer appropriate, that certain alternatives are 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 this is always the best alternative. In fact, the discrete Kolmogorov-Smirnov and related statistics used by Clauset et al. (2009) and D'Agostino and Stephens (1986) can be more powerful than the root-mean-square when there is a natural data-independent ordering (or partial order) for the bins — see Chapter 2; 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 straightforward, even when estimating continuous parameters via maximum-likelihood methods, as in Chapters 6 and 7. Moreover, the classic χ^2 statistic is just a weighted version of the root-mean-square, facilitating their comparison. Classic statistics such as χ^2 focus on relative (rather than absolute) discrepancies. Finally, χ^2 and the root-mean-square coincide when the model distribution is uniform.

Please note that all statistical tests reported in this monograph (including those involving the χ^2 statistic) are exact; we compute P-values via Monte-Carlo simulations providing guaranteed error bounds (see Section 1.3 below). For all numerical results, we generated random numbers via the C programming language procedure given on page 9 of Marsaglia (2003), implementing the recommended complementary multiply with carry.

The problem with χ^2 is neither subtle nor esoteric. The present chapter surveys many examples; Chapter 3 below details a specific example from elementary population genetics.

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 1.1.1. In many of our examples, there is a bin for which the expected number of draws is very small under the model. 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, Subsection 1.5.1.1 or Subsection 1.5.2.4.

Remark 1.1.2. 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 the 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 (1900) introducing χ^2; see also the modern treatments by Gelman et al. (1995) and Cox (2006).

Remark 1.1.3. The assumption that the given observations are i.i.d. draws is not necessary. Indeed, all other chapters in this book do not make this assumption. This first chapter focuses on the i.i.d. case for simplicity. See the appendix for a more general treatment.

1.2 Definitions of the divergences

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 Read and Cressie (1988). We use $p_0^{(1)}$, $p_0^{(2)}$, …, $p_0^{(m)}$ to denote the modeled fractions of n i.i.d. draws falling in m bins, numbered 1, 2, …, m, respectively, and we use $\hat{p}^{(1)}$, $\hat{p}^{(2)}$, …, $\hat{p}^{(m)}$ to denote the observed fractions of the n draws falling in the respective bins. That is, $p_0^{(1)}$, $p_0^{(2)}$, …, $p_0^{(m)}$ are the probabilities associated with the respective bins in the *model* distribution, whereas $\hat{p}^{(1)}$, $\hat{p}^{(2)}$, …, $\hat{p}^{(m)}$ 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, …, 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, …, i_n falling in bin j, for $j = 1, 2, \ldots, m$. If the model is parameterized by a parameter θ, then the probabilities $p_0^{(1)}$, $p_0^{(2)}$, …, $p_0^{(m)}$ are functions of θ; if the model is fully specified, then we can view the probabilities $p_0^{(1)}$, $p_0^{(2)}$, …, $p_0^{(m)}$ as constant as functions of θ. We use $\hat{\theta}$ to denote a maximum-likelihood estimate of θ obtained from $\hat{p}^{(1)}$, $\hat{p}^{(2)}$, …, $\hat{p}^{(m)}$.

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

$$x = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{p}^{(j)} - p_0^{(j)}(\hat{\theta}))^2}. \tag{1.1}$$

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

The classical Pearson χ^2 statistic is

$$\chi^2 = n \sum_{j=1}^{m} \frac{(\hat{p}^{(j)} - p_0^{(j)}(\hat{\theta}))^2}{p_0^{(j)}(\hat{\theta})}, \tag{1.2}$$

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

$$g^2 = 2n \sum_{j=1}^{m} \hat{p}^{(j)} \ln\left(\frac{\hat{p}^{(j)}}{p_0^{(j)}(\hat{\theta})}\right), \tag{1.3}$$

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

$$h^2 = 4n \sum_{j=1}^{m} \left(\sqrt{\hat{p}^{(j)}} - \sqrt{p_0^{(j)}(\hat{\theta})}\right)^2 = 4n \sum_{j=1}^{m} \left[(\hat{p}^{(j)} - p_0^{(j)}(\hat{\theta}))^2 \Big/ \left(\sqrt{\hat{p}^{(j)}} + \sqrt{p_0^{(j)}(\hat{\theta})}\right)^2 \right]. \tag{1.4}$$

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 (1.2), g^2 defined in (1.3), and h^2 defined in (1.4) are all the same when the actual underlying distribution of the draws comes from the model, as discussed, for example, by Rao (2002). 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.

1.3 Computation of P-values

We need to compute the P-value assessing the consistency of the experimental data with assuming

$$H_0 : \text{the data arises as i.i.d. draws from } p_0(\hat{\theta}), \text{ for the particular observed value of } \hat{\theta}, \tag{1.5}$$

where $\hat{\theta}$ is a maximum-likelihood estimate of θ (if the model is fully specified, then the probability distribution $p_0(\theta)$ is the same for all θ); see the appendix for discussion of this hypothesis and several alternatives (including "P-values" that differ from those used below).

Remark 1.3.1. 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 Subsection 1.4.2 for details. With the Monte-Carlo scheme described below, we need not contemplate how many degrees of freedom are in a permutation.

For the computation of P-values, we can use Monte-Carlo simulations (very similar to those used by Clauset et al. (2009)). 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 (1.5) 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 1.3.2 below. This procedure works since, by definition, the P-value is the probability that

$$d\left[\begin{pmatrix} \hat{P}^{(1)} \\ \hat{P}^{(2)} \\ \vdots \\ \hat{P}^{(m)} \end{pmatrix}, \begin{pmatrix} p_0^{(1)}(\hat{\Theta}) \\ p_0^{(2)}(\hat{\Theta}) \\ \vdots \\ p_0^{(m)}(\hat{\Theta}) \end{pmatrix}\right] \quad \geq \quad d\left[\begin{pmatrix} \hat{p}^{(1)} \\ \hat{p}^{(2)} \\ \vdots \\ \hat{p}^{(m)} \end{pmatrix}, \begin{pmatrix} p_0^{(1)}(\hat{\theta}) \\ p_0^{(2)}(\hat{\theta}) \\ \vdots \\ p_0^{(m)}(\hat{\theta}) \end{pmatrix}\right], \tag{1.6}$$

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),

- $\hat{p}^{(1)}$, $\hat{p}^{(2)}$, …, $\hat{p}^{(m)}$ are the fractions of the n given experimental draws falling in the respective bins,

- $\hat{\theta}$ is the estimate of θ obtained from $\hat{p}^{(1)}$, $\hat{p}^{(2)}$, …, $\hat{p}^{(m)}$,

- $\hat{P}^{(1)}$, $\hat{P}^{(2)}$, …, $\hat{P}^{(m)}$ 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 (1.5), and

- $\hat{\Theta}$ is the estimate of the parameter θ obtained from $\hat{P}^{(1)}$, $\hat{P}^{(2)}$, ..., $\hat{P}^{(m)}$ (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; Remark A.4.2 discusses a consequence).

When taking the probability that (1.6) occurs, only the left-hand side is random — we regard the left-hand side of (1.6) 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 (1.6) occurs, we can calculate many independent realizations of the random variable and observe that the fraction which satisfy (1.6) is a good approximation to the probability when the number of realizations is large; Remark 1.3.2 details the accuracy of the approximation. (The procedure in the present section follows this prescription to estimate P-values.)

Remark 1.3.2. The standard error of the estimate from the present section for an exact P-value P is $\sqrt{P(1-P)/\ell}$, where ℓ is the number of Monte-Carlo 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 greater than or equal to that corresponding to P-value P is therefore $\sqrt{\ell P(1-P)}$, and so the standard deviation of the *fraction* of the simulations 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 $\sqrt{P(1-P)/\ell}$).

1.4 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 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 Section 1.3; Remark 1.3.2 details the guaranteed accuracy of the computed P-values.

1.4.1 Synthetic examples

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

$$p_0^{(1)} = \frac{1}{4}, \tag{1.7}$$

$$p_0^{(2)} = \frac{1}{4}, \tag{1.8}$$

and

$$p_0^{(j)} = \frac{1}{2m-4} \tag{1.9}$$

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 (1.7)–(1.9), but we would like to see exactly how well the various goodness-of-fit statistics detect the discrepancy.

Figure 1.1 plots the P-values for testing whether the empirical data arises from the model specified in (1.7)–(1.9). 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 (1.9) from 0 is usually less than the discrepancy of (1.9) from a typical realization drawn from the model, since under the model the sum of the expected numbers of draws in bins 3, 4, ..., m is $n/2$ for any m.

Figure 1.1 demonstrates that the root-mean-square rejects the invalid model with nearly 100% confidence while the classical statistics report nearly 0% confidence for rejection. Also, 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 depends dramatically on the number $(m - 2)$ of unlikely bins in the model, even though the data are the same for all m. This suggests a 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 (1.7)–(1.9). 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, ..., 50, and no draw in any other bin. As before, this data is unlikely to arise from the model specified in (1.7)–(1.9), but we would like to see exactly how well the various goodness-of-fit statistics detect the discrepancy.

Figure 1.2 plots the P-values for testing whether the empirical data arises from the model specified in (1.7)–(1.9). 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 rejects the hypothesis that the data arises from the model, whereas the classical statistics find little evidence for rejecting the manifestly invalid model.

1.4.2 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 (1935)); the chief source for the English language is Eldridge (2010). We revisit the data of Eldridge (2010) in the present subsection to assess the performance of the goodness-of-fit statistics.

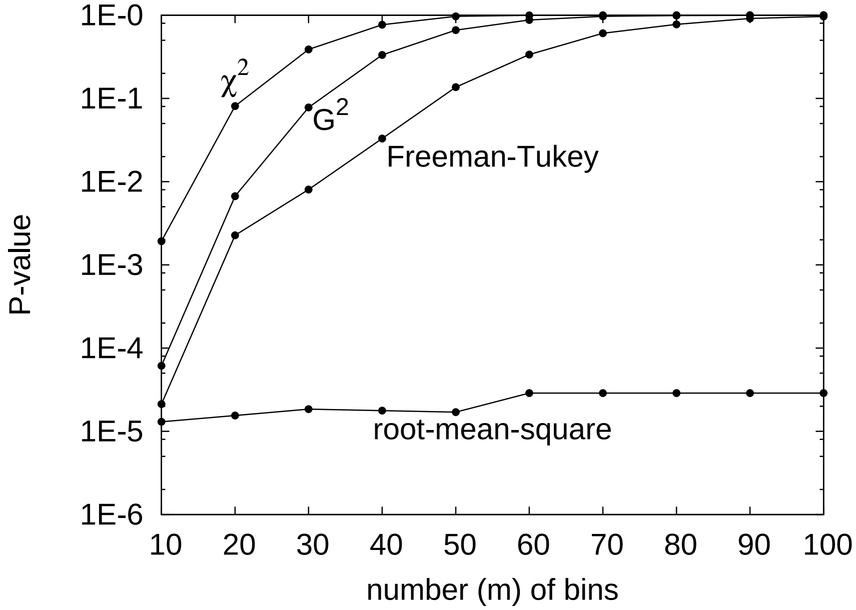


Figure 1.1: P-values for the hypothesis that the model (1.7)–(1.9) agrees with the data of 15 draws in the first bin, 5 draws in the second bin, and no draw in any other bin

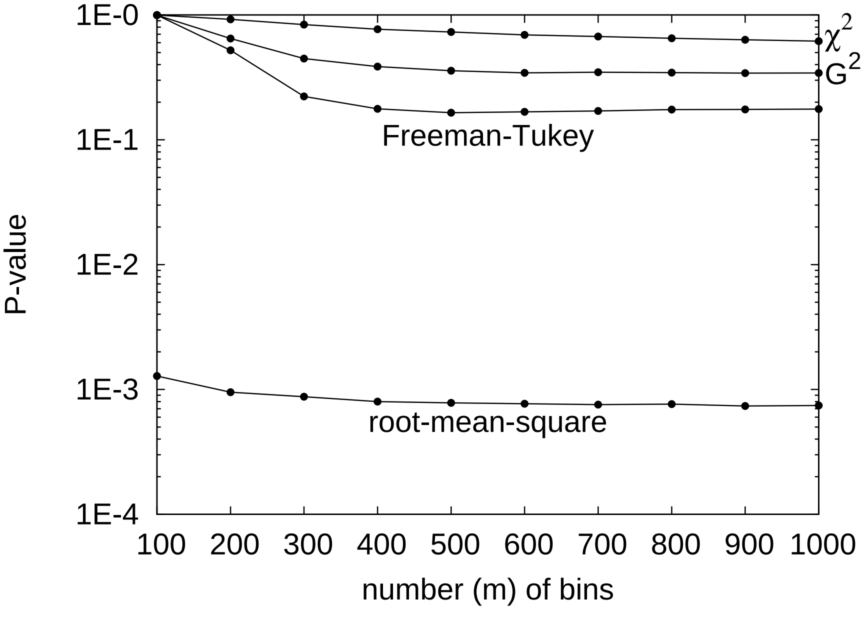


Figure 1.2: P-values for the hypothesis that the model (1.7)–(1.9) agrees with the data of 36 draws in the first bin, 12 draws in the second bin, 1 draw each in bins 3, 4, . . . , 50, and no draw in any other bin

We first analyze List 1 of Eldridge (2010), 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 $n = 10{,}000$ of the 13,825 words to obtain a corpus of $n = 10{,}000$ draws over 2,890 bins. Figure 1.3 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

$$p_0^{(j)}(\theta) = \frac{C_1}{\theta(j)} \tag{1.10}$$

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

$$C_1 = \frac{1}{\sum_{j=1}^{m} 1/j}; \tag{1.11}$$

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 $\theta(j_1) = 1$, $\theta(j_2) = 2$, $\dots$, $\theta(j_m) = m$. 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 Figure 1.4 we plot P-values (each computed via 40,000 Monte-Carlo simulations) for varying values of m; although List 1 of Eldridge (2010) 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 (2010). Note that Figure 1.4 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 (2010) was drawn. It is not clear a priori which dictionary is appropriate. The P-values for 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 (1.10) and (1.11)) for any m. In contrast, the P-values for the classical statistics vary significantly 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 give no information.

Furthermore, analyzing List 5 of Eldridge (2010) 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 (2010). We randomly choose $n = 20{,}000$ of the 43,989 words to obtain a corpus of $n = 20{,}000$ draws over 6,002 bins. Figure 1.5 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 Figure 1.6 we plot P-values (each computed via 40,000 Monte-Carlo simulations) for varying values of m; although List 5 of Eldridge (2010) 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 (2010). Please note that Figure 1.6 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 Figures 1.4 and 1.6 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 give no information.

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 (2010), with the number m of bins set to 7,500. Indeed, let us consider the model

$$p_0^{(j)}(\theta_0,\theta_1,\theta_2,\theta_3,\theta_4) = \begin{cases} \theta_1, & \theta_0(j) = 1 \\ \theta_2, & \theta_0(j) = 2 \\ \theta_3, & \theta_0(j) = 3 \\ C/(\theta_0(j))^{\theta_4}, & \theta_0(j) = 4, 5, \dots, 7500 \end{cases}, \tag{1.12}$$

where

$$C = C_{\theta_1,\theta_2,\theta_3,\theta_4} = \frac{1-\theta_1-\theta_2-\theta_3}{\sum_{j=4}^{7500} 1/j^{\theta_4}}, \tag{1.13}$$

with θ_0 being a permutation of the integers $1, 2, \dots, 7500$, and θ_1, θ_2, θ_3, θ_4 being nonnegative real numbers; we estimate θ_0, θ_1, θ_2, θ_3, θ_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 (1.12) and (1.13), the P-values calculated via 4,000,000 Monte-Carlo simulations are

- χ^2: .510
- G^2 (the log–likelihood-ratio): .998
- Freeman-Tukey (the Hellinger distance): 1.000
- root-mean-square: .587

Thus, all four statistics indicate that the truncated power-law model defined in (1.12) and (1.13) is a good fit. This is in accord with Figure 1.5, in which all but the three greatest frequencies appear to follow a truncated power-law.

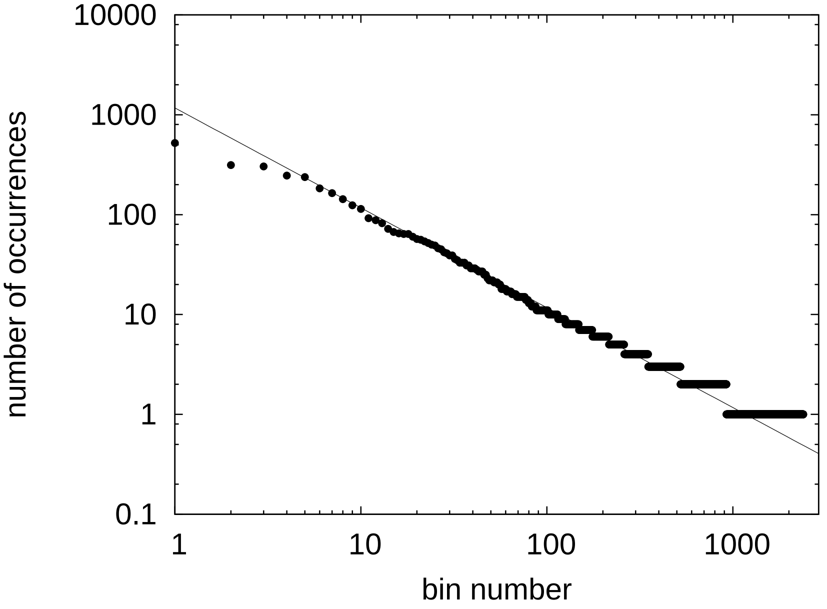


Figure 1.3: Numbers of occurrences of the various words (one bin for each distinct word) in a corpus of 10,000 random draws from List 1 of Eldridge (2010)

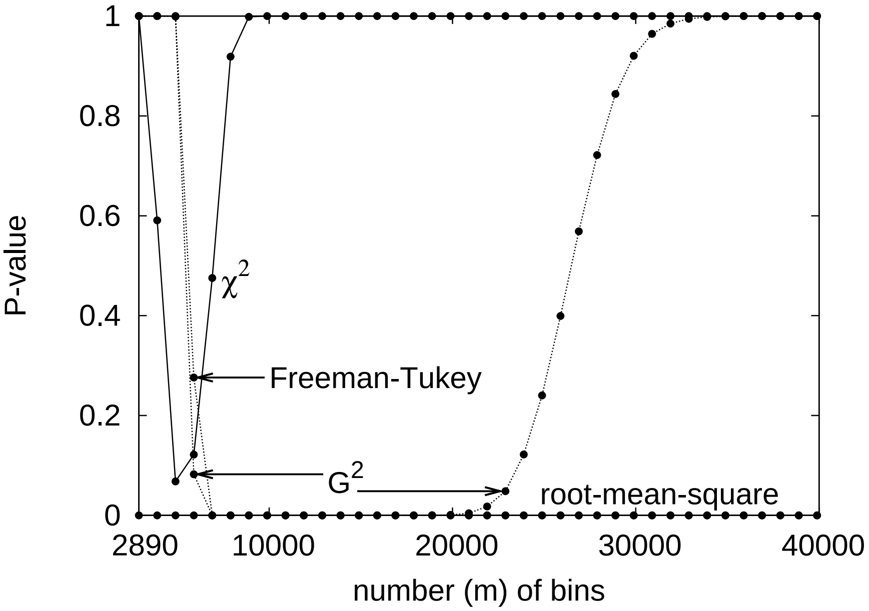


Figure 1.4: P-values for the data plotted in Figure 1.3 to follow the Zipf distribution

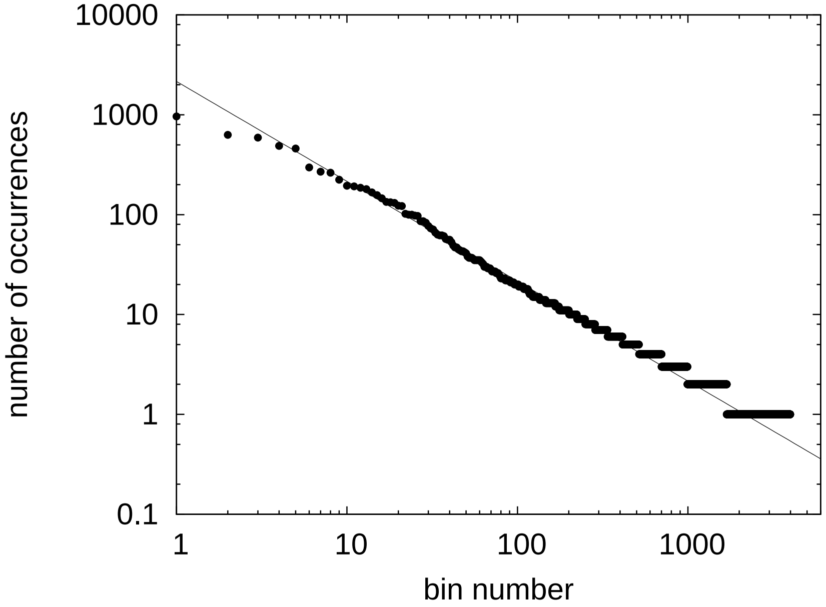


Figure 1.5: Numbers of occurrences of the various words (one bin for each distinct word) in a corpus of 20,000 random draws from List 5 of Eldridge (2010)

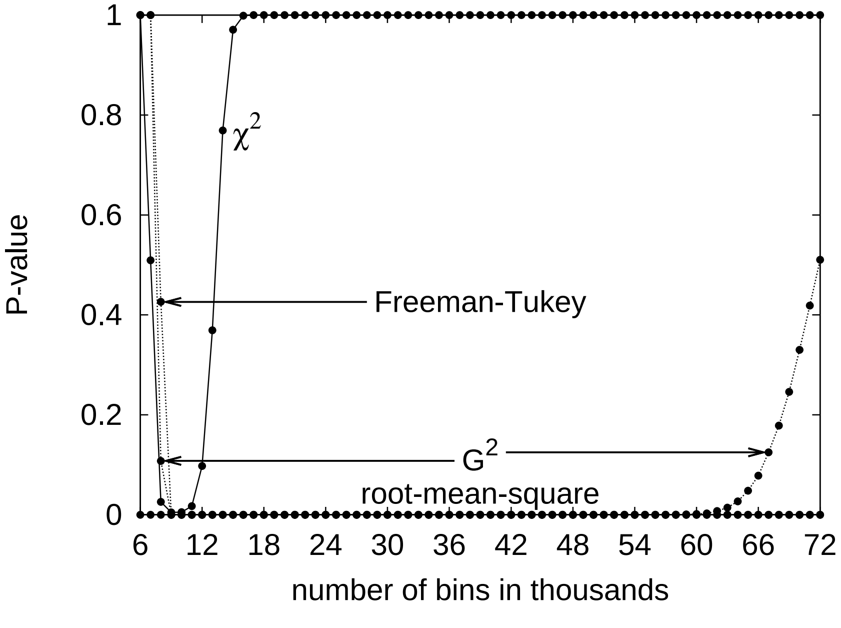


Figure 1.6: P-values for the data plotted in Figure 1.5 to follow the Zipf distribution

Table 1.1: 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

1.4.3 A Poisson law for radioactive decays

Table 1.1 summarizes the classic example of a Poisson-distributed experiment in radioactive decay of Rutherford et al. (1910); Figure 1.7 plots the data, along with the Poisson distribution whose mean is the same as the data's. Figure 1.8 reports the P-values for testing whether the data, while retaining only bins 1, 2, …, 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, …, the truncation amounts to ignoring draws in bins $m+1$, $m+2$, $m+3$, … when $m \geq 12$, 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.

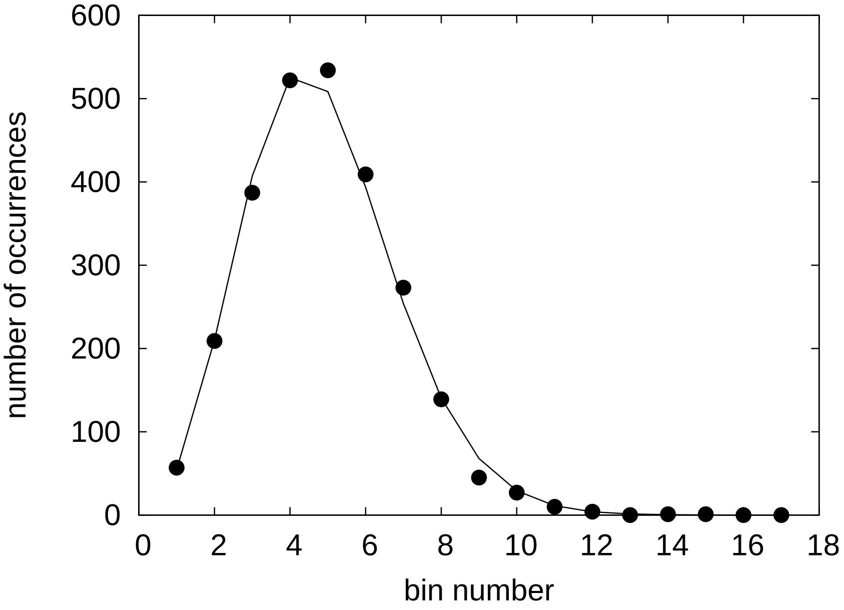


Figure 1.7: The data in Table 1.1 (the dots) and the best-fit Poisson distribution (the lines)

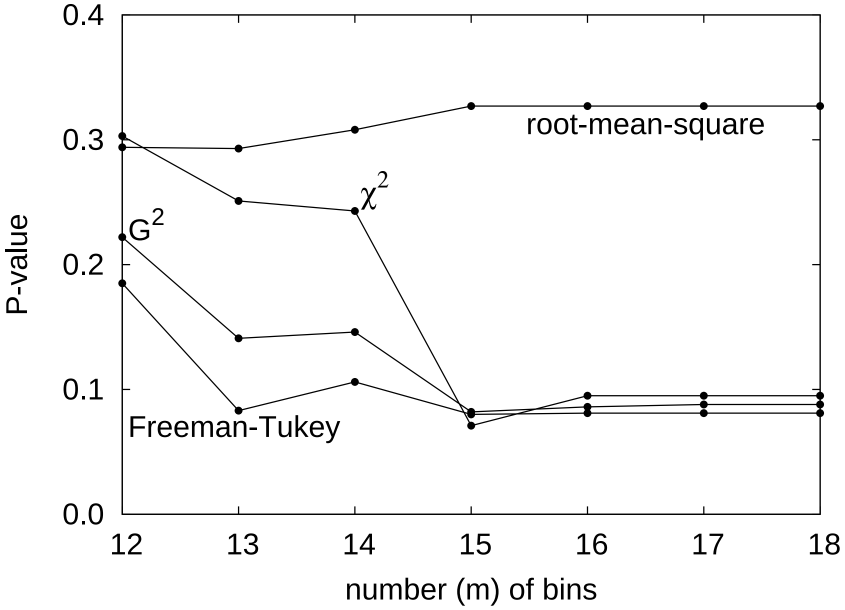


Figure 1.8: P-values for the distribution of Table 1.1 to be Poisson

Table 1.2: 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

1.4.4 A Poisson law for counting with a hæmacytometer

Page 357 of Student (1907) reports on the number of yeast cells observed in each of 400 squares in a hæmacytometer microscope slide. Table 1.2 displays the counts; Figure 1.9 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: .627
- G^2 (the log–likelihood-ratio): .365
- Freeman-Tukey (the Hellinger distance): .111
- root-mean-square: .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.

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

Using propensity scores, Erosheva et al. (2007) matched each of 335 surveyed foreign-born Asian Americans to a similar surveyed US-born Asian American. Table 1.3 duplicates Table 4 of Erosheva et al. (2007), tabulating the numbers of matched pairs reporting various

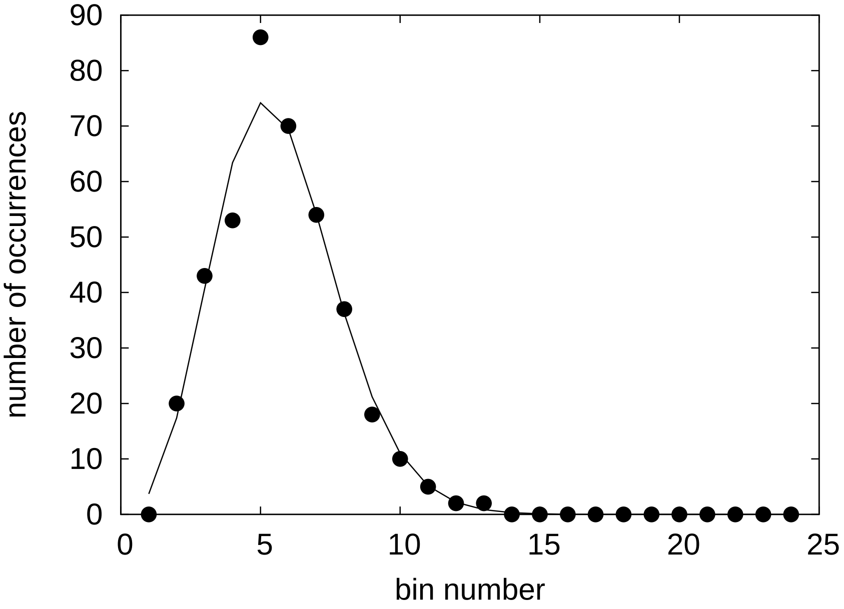


Figure 1.9: The data in Table 1.2 (the dots) and the best-fit Poisson distribution (the lines)

combinations of physical health; the propensity scores were generated without reference to the health ratings. Table 1.3 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 1.3 are

- χ^2: .784
- G^2 (the log–likelihood-ratio): .739
- Freeman-Tukey (the Hellinger distance): .642
- root-mean-square: .973

After noting that χ^2 does not reveal any statistically significant asymmetry in Table 1.3, Erosheva et al. (2007) 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 1.3. Table 1.4 provides one possible modification. The P-values calculated via 4,000,000 Monte-Carlo simulations for testing the symmetry of Table 1.4 are

- χ^2: .109
- G^2 (the log–likelihood-ratio): .123
- Freeman-Tukey (the Hellinger distance): .155
- root-mean-square: .014

Table 1.3: 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 1.4: A variation on Table 1.3

		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

Evidently, the root-mean-square is more powerful for detecting the asymmetry of Table 1.4.

Table 1.5 provides another hypothetical cross-tabulation. The P-values calculated via 64,000,000 Monte-Carlo simulations for testing the symmetry of Table 1.5 are

- χ^2: .0015
- G^2 (the log–likelihood-ratio): .00016
- Freeman-Tukey (the Hellinger distance): .000006, i.e., 6E–6
- root-mean-square: .131

The classical statistics are much more powerful for detecting the asymmetry of Table 1.5, contrasting how the root-mean-square is more powerful for detecting the asymmetry of Table 1.4. 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.

Table 1.5: Another variation on Table 1.3

		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

1.4.6 A modified geometric law for the species of butterflies

Fisher et al. (1943) 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. Figure 1.10 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 Figure 1.10, we must include a permutation of the bins as a parameter, since we have sorted the data (see Subsection 1.4.2 for further discussion of sorting and permutations). We take the model to be

$$p_0^{(j)}(\theta_0, \theta_1) = A_{\theta_1} \frac{(\theta_1)^{\theta_0(j)}}{\sqrt{\theta_0(j) + 23}} \tag{1.14}$$

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

$$A_{\theta_1} = \frac{1}{\sum_{j=1}^{217} (\theta_1)^j / \sqrt{j + 23}}; \tag{1.15}$$

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: .0050
- G^2 (the log–likelihood-ratio): .349
- Freeman-Tukey (the Hellinger distance): .951
- root-mean-square: .00002, i.e., 2E–5

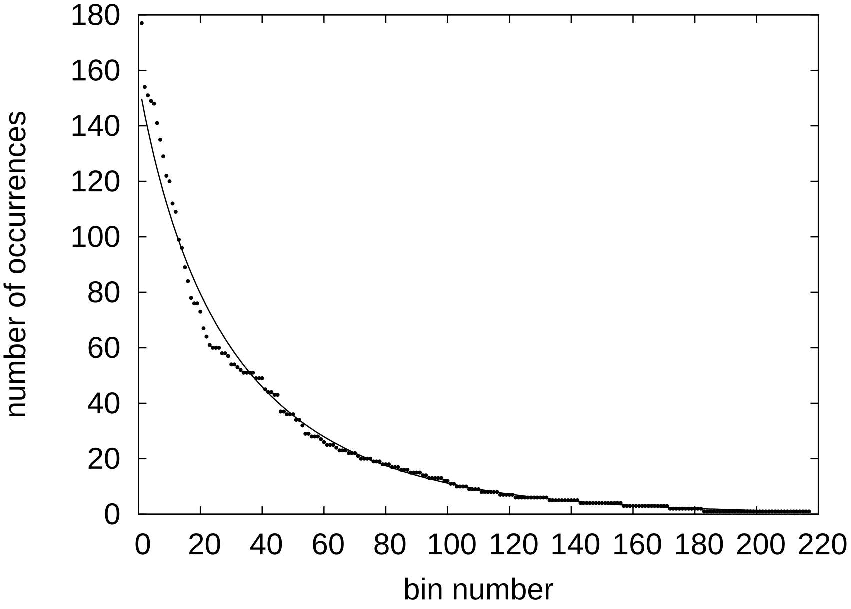


Figure 1.10: Numbers of specimens (the dots) from 217 species of butterflies (one bin per species), and the best-fit distribution (the lines)

As Figure 1.10 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.

1.5 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 success at detecting discrepancies from the models, we use the formulation of the following remark.

Remark 1.5.1. In the present section, 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. Section 1.6 uses a weaker notion of "distinguish" — in Section 1.6 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.

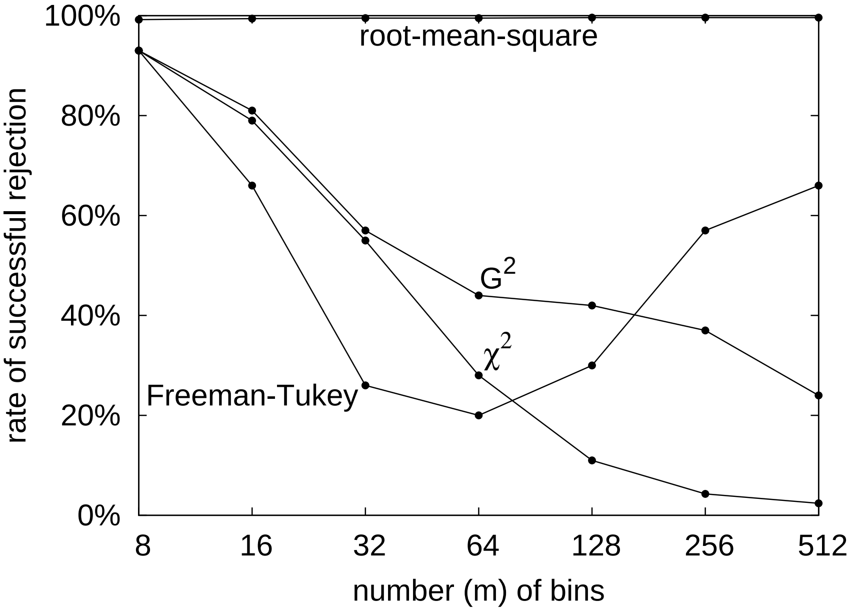


Figure 1.11: First example, with $n = 200$ draws; see Subsection 1.5.1.1.

Remark 1.5.2. To compute the P-values for each example in Subsection 1.5.2, 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 Section 1.3. The computational costs for generating the plots in Subsection 1.5.2 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 for 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 Section 1.3, using the single value of $\hat{\theta}$ (but many values of $\tilde{\theta}$ from Section 1.3). 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 Subsection 1.5.2.6, 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 Section 1.3) 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 1.5.3. 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 and an asymptotically equivalent variation of χ^2, such as the log–likelihood-ratio "G^2."

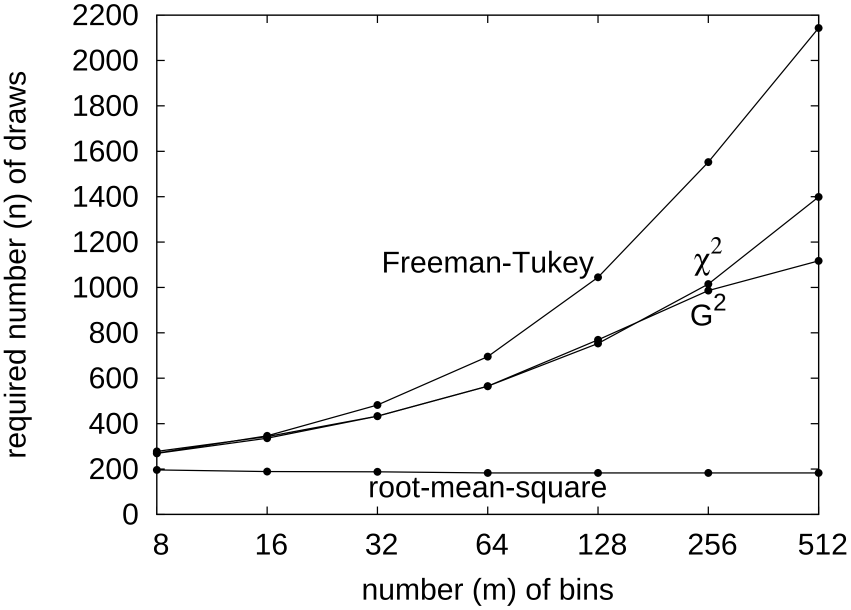


Figure 1.12: First example (statistical "efficiency"); see Subsection 1.5.1.1.

1.5.1 Examples without parameter estimation

1.5.1.1 A simple, illustrative example

Let us first specify the model distribution to be

$$p_0^{(1)} = \frac{1}{4}, \qquad p_0^{(2)} = \frac{1}{4}, \qquad p_0^{(j)} = \frac{1}{2m-4} \tag{1.16}$$

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

$$p^{(1)} = \frac{3}{8}, \qquad p^{(2)} = \frac{1}{8}, \qquad p^{(j)} = p_0^{(j)} \tag{1.17}$$

for $j = 3, 4, \dots, m$, where $p_0^{(3)}, p_0^{(4)}, \dots, p_0^{(m)}$ are the same as in (1.16).

Figure 1.11 plots the percentage of 40,000 simulations, each generating 200 i.i.d. draws according to the actual distribution defined in (1.17), 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 (1.16). Figure 1.11 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 $m \geq 256$.

Figure 1.12 plots the number n of draws required to distinguish the actual distribution defined in (1.17) from the model distribution defined in (1.16), where "distinguish" is defined in Remark 1.5.1 above. Figure 1.12 shows the root-mean-square requires 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 $m \geq 128$. The classical χ^2 statistic requires increasingly many draws as the number m of bins increases, unlike the root-mean-square.

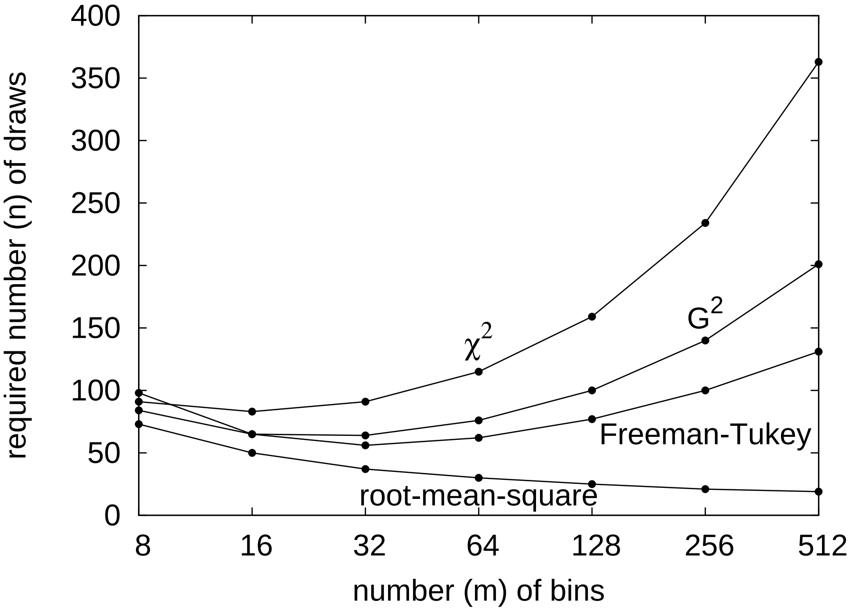


Figure 1.13: Second example; see Subsection 1.5.1.2.

1.5.1.2 Truncated power-laws

Next, let us specify the model distribution to be

$$p_0^{(j)} = \frac{C_1}{j} \tag{1.18}$$

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

$$C_1 = \frac{1}{\sum_{j=1}^m 1/j}. \tag{1.19}$$

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

$$p^{(j)} = \frac{C_2}{j^2} \tag{1.20}$$

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

$$C_2 = \frac{1}{\sum_{j=1}^m 1/j^2}. \tag{1.21}$$

Figure 1.13 plots the number n of draws required to distinguish the actual distribution defined in (1.20) and (1.21) from the model distribution defined in (1.18) and (1.19). (Remark 1.5.1 above specifies what we mean by "distinguish.") Figure 1.13 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.

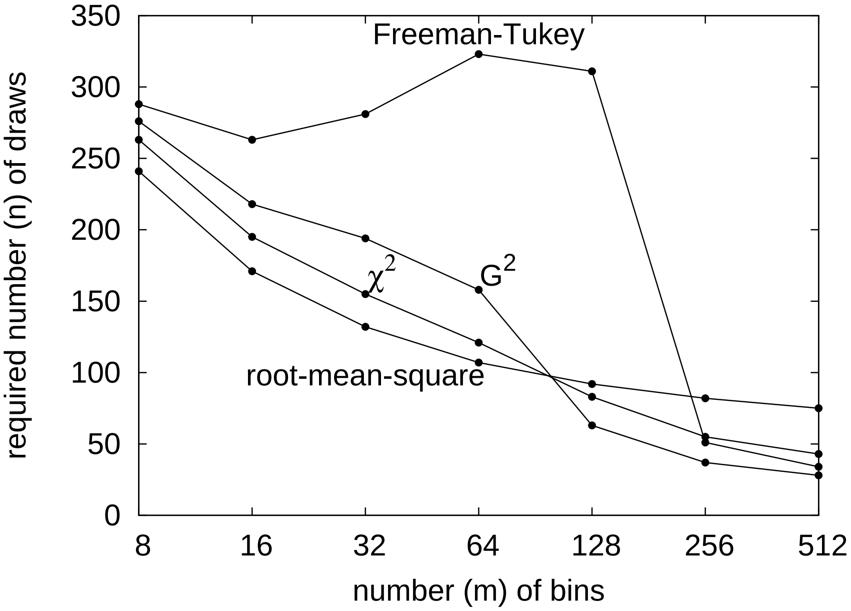


Figure 1.14: Third example; see Subsection 1.5.1.3.

1.5.1.3 Additional truncated power-laws

Let us again specify the model distribution to be

$$p_0^{(j)} = \frac{C_1}{j} \tag{1.22}$$

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

$$C_1 = \frac{1}{\sum_{j=1}^{m} 1/j}. \tag{1.23}$$

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

$$p^{(j)} = \frac{C_{1/2}}{\sqrt{j}} \tag{1.24}$$

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

$$C_{1/2} = \frac{1}{\sum_{j=1}^{m} 1/\sqrt{j}}. \tag{1.25}$$

Figure 1.14 plots the number n of draws required to distinguish the actual distribution defined in (1.24) and (1.25) from the model distribution defined in (1.22) and (1.23). (Remark 1.5.1 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 1.5.3 at the beginning of the present section.

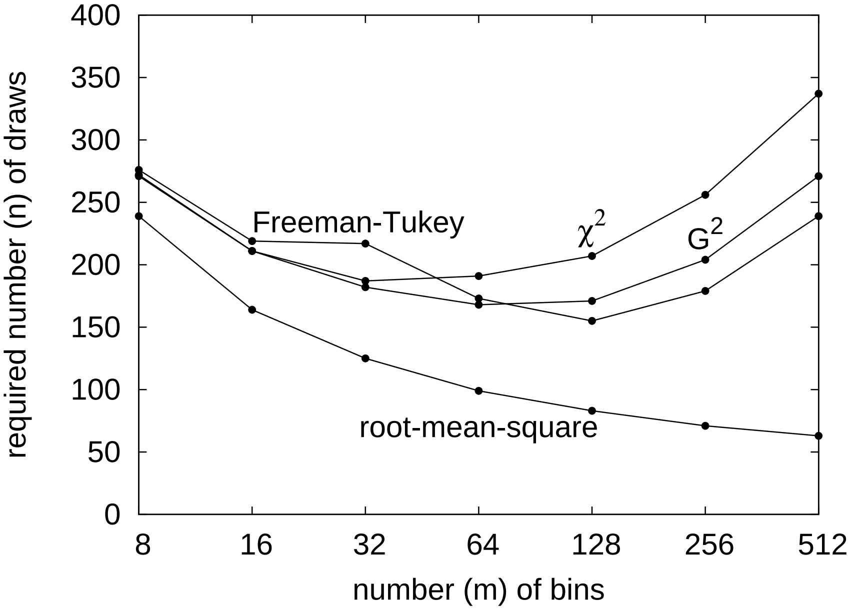


Figure 1.15: Fourth example; see Subsection 1.5.1.4.

1.5.1.4 Additional truncated power-laws, reversed

Let us next specify the model distribution to be

$$p_0^{(j)} = \frac{C_{1/2}}{\sqrt{j}} \tag{1.26}$$

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

$$C_{1/2} = \frac{1}{\sum_{j=1}^{m} 1/\sqrt{j}}. \tag{1.27}$$

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

$$p^{(j)} = \frac{C_1}{j} \tag{1.28}$$

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

$$C_1 = \frac{1}{\sum_{j=1}^{m} 1/j}. \tag{1.29}$$

Figure 1.15 plots the number n of draws required to distinguish the actual distribution defined in (1.28) and (1.29) from the model distribution defined in (1.26) and (1.27). (Remark 1.5.1 above specifies what we mean by "distinguish.") Figure 1.15 shows that the classical χ^2 statistic requires many times more draws than the root-mean-square, as the number m of bins increases.

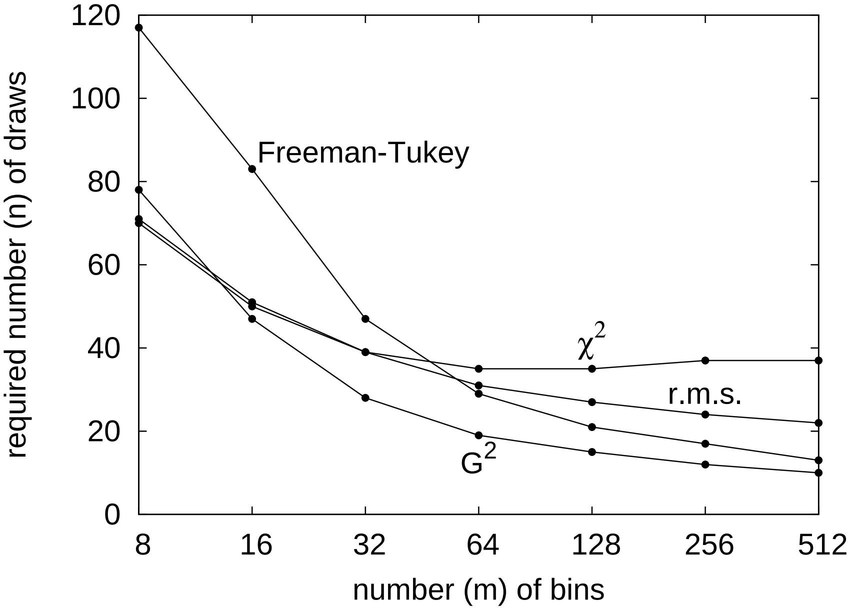


Figure 1.16: Fifth example; see Subsection 1.5.1.5.

1.5.1.5 A final example with fully specified truncated power-laws

Let us next specify the model distribution to be

$$p_0^{(j)} = \frac{C_2}{j^2} \tag{1.30}$$

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

$$C_2 = \frac{1}{\sum_{j=1}^m 1/j^2}. \tag{1.31}$$

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

$$p^{(j)} = \frac{C_1}{j} \tag{1.32}$$

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

$$C_1 = \frac{1}{\sum_{j=1}^m 1/j}. \tag{1.33}$$

Figure 1.16 plots the number n of draws required to distinguish the actual distribution defined in (1.32) and (1.33) from the model distribution defined in (1.30) and (1.31). (Remark 1.5.1 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 1.5.3 at the beginning of the present section.

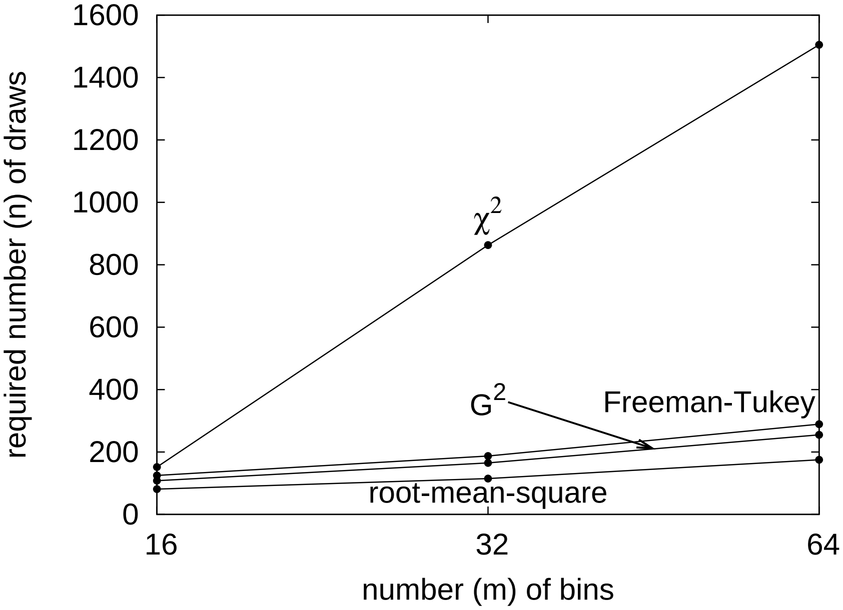


Figure 1.17: Sixth example; see Subsection 1.5.1.6.

1.5.1.6 Modified Poisson distributions

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

$$p_0^{(j)} = \frac{B_{3m/8}\,\left(\frac{3m}{8}\right)^{j-1}}{(j-1)!} \tag{1.34}$$

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

$$B_{3m/8} = \frac{1}{\sum_{j=1}^{m} \left(\frac{3m}{8}\right)^{j-1}/(j-1)!}. \tag{1.35}$$

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

$$p^{((3m/8)-1)} = S/10, \tag{1.36}$$

$$p^{(3m/8)} = 4S/5, \tag{1.37}$$

$$p^{((3m/8)+1)} = S/10, \tag{1.38}$$

$$S = p_0^{((3m/8)-1)} + p_0^{(3m/8)} + p_0^{((3m/8)+1)}, \tag{1.39}$$

$$p^{(j)} = p_0^{(j)} \tag{1.40}$$

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

Figure 1.17 plots the number n of draws required to distinguish the actual distribution defined in (1.36)–(1.40) from the model distribution defined in (1.34) and (1.35). (Remark 1.5.1 above specifies what we mean by “distinguish.”)

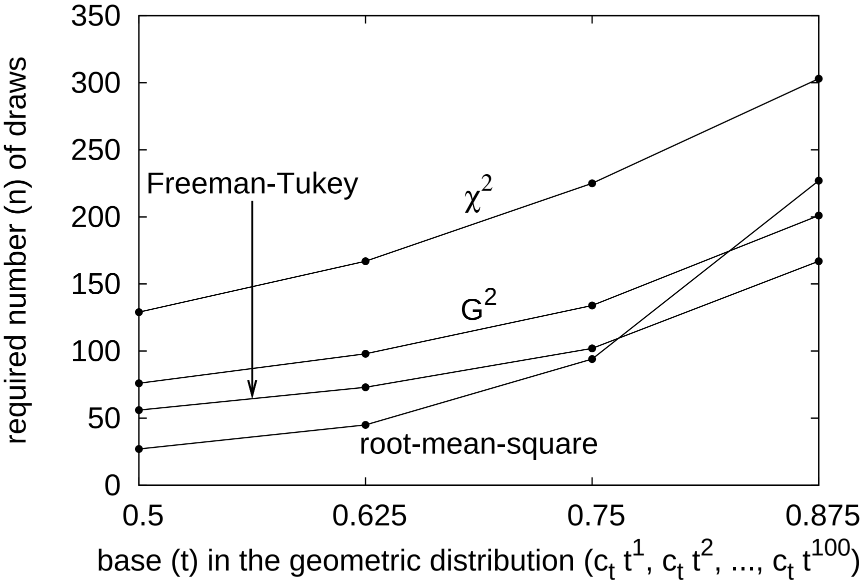


Figure 1.18: Seventh example; see Subsection 1.5.1.7.

1.5.1.7 A truncated power-law and a truncated geometric distribution

Let us finally specify the model distribution to be

$$p_0^{(j)} = \frac{C_1}{j} \tag{1.41}$$

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

$$C_1 = \frac{1}{\sum_{j=1}^{100} 1/j}. \tag{1.42}$$

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

$$p^{(j)} = c_t\, t^j \tag{1.43}$$

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

$$c_t = \frac{1}{\sum_{j=1}^{100} t^j}; \tag{1.44}$$

Figure 1.18 considers several values for t.

Figure 1.18 plots the number n of draws required to distinguish the actual distribution defined in (1.43) and (1.44) from the model distribution defined in (1.41) and (1.42). (Remark 1.5.1 above specifies what we mean by "distinguish.") See the next section, Subsection 1.5.2.1, for a similar example, this time involving parameter estimation.

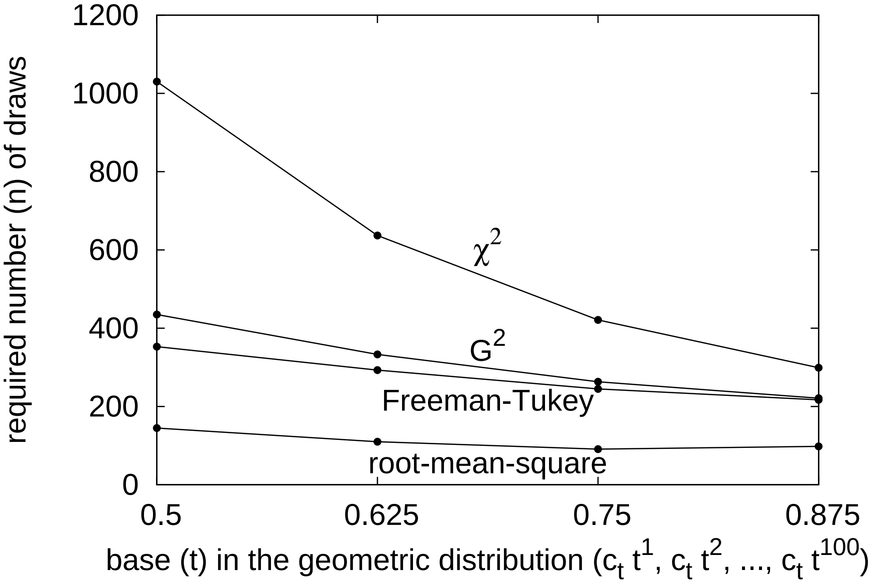


Figure 1.19: First example; see Subsection 1.5.2.1.

1.5.2 Examples with parameter estimation

1.5.2.1 A truncated power-law and a truncated geometric distribution

We turn now to models involving parameter estimation. Let us specify the model distribution to be the Zipf distribution

$$p_0^{(j)}(\theta) = \frac{C_\theta}{j^\theta} \tag{1.45}$$

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

$$C_\theta = \frac{1}{\sum_{j=1}^{100} 1/j^\theta}; \tag{1.46}$$

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

$$p^{(j)} = c_t\, t^j \tag{1.47}$$

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

$$c_t = \frac{1}{\sum_{j=1}^{100} t^j}; \tag{1.48}$$

Figure 1.19 considers several values for t.

Figure 1.19 plots the number n of draws required to distinguish the actual distribution defined in (1.47) and (1.48) from the model distribution defined in (1.45) and (1.46), estimating the parameter θ in (1.45) and (1.46) via maximum-likelihood methods. (Remark 1.5.1 above specifies what we mean by "distinguish.")

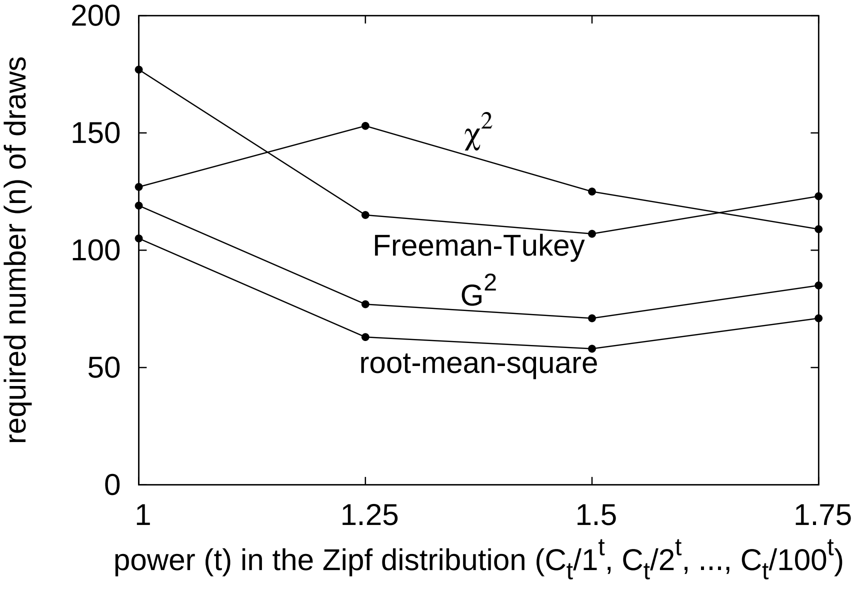


Figure 1.20: Second example; see Subsection 1.5.2.2.

1.5.2.2 A rebinned geometric distribution and a truncated power-law

Let us specify the model distribution to be

$$p_0^{(j)}(\theta) = \theta^{j-1}(1-\theta) \tag{1.49}$$

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

$$p_0^{(100)}(\theta) = \theta^{99}; \tag{1.50}$$

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

$$p^{(j)} = \frac{C_t}{j^t} \tag{1.51}$$

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

$$C_t = \frac{1}{\sum_{j=1}^{100} 1/j^t}; \tag{1.52}$$

Figure 1.20 considers several values for t.

Figure 1.20 plots the number n of draws required to distinguish the actual distribution defined in (1.51) and (1.52) from the model distribution defined in (1.49) and (1.50), estimating the parameter θ in (1.49) and (1.50) via maximum-likelihood methods. (Remark 1.5.1 above specifies what we mean by "distinguish.")

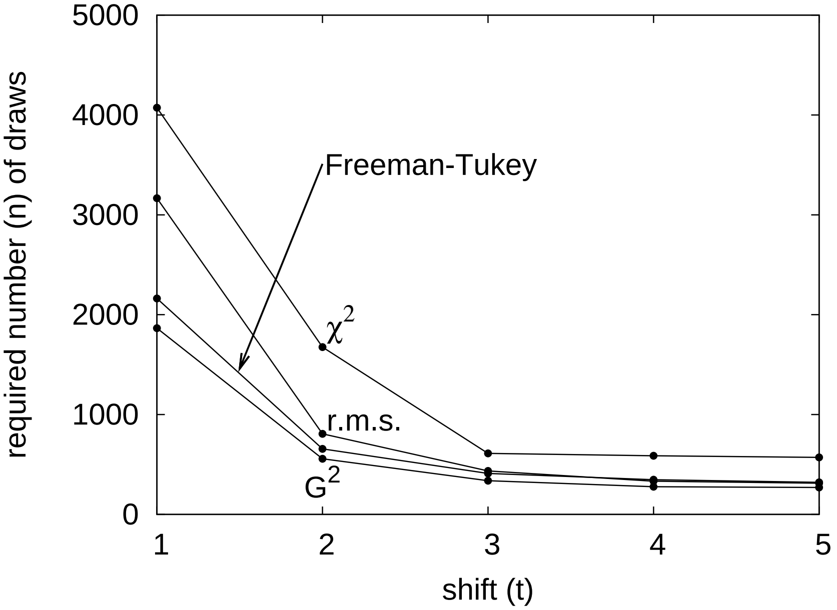


Figure 1.21: Third example; see Subsection 1.5.2.3.

1.5.2.3 Truncated shifted Poisson distributions

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

$$p_0^{(j)}(\theta) = \frac{B_\theta\, \theta^{j-1}}{(j-1)!} \tag{1.53}$$

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

$$B_\theta = \frac{1}{\sum_{j=1}^{21} \theta^{j-1}/(j-1)!}; \tag{1.54}$$

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

$$p^{(j)} = \frac{\tilde{B}_t\, 5^{j-1+t}}{(j-1+t)!} \tag{1.55}$$

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

$$\tilde{B}_t = \frac{1}{\sum_{j=1}^{21} 5^{j-1+t}/(j-1+t)!}; \tag{1.56}$$

Figure 1.21 considers several values for t. Clearly, $p^{(j)} = p_0^{(j)}(5)$ for $j = 1, 2, \dots, 21$, if $t = 0$.

Figure 1.21 plots the number n of draws required to distinguish the actual distribution defined in (1.55) and (1.56) from the model distribution defined in (1.53) and (1.54), estimating the parameter θ in (1.53) and (1.54) via maximum-likelihood methods. (Remark 1.5.1 above specifies what we mean by "distinguish.")

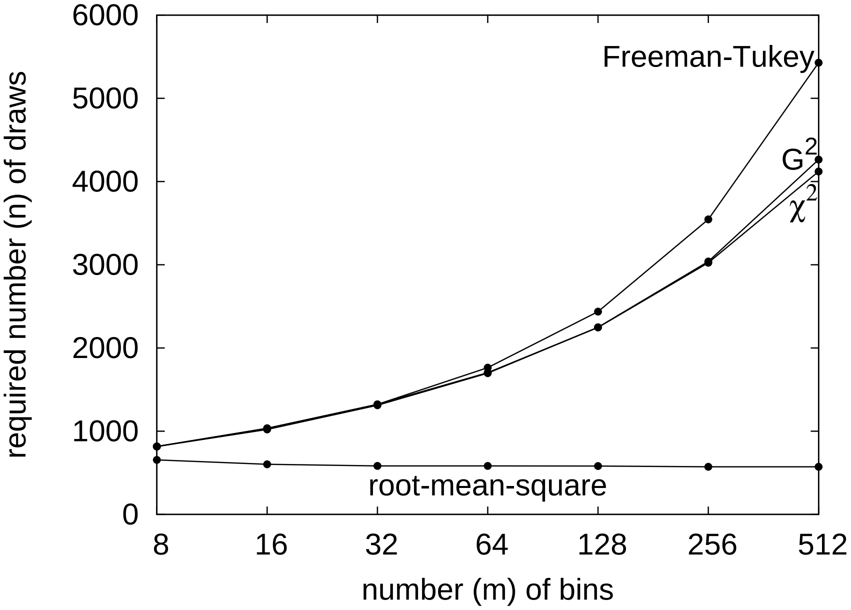


Figure 1.22: Fourth example; see Subsection 1.5.2.4.

1.5.2.4 An example with a uniform tail

Let us specify the model distribution to be

$$p_0^{(1)}(\theta) = \theta, \tag{1.57}$$

$$p_0^{(2)}(\theta) = \theta, \tag{1.58}$$

$$p_0^{(3)}(\theta) = \frac{1}{2} - 2\theta, \tag{1.59}$$

$$p_0^{(j)}(\theta) = \frac{1}{2m-6} \tag{1.60}$$

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

$$p^{(1)} = \frac{1}{4}, \tag{1.61}$$

$$p^{(2)} = \frac{1}{8}, \tag{1.62}$$

$$p^{(3)} = \frac{1}{8}, \tag{1.63}$$

$$p^{(j)} = \frac{1}{2m-6} \tag{1.64}$$

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

Figure 1.22 plots the number n of draws required to distinguish the actual distribution defined in (1.61)–(1.64) from the model distribution defined in (1.57)–(1.60), estimating the parameter θ in (1.57)–(1.60) via maximum-likelihood methods. (Remark 1.5.1 above specifies what we mean by “distinguish.”)

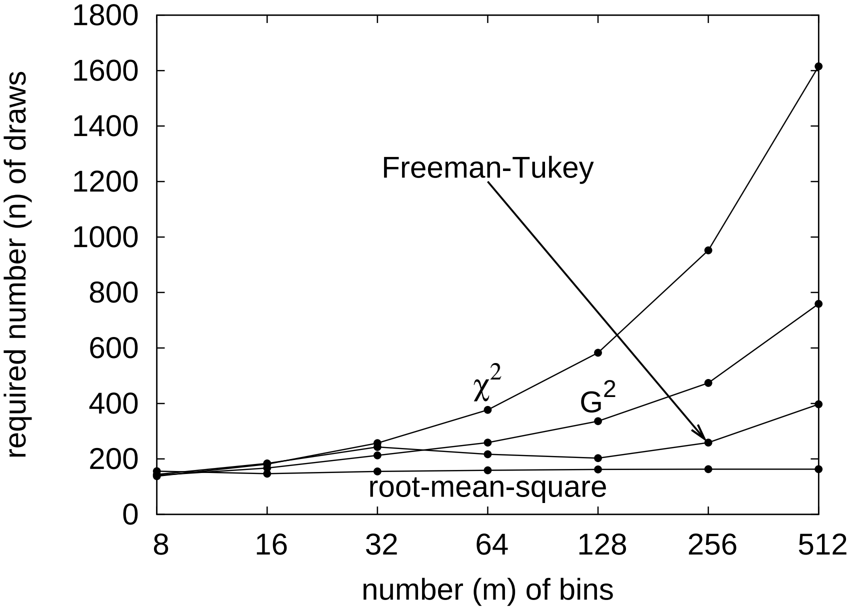


Figure 1.23: Fifth example; see Subsection 1.5.2.5.

1.5.2.5 A model with an integer-valued parameter

Let us specify the model distribution to be

$$p_0^{(j)}(\theta) = \frac{1}{2\theta} \tag{1.65}$$

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

$$p_0^{(j)}(\theta) = \frac{1}{2(m-\theta)} \tag{1.66}$$

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

$$p^{(1)} = \frac{1}{4}, \tag{1.67}$$

$$p^{(2)} = \frac{1}{4}, \tag{1.68}$$

$$p^{(3)} = \frac{1}{4}, \tag{1.69}$$

and

$$p^{(j)} = \frac{1}{4m - 12} \tag{1.70}$$

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

Figure 1.23 plots the number n of draws required to distinguish the actual distribution defined in (1.67)–(1.70) from the model distribution defined in (1.65) and (1.66), estimating the parameter θ in (1.65) and (1.66) via maximum-likelihood methods. (Remark 1.5.1 above specifies what we mean by “distinguish.”)

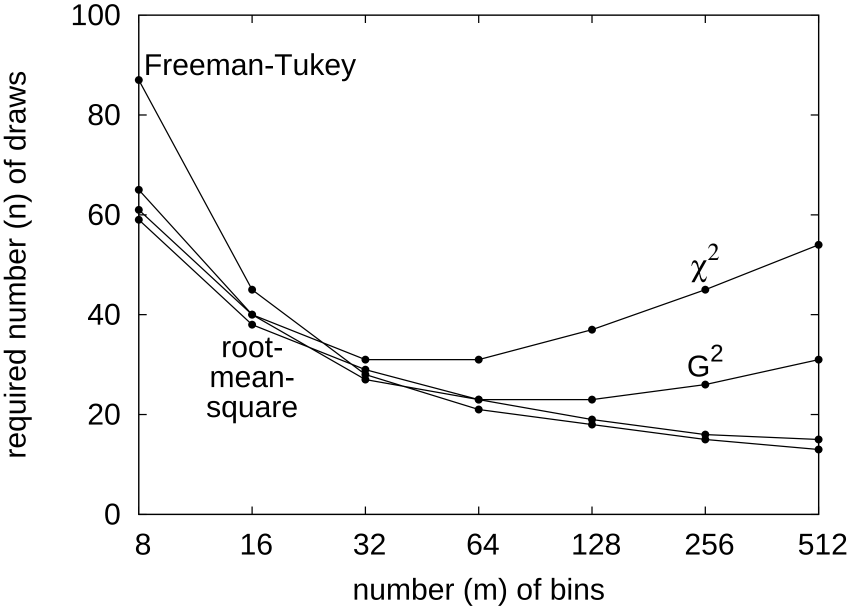


Figure 1.24: Sixth example; see Subsection 1.5.2.6.

1.5.2.6 Truncated power-laws parameterized with a permutation

Let us specify the model to be the Zipf distribution

$$p_0^{(j)}(\theta) = \frac{C_1}{\theta(j)} \tag{1.71}$$

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

$$C_1 = \frac{1}{\sum_{j=1}^{m} 1/j}; \tag{1.72}$$

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

$$p^{(j)} = \frac{C_2}{j^2} \tag{1.73}$$

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

$$C_2 = \frac{1}{\sum_{j=1}^{m} 1/j^2}. \tag{1.74}$$

Figure 1.24 plots the number n of draws required to distinguish the actual distribution defined in (1.73) and (1.74) from the model distribution defined in (1.71) and (1.72), estimating the parameter θ in (1.71) via maximum-likelihood methods (that is, by sorting). (Remark 1.5.1 above specifies what we mean by “distinguish.”)

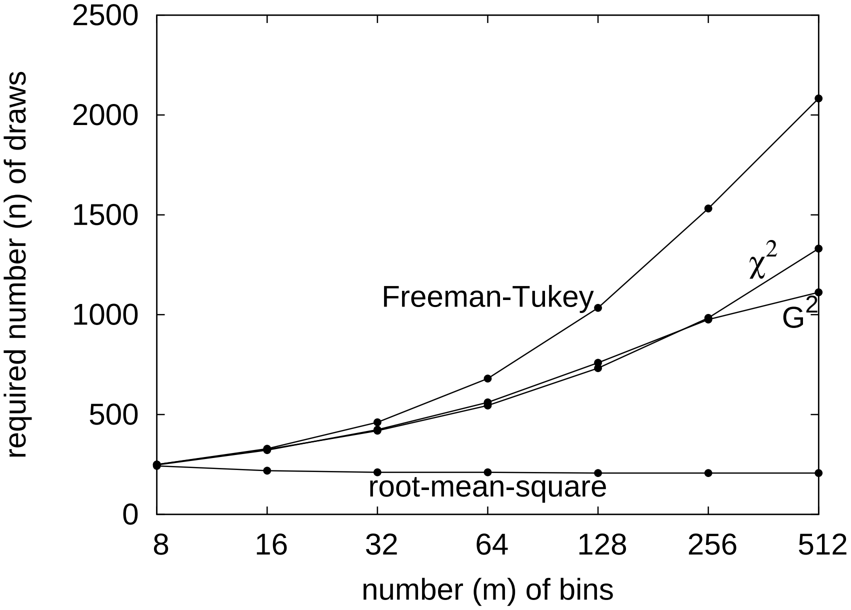


Figure 1.25: Seventh example; see Subsection 1.5.2.7.

1.5.2.7 A model with two parameters

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

$$p_0^{(1)}(\theta_1, \theta_2) = \theta_1, \tag{1.75}$$

$$p_0^{(2)}(\theta_1, \theta_2) = \theta_1, \tag{1.76}$$

$$p_0^{(3)}(\theta_1, \theta_2) = \theta_2, \tag{1.77}$$

$$p_0^{(4)}(\theta_1, \theta_2) = \theta_2, \tag{1.78}$$

and

$$p_0^{(j)}(\theta_1, \theta_2) = \frac{1 - 2\theta_1 - 2\theta_2}{m - 4} \tag{1.79}$$

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

$$p^{(1)} = \frac{9}{32}, \tag{1.80}$$

$$p^{(2)} = \frac{3}{32}, \tag{1.81}$$

$$p^{(3)} = \frac{3}{32}, \tag{1.82}$$

$$p^{(4)} = \frac{1}{32}, \tag{1.83}$$

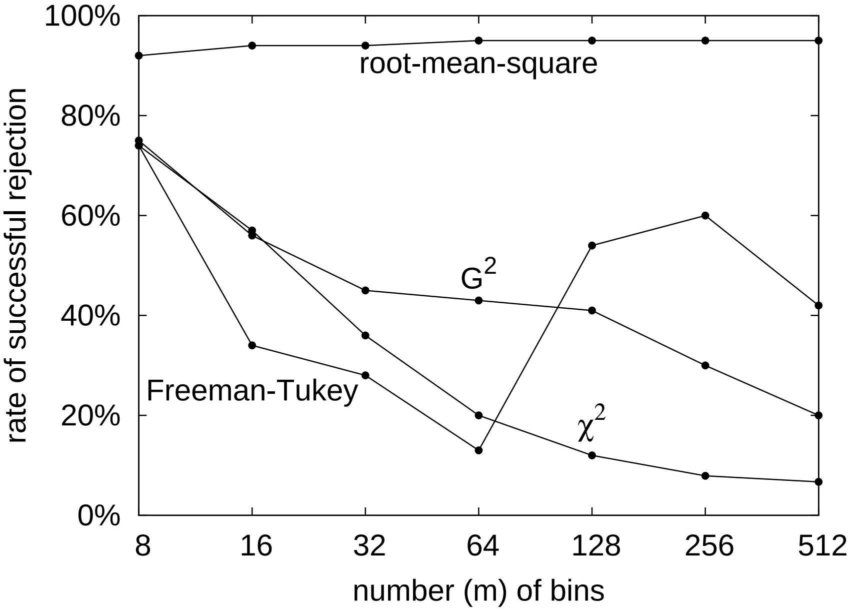


Figure 1.26: First example, with $m = 100$ draws; see Subsection 1.5.1.1.

and

$$p^{(j)} = \frac{1}{2m - 8} \tag{1.84}$$

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

Figure 1.25 plots the number n of draws required to distinguish the actual distribution defined in (1.80)–(1.84) from the model distribution defined in (1.75)–(1.79), estimating the parameters θ_1 and θ_2 in (1.75)–(1.79) via maximum-likelihood methods. (Remark 1.5.1 above specifies what we mean by "distinguish.")

1.6 Additional plots of power and efficiency

For each plot in Section 1.5, the present section provides a corresponding plot based on a confidence level of 95% (that is, a significance level of 5%), rather than a confidence level of 99% (that is, a significance level of 1%). In this section, Figures 1.27–1.40 set the probabilities of false positives and false negatives both to be 5% in order to determine the required number n of draws, whereas in Section 1.5 above Figures 1.12–1.25 set the probabilities of false positives and false negatives both to be 1% (see Remark 1.5.1). Similarly, a rejection is deemed successful for Figure 1.26 at the 5% significance level (or better), whereas a rejection is deemed successful for Figure 1.11 only at the stricter 1% significance level (or better).

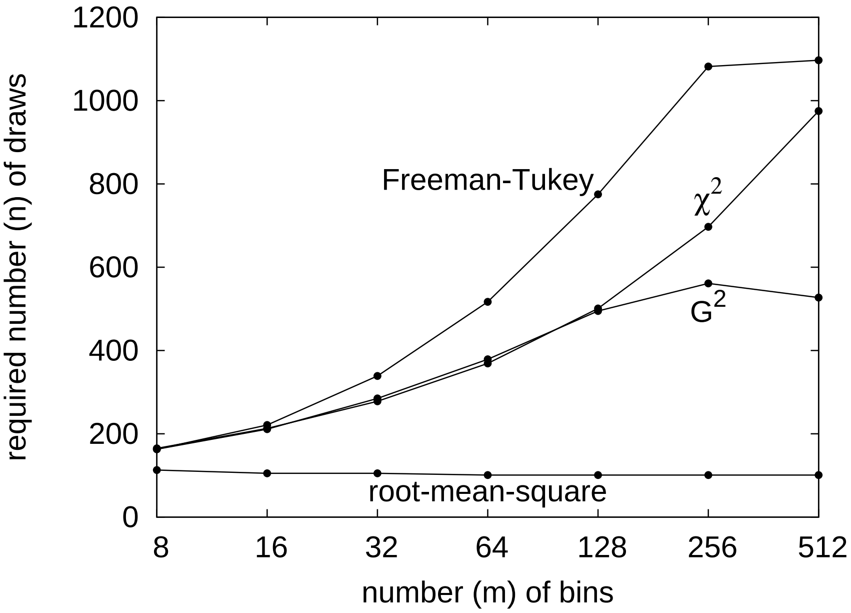


Figure 1.27: First example (statistical “efficiency”); see Subsection 1.5.1.1.

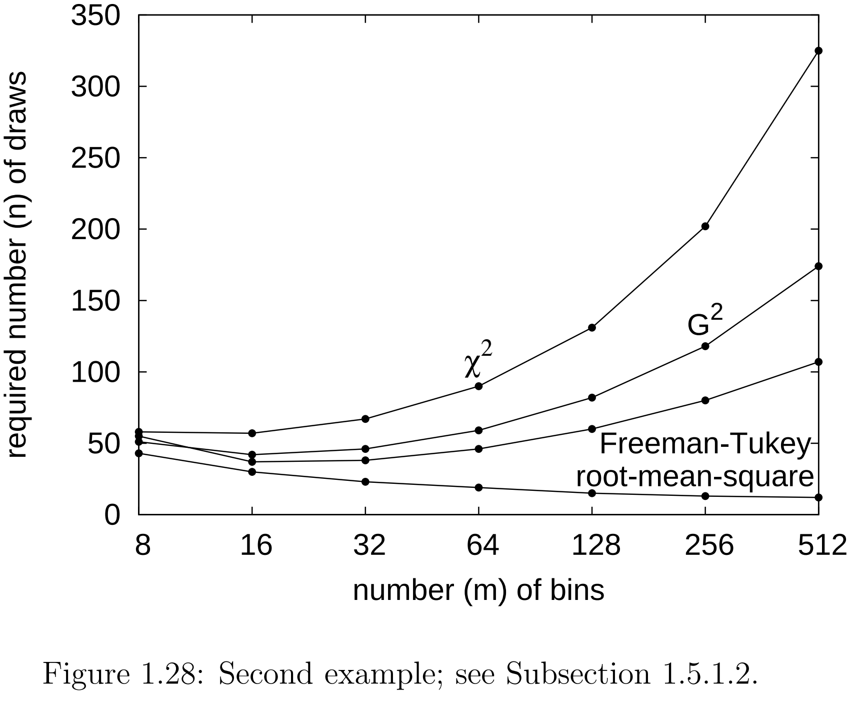


Figure 1.28: Second example; see Subsection 1.5.1.2.

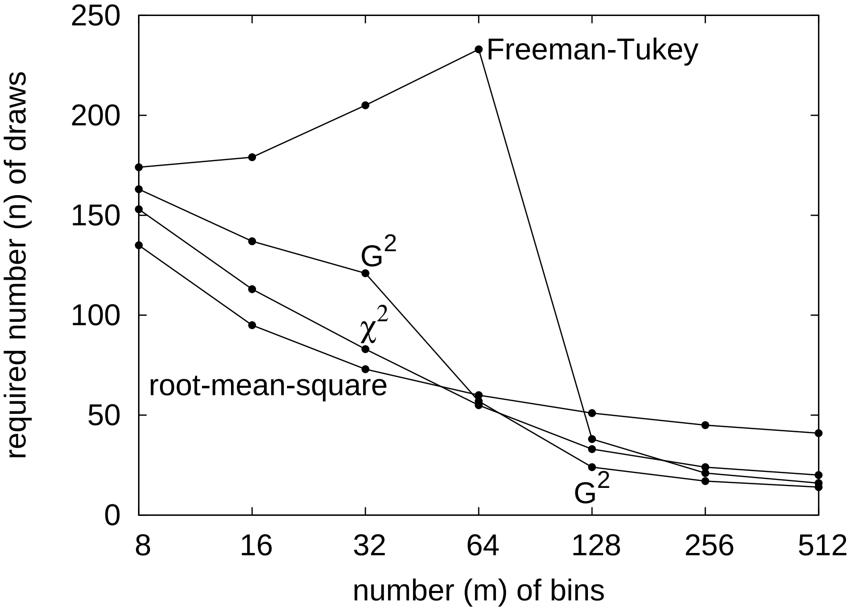


Figure 1.29: Third example; see Subsection 1.5.1.3.

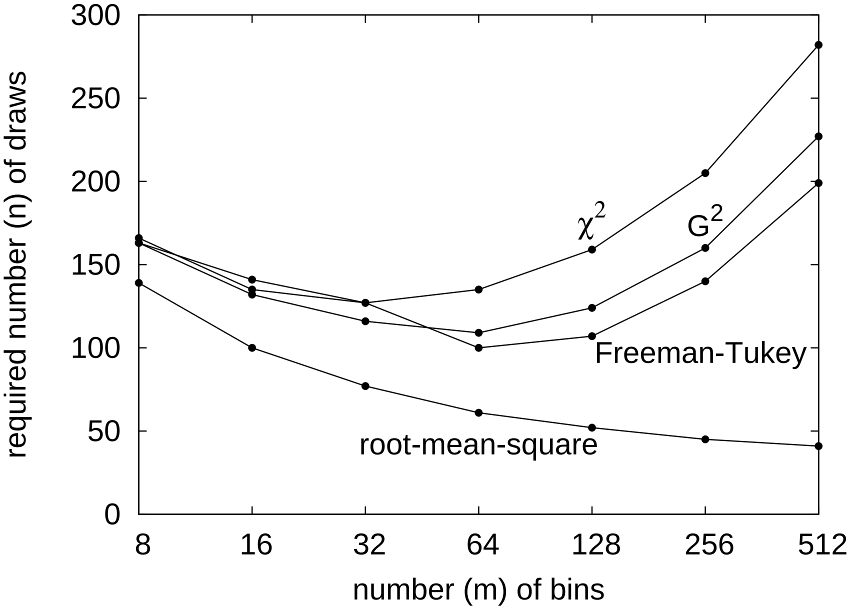


Figure 1.30: Fourth example; see Subsection 1.5.1.4.

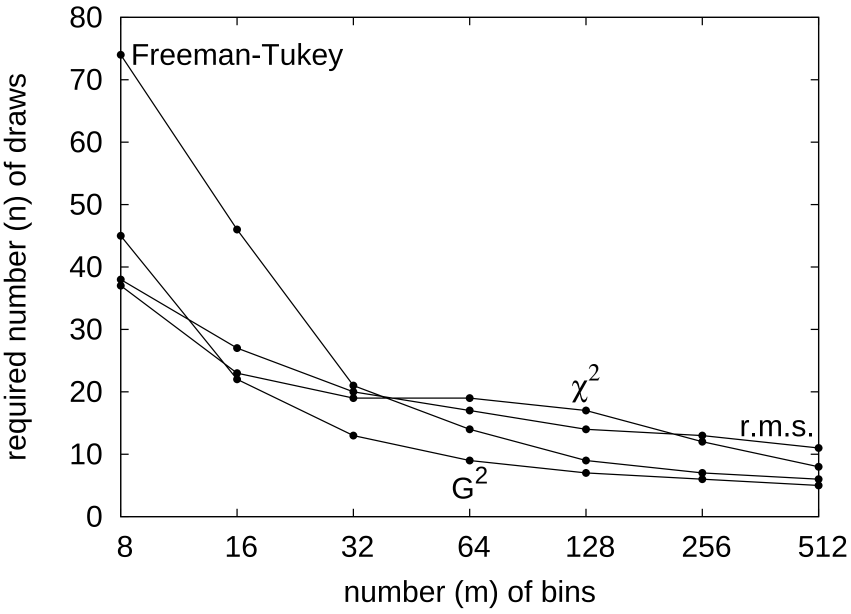


Figure 1.31: Fifth example; see Subsection 1.5.1.5.

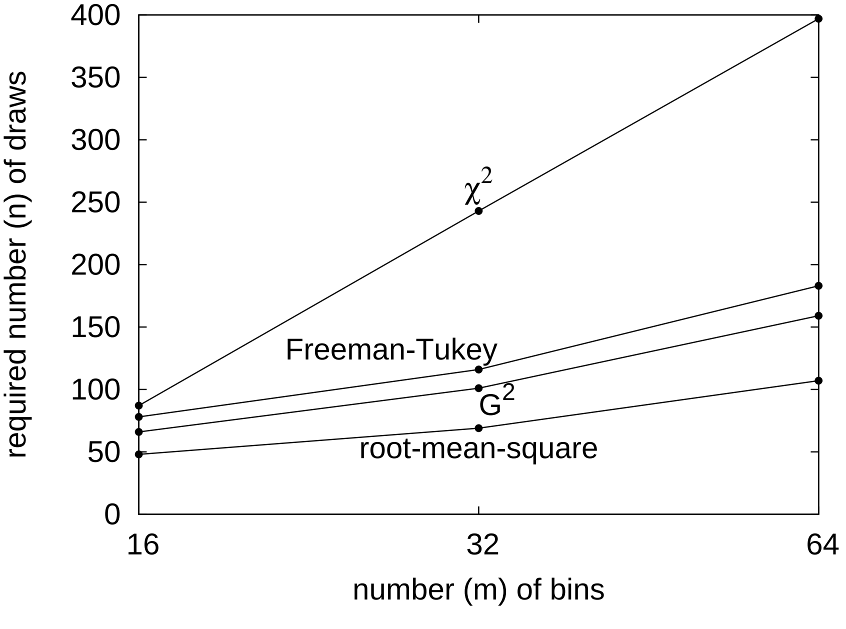


Figure 1.32: Sixth example; see Subsection 1.5.1.6.

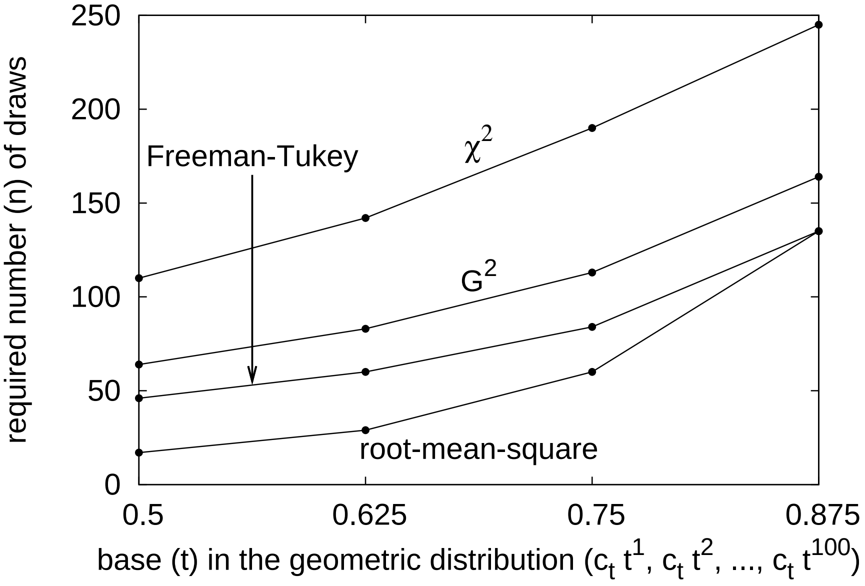


Figure 1.33: Seventh example; see Subsection 1.5.1.7.

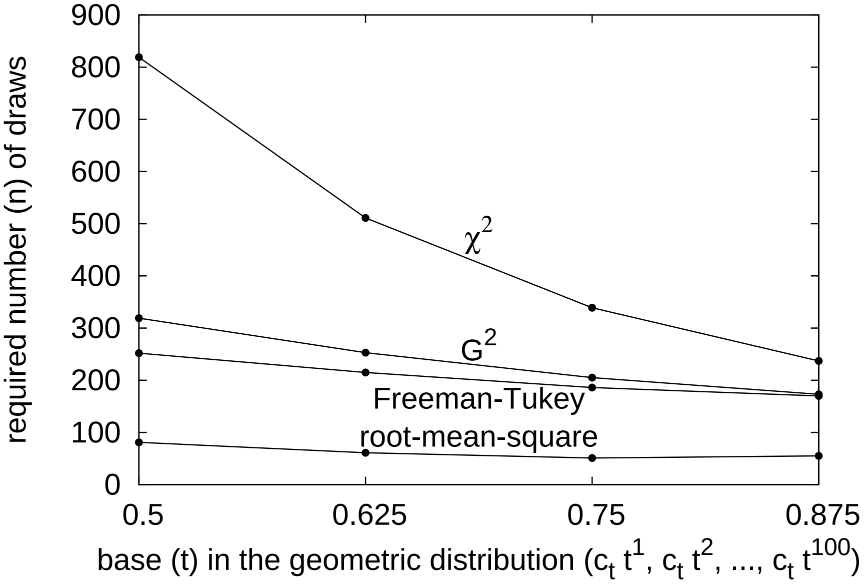


Figure 1.34: First example; see Subsection 1.5.2.1.

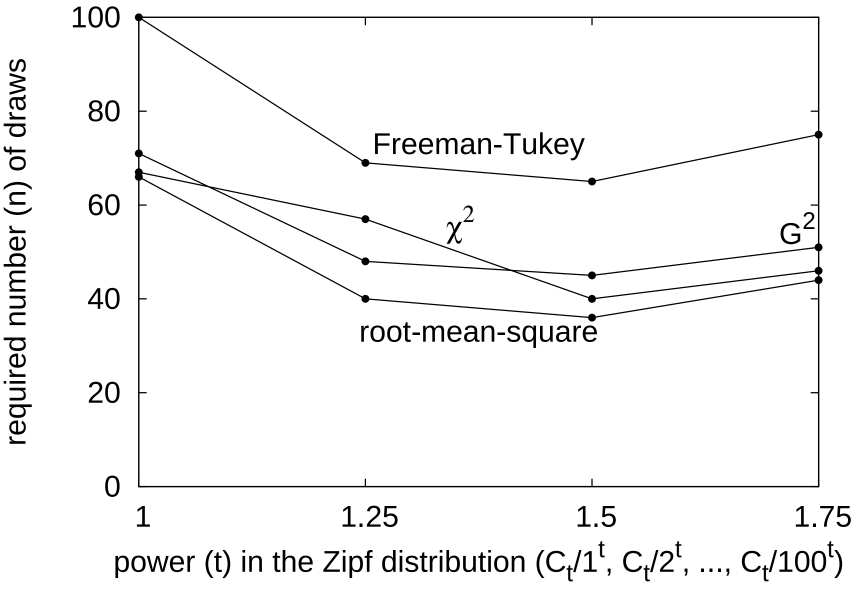


Figure 1.35: Second example; see Subsection 1.5.2.2.

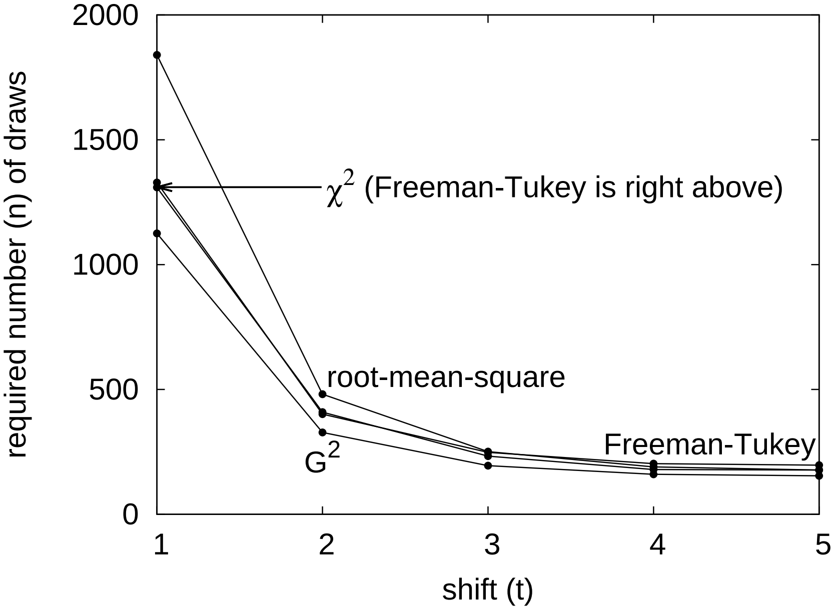


Figure 1.36: Third example; see Subsection 1.5.2.3.

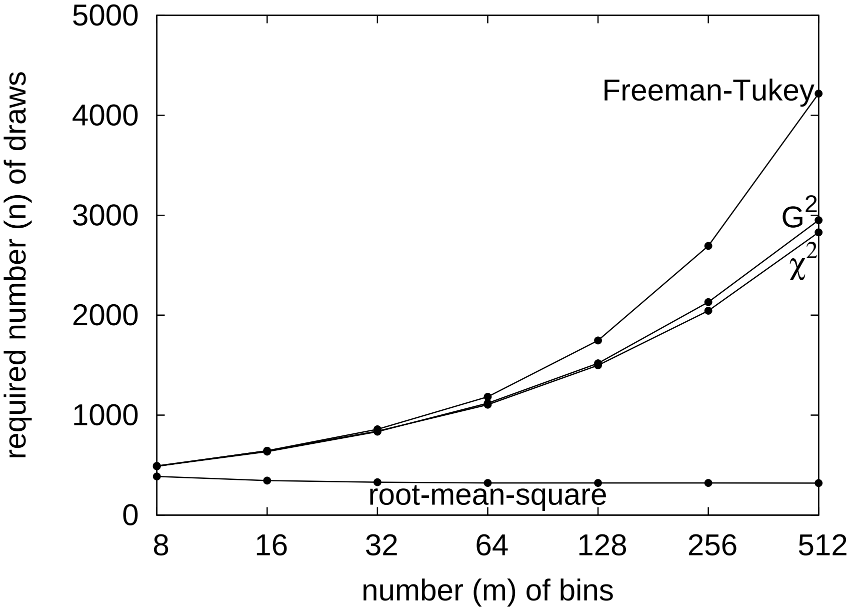


Figure 1.37: Fourth example; see Subsection 1.5.2.4.

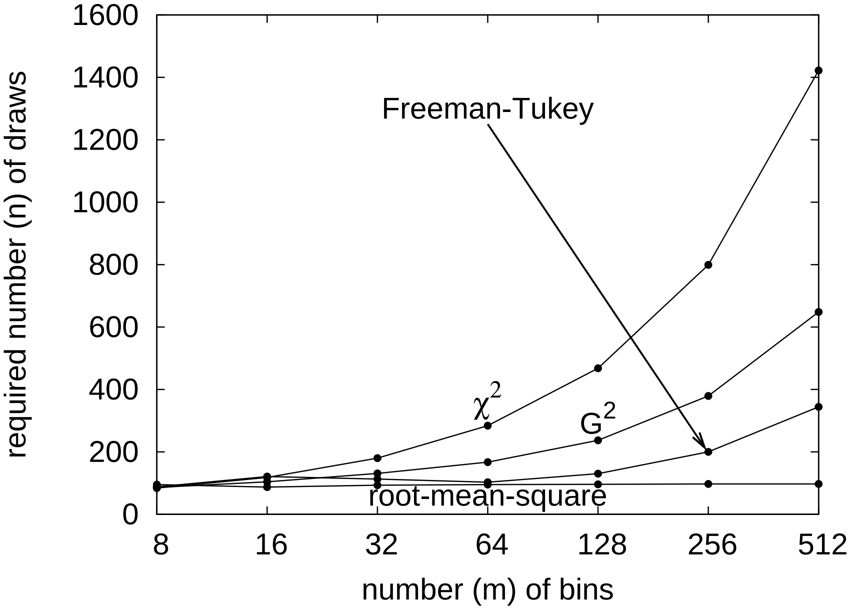


Figure 1.38: Fifth example; see Subsection 1.5.2.5.

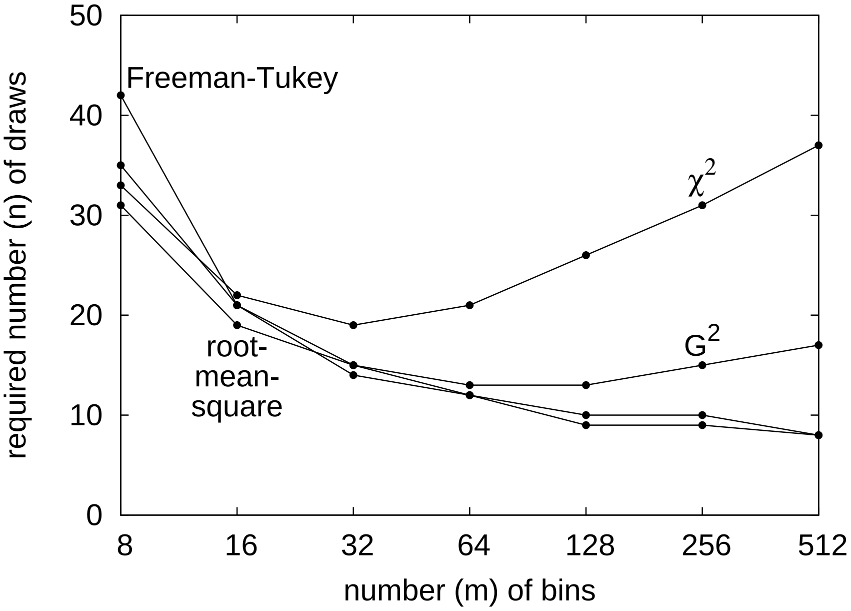


Figure 1.39: Sixth example; see Subsection 1.5.2.6.

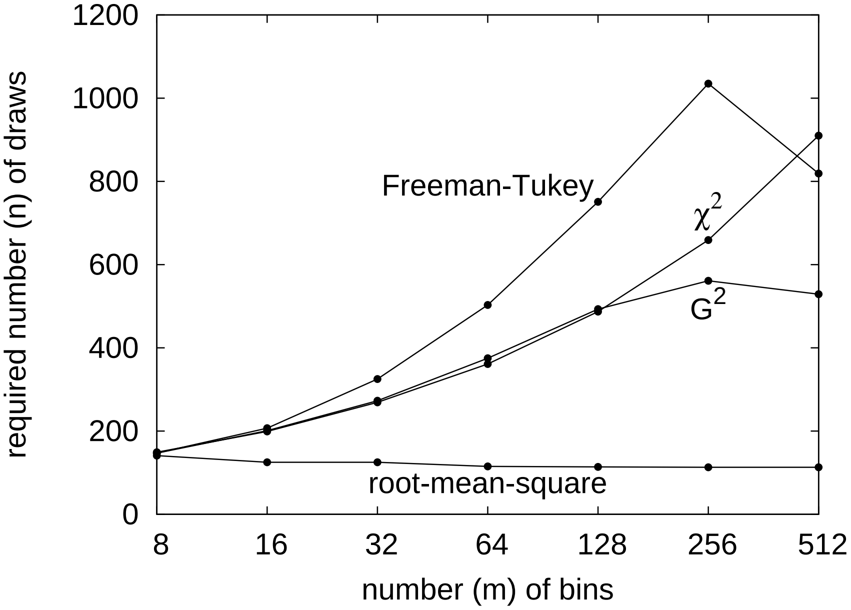


Figure 1.40: Seventh example; see Subsection 1.5.2.7.

Chapter 2

Nominal versus ordinal

Goodness-of-fit tests gauge whether a given set of observations is consistent (up to expected random fluctuations) with arising as independent and identically distributed (i.i.d.) draws from a user-specified probability distribution known as the "model." The standard gauges involve the discrepancy between the model and the empirical distribution of the observed draws. Some measures of discrepancy are cumulative; others are not. The most popular cumulative measure is the Kolmogorov-Smirnov statistic; when all probability distributions under consideration are discrete, a natural noncumulative measure is the root-mean-square distance between the model and the empirical distributions. In the present chapter, both mathematical analysis and its illustration via various data sets indicate that the Kolmogorov-Smirnov statistic tends to be more powerful than the root-mean-square distance when there is a natural ordering for the values that the draws can take — that is, when the data is ordinal — whereas the root-mean-square distance is more reliable and more easily understood than the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order) — that is, when the data is nominal.

2.1 Recap

Testing goodness-of-fit is one of the foundations of modern statistics, as elaborated in Chapter 1 and by Rao (2002), for example. The formulation in the discrete setting involves distributions over m bins ("categories," "cells," and "classes" are common synonyms for "bins"). In accordance with the standard conventions, we will use p_0 to denote a user-specified distribution, usually called the "model," with $p_0 = (p_0^{(1)}, p_0^{(2)}, \dots, p_0^{(m)})$, such that $p_0^{(1)}$, $p_0^{(2)}$, …, $p_0^{(m)}$ are nonnegative and

$$\sum_{j=1}^{m} p_0^{(j)} = 1. \tag{2.1}$$

A goodness-of-fit test produces a value — the "P-value" — that gauges the consistency of n observations with arising as independent and identically distributed (i.i.d.) draws from p_0. In many formulations, the user-specified model p_0 consists of a family of probability distributions parameterized by θ, where θ can be integer-valued, real-valued, complex-valued, vector-valued, matrix-valued, or any combination of the many possibilities. In such cases, the P-value gauges the consistency of the observed data with arising as i.i.d. draws from $p_0(\hat{\theta})$, where $\hat{\theta}$ is an estimate (taken to be the maximum-likelihood estimate throughout the present chapter). We now review the definition of P-values.

P-values are defined via the empirical distribution $\hat{p}$, where $\hat{p} = (\hat{p}^{(1)}, \hat{p}^{(2)}, \dots, \hat{p}^{(m)})$, with $\hat{p}^{(j)}$ being the proportion of the n observed draws that fall in the jth bin, that is, $\hat{p}^{(j)}$ is the number of draws falling in the jth bin, divided by n. P-values involve a hypothetical experiment taking n i.i.d. draws from the assumed actual underlying distribution $p_0(\hat{\theta})$. We denote by $\hat{P}$ the empirical distribution of the draws from the hypothetical experiment; we denote by $\hat{\Theta}$ a maximum-likelihood estimate of θ obtained from the hypothetical experiment. The P-value is then the probability that the discrepancy between the random variables $\hat{P}$ and $p_0(\hat{\Theta})$ is at least as large as the observed discrepancy between $\hat{p}$ and $p_0(\hat{\theta})$, calculating the probability under the assumption that the hypothetical observations arise as i.i.d. draws from $p_0(\hat{\theta})$.

2.2 Definitions of the divergences

To complete the definition of P-values, we must choose a measure of discrepancy. In the present chapter, we consider the (discrete) Kolmogorov-Smirnov and root-mean-square distances,

$$d_1(a, b) = \max_{1 \le k \le m} \left| \sum_{j=1}^{k} a^{(j)} - \sum_{j=1}^{k} b^{(j)} \right| \tag{2.2}$$

and

$$d_2(a, b) = \sqrt{\sum_{j=1}^{m} (a^{(j)} - b^{(j)})^2 / m}, \tag{2.3}$$

respectively. The P-value for the Kolmogorov-Smirnov statistic is the probability that $d_1(\hat{P}, p_0(\hat{\Theta})) \ge d_1(\hat{p}, p_0(\hat{\theta}))$; the P-value for the root-mean-square is the probability that $d_2(\hat{P}, p_0(\hat{\Theta})) \ge d_2(\hat{p}, p_0(\hat{\theta}))$. When evaluating the probabilities, we view $\hat{P}$ and $\hat{\Theta}$ as random variables, constructed with i.i.d. draws from the assumed distribution $p_0(\hat{\theta})$, while viewing the observed $\hat{p}$ and $\hat{\theta}$ as fixed, not random.

If a P-value is very small, then we can be confident that the given observed draws are inconsistent with the assumed model, are not i.i.d., or are both inconsistent and not i.i.d.

Needless to say, the Kolmogorov-Smirnov distance defined in (2.2) is the maximum absolute difference between *cumulative* distribution functions. The Kolmogorov-Smirnov statistic depends on the ordering of the bins, unlike the root-mean-square distance.

As supported by the investigations below, we recommend using the Kolmogorov-Smirnov statistic when there is a natural ordering of the bins, while the root-mean-square distance is more reliable and more easily understood than the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order). Unlike the root-mean-square distance, the Kolmogorov-Smirnov statistic utilizes the information in a natural ordering of the bins, when the latter is available. Horn (1977) gave similar recommendations when comparing the χ^2

and Kolmogorov-Smirnov statistics. Detailed comparisons between the root-mean-square distance and χ^2 statistics are available in Chapter 1.

The Kolmogorov-Smirnov statistic is cumulative; it accentuates low-frequency differences between the model and the empirical distribution of the draws, but tends to average away and otherwise obscure high-frequency differences. Similar observations have been made by Pettitt and Stephens (1977), D'Agostino and Stephens (1986), Choulakian et al. (1994), From (1996), Best and Rayner (1997), Haschenburger and Spinelli (2005), Steele and Chaseling (2006), Lockhart et al. (2007), Ampadu (2008), and Ampadu et al. (2009), among others. Our suggestions appear to be closest to those of Horn (1977).

There are many cumulative approaches similar to the Kolmogorov-Smirnov statistic. These include the Cramér–von-Mises, Watson, Kuiper, and Rényi statistics, as well as their Anderson-Darling variants; Section 14.3.4 of Press et al. (2007), Stephens (1970), and Rényi (1953) review these statistics. We ourselves are fond of the Kuiper approach. However, the present chapter focuses on the popular Kolmogorov-Smirnov statistic; the Cramér–von-Mises, Watson, and Kuiper variants are very similar.

Please note that multidimensional problems deserve special methods; see, for example, Section 14.8 ("Do two-dimensional distributions differ?") of Press et al. (2007). Alan Sokal, Salil Vadhan, and others have pointed out to us that the earth-mover distance (also known as the Wasserstein, Monge-Kantorovich, Mallows, Kantorovich-Rubinstein, or Hutchinson distance) is natural for many application domains, including image analysis; see, for example, the recent work by del Barrio et al. (2000) and Munk and Czado (1998).

The remainder of the present chapter has the following structure: Section 2.3 describes how the root-mean-square is generally preferable to the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order) of the bins. Section 2.4 describes how the Kolmogorov-Smirnov statistic is generally preferable to the root-mean-square distance when there is a natural ordering of the bins. Section 2.5 illustrates both cases with examples of data sets and the associated P-values, computing the P-values via Monte-Carlo simulations with guaranteed error bounds. The reader may wish to begin with Section 2.5, referring back to earlier sections as needed.

2.3 The case when the bins do not have a natural order

The root-mean-square distance is generally preferable to the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order) of the bins. As discussed in Chapter 7, the interaction of parameter estimation and the root-mean-square distance is easy to understand and quantify, at least asymptotically, in the limit of large numbers of draws. In contrast, the interaction of parameter estimation and the Kolmogorov-Smirnov statistic can be very complicated, though Choulakian et al. (1994) and Lockhart et al. (2007) have pointed out that the interaction is somewhat simpler with Cramér's and von Mises', Watson's, and some of Anderson's and Darling's very similar statistics. That said, the root-mean-square distance can be more reliable even when there are no parameters in the model, that is, when the model p_0 is a single, fixed, fully specified probability distribution; the remainder of the present section describes why.

The basis of the analysis is the following lemma, a reformulation of the fact that the expected maximum absolute deviation from zero of the standard Brownian bridge is $\sqrt{\pi/2} \cdot \ln(2) \approx .8687$ (see, for example, Section 3 of Marsaglia et al., 2003).

Lemma 2.3.1. *Suppose that m is even and that $D^{(1)}$, $D^{(2)}$, ..., $D^{(m)}$ form a randomly ordered list of $m/2$ positive ones and $m/2$ negative ones (with the ordering drawn uniformly at random). Then,*

$$\mathbf{E} \max_{1 \le k \le m} \left| \sum_{j=1}^{k} D^{(j)} \right| \Big/ \sqrt{m} \quad \longrightarrow \quad \sqrt{\pi/2} \cdot \ln(2) \tag{2.4}$$

in the limit that $m \to \infty$, where (as usual) $\mathbf{E}$ produces the expected value.

We will assume that the given observations arise as i.i.d. draws from an (unknown) actual underlying probability distribution denoted p. We denote by p_0 the model distribution. We denote by $\hat{P}$ the empirical distribution of the n draws. These are all probability distributions, that is, $p^{(j)} \ge 0$, $p_0^{(j)} \ge 0$, and $\hat{P}^{(j)} \ge 0$ for $j = 1, 2, \ldots, m$, and (2.1) and its analogues for p and $\hat{P}$ hold.

Suppose that the actual underlying distribution $p^{(1)}$, $p^{(2)}$, ..., $p^{(m)}$ of the draws is the same as the model distribution $p_0^{(1)}$, $p_0^{(2)}$, ..., $p_0^{(m)}$; the random variables $\hat{P}^{(1)}$, $\hat{P}^{(2)}$, ..., $\hat{P}^{(m)}$ are then the proportions of n i.i.d. draws from p_0 that fall in the respective m bins. The root-sum-square distance is

$$U = \sqrt{\sum_{j=1}^{m} (\hat{P}^{(j)} - p_0^{(j)})^2}. \tag{2.5}$$

The Kolmogorov-Smirnov statistic is

$$V = \max_{1 \le k \le m} \left| \sum_{j=1}^{k} (\hat{P}^{(j)} - p_0^{(j)}) \right|. \tag{2.6}$$

The expected value of the square of the root-sum-square distance is

$$\mathbf{E}\, U^2 = \sum_{j=1}^{m} \mathbf{E}\, (\hat{P}^{(j)} - p_0^{(j)})^2 = \sum_{j=1}^{m} \frac{p_0^{(j)}}{n} = \frac{1}{n}. \tag{2.7}$$

As shown, for example, by Durbin (1972) using Lemma 2.3.1 above, the expected value of $\sqrt{n}$ times the Kolmogorov-Smirnov statistic is

$$\mathbf{E}\, V \sqrt{n} \to \sqrt{\pi/2} \cdot \ln(2) \approx .8687 \tag{2.8}$$

in the limit that $n \to \infty$ and $\max_{1 \le j \le m} p_0^{(j)} \to 0$. Comparing (2.7) and (2.8), we see that U and V are roughly the same size (inversely proportional to $\sqrt{n}$) when the actual underlying distribution of the draws is the same as the model distribution.

However, when the actual underlying distribution of the draws differs from the model distribution, the root-sum-square distance and the Kolmogorov-Smirnov statistic can be very different. If the number n of draws is large, then the empirical distribution $\hat{P}$ will be very close to the actual distribution p. Therefore, to study the performance of the goodness-of-fit statistics as $n \to \infty$ when the actual distribution p differs from the model distribution p_0 (and both are independent of n), we can focus on the difference between p and p_0 (rather than the difference between $\hat{P}$ and p_0). We now define and study the difference

$$d^{(j)} = p^{(j)} - p_0^{(j)} \tag{2.9}$$

for $j = 1, 2, \ldots, m$. The root-sum-square distance between p and p_0 is

$$u = \sqrt{\sum_{j=1}^{m} (d^{(j)})^2}. \tag{2.10}$$

The Kolmogorov-Smirnov statistic (the maximum absolute cumulative difference) is

$$v = \max_{1 \le k \le m} \left| \sum_{j=1}^{k} d^{(j)} \right|. \tag{2.11}$$

For simplicity (and because the following analysis generalizes straightforwardly), let us consider the illustrative case in which $|d^{(1)}| = |d^{(2)}| = \cdots = |d^{(m)}|$, that is,

$$|d^{(j)}| = c_m \tag{2.12}$$

for all $j = 1, 2, \ldots, m$, where c_m is a positive real number (c_m must always satisfy $m \cdot c_m \le 2$, since $m \cdot c_m = \sum_{j=1}^{m} c_m = \sum_{j=1}^{m} |d^{(j)}| \le \sum_{j=1}^{m} [p^{(j)} + p_0^{(j)}] = 2$). Combining (2.9), (2.1) and its analogue for p yields that

$$\sum_{j=1}^{m} d^{(j)} = 0. \tag{2.13}$$

Together, (2.13) and (2.12) imply that m is even and that half of $d^{(1)}$, $d^{(2)}$, $\ldots$, $d^{(m)}$ are equal to $+c_m$, and the other half are equal to $-c_m$.

Combining (2.12) and (2.10) yields that the root-sum-square distance is

$$u = \sqrt{m} \cdot c_m. \tag{2.14}$$

The fact that half of $d^{(1)}$, $d^{(2)}$, $\ldots$, $d^{(m)}$ are equal to $+c_m$, and the other half are equal to $-c_m$, yields that the Kolmogorov-Smirnov statistic v defined in (2.11) could be as small as c_m or as large as $m \cdot c_m/2$, depending on the ordering of the signs in $d^{(1)}$, $d^{(2)}$, $\ldots$, $d^{(m)}$. If all orderings are equally likely (which is equivalent to ordering the bins uniformly at random), then by Lemma 2.3.1 the mean value for v is $\sqrt{m\pi/2} \cdot \ln(2) \cdot c_m \approx \sqrt{m} \cdot .8687 \cdot c_m$ in the limit that m is large (this is the expected maximum absolute deviation from zero of a tied-down random walk with m steps, each of length c_m, that starts and ends at zero; the random walk ends at zero due to (2.13)).

Thus, in the limit that the number n of draws is large (and $\max_{1 \le j \le m} p_0^{(j)} \to 0$, while both the model p_0 and the alternative distribution p are independent of n), the root-mean-square distance and the Kolmogorov-Smirnov statistic have similar statistical power on average, if all orderings of the bins are equally likely. However, the root-mean-square distance is the same for any ordering of the bins, whereas the power of the Kolmogorov-Smirnov statistic depends strongly on the ordering. We see, then, that the root-mean-square distance is more reliable than the Kolmogorov-Smirnov statistic when there is no especially natural ordering for the bins.

Remark 2.3.2. It is possible to use an ordering for which the Kolmogorov-Smirnov statistic attains its greatest value (this corresponds to renumbering the bins such that the differences $D^{(j)} = \hat{P}^{(j)} - p_0^{(j)}$ satisfy $D^{(1)} \ge D^{(2)} \ge \cdots \ge D^{(m)}$ or $D^{(1)} \le D^{(2)} \le \cdots \le D^{(m)}$). However, this data-dependent ordering produces a statistic which is proportional to the l^1 distance $\sum_{j=1}^{m} |D^{(j)}|$ (whereas the root-sum-square distance is the l^2 distance), as remarked at the top of page 396 of Hoeffding (1965). This distance is not so obviously cumulative.

2.4 The case when the bins have a natural order

The Kolmogorov-Smirnov statistic is often preferable to the root-mean-square distance when there is a natural ordering of the bins. In fact, the Kolmogorov-Smirnov statistic is always preferable when the data is very sparse and there is a natural ordering of the bins. In the limit that the maximum expected number of draws per bin tends to zero, the root-mean-square distance always takes the same value under the null hypothesis, providing no discriminative power: indeed, when the draws producing the empirical distribution $\hat{P}$ are taken i.i.d. from the model distribution p_0, the root-mean-square distance is almost surely $1/\sqrt{mn}$,

$$\sqrt{\sum_{j=1}^{m} (\hat{P}^{(j)} - p_0^{(j)})^2/m} = \frac{1}{\sqrt{mn}}, \tag{2.15}$$

in the limit that $n \cdot \max_{1 \le j \le m} p_0^{(j)} \to 0$ (the reason is that, in this limit, $\max_{1 \le j \le m} p_0^{(j)} \to 0$ and moreover almost every realization of the experiment satisfies that, for all j = 1, 2, ..., m, $\hat{P}^{(j)} = 0$ or $\hat{P}^{(j)} = 1/n$, that is, there is at most one observed draw per bin). In contrast, the Kolmogorov-Smirnov statistic is nontrivial even in the limit that the maximum expected number of draws per bin tends to zero — in fact, this is exactly the continuum limit for the original Kolmogorov-Smirnov statistic involving continuous cumulative distribution functions (as opposed to the discontinuous cumulative distribution functions arising from the discrete distributions considered in the present chapter). Furthermore, the Kolmogorov-Smirnov statistic is sensitive to symmetry (or asymmetry) in a distribution, and can detect other interesting properties of distributions that depend on the ordering of the bins.

2.5 Data analysis

This section gives four examples illustrating the performance of the Kolmogorov-Smirnov statistic and the root-mean-square distance in various circumstances. The Kolmogorov-Smirnov statistic is more powerful than the root-mean-square distance in the first two examples, for which there are natural orderings of the bins. The root-mean-square distance is more reliable than the Kolmogorov-Smirnov statistic in the last two examples, for which any ordering of the bins is necessarily rather arbitrary. We computed all P-values via Monte-Carlo simulations with guaranteed error bounds, as in Section 1.3. Remark 1.3.2 proves that the standard error of the obtained estimate for a P-value P is $\sqrt{P(1-P)/\ell}$, where ℓ is the number of simulations conducted to calculate the P-value.

2.5.1 A test of randomness

A particular random number generator is supposed to produce an integer from 1 to 2^{32} uniformly at random. The model distribution for such a generator is

$$p_0^{(j)} = 2^{-32} \tag{2.16}$$

for $j = 1,\ 2,\ \dots,\ 2^{32}$. We test the (obviously poor) generator which produces the numbers 1, 2, 3, …, n, in that order, so that the observed distribution of the generated numbers is

$$\hat{p}^{(j)} = \begin{cases} 1/n, & j = 1,\ 2,\ \dots,\ n \\ 0, & j = n+1,\ n+2,\ \dots,\ 2^{32} \end{cases} \tag{2.17}$$

for $j = 1, 2, \dots, 2^{32}$. For this data, the P-value for the root-mean-square is 1 to several digits of precision, while the P-value for the Kolmogorov-Smirnov statistic is 0 to several digits, at least for n between a hundred and a million. So, as expected, the root-mean-square has almost no discriminative power for such sparse data, whereas the Kolmogorov-Smirnov statistic easily discerns that the data (2.17) is inconsistent with the model (2.16).

Remark 2.5.1. Like the root-mean-square distance, classical goodness-of-fit statistics such as χ^2, G^2 (the log–likelihood-ratio), and the Freeman-Tukey/Hellinger distance are invariant to the ordering of the bins, and also produce P-values that are equal to 1 to several digits of precision, at least for n between a hundred and a million. For definitions and further discussion of the χ^2, G^2, and Freeman-Tukey statistics, see Section 1.2.

2.5.2 A test of Poissonity

A Poisson-distributed random number generator with mean 100 is supposed to produce a nonnegative integer according to the model

$$p_0^{(j)} = \frac{100^j}{j! \cdot \exp(100)} \tag{2.18}$$

for $j = 0,\ 1,\ 2,\ 3,\ \dots$. We test the (obviously poor) generator which produces the numbers 100, 101, 102, …, 109, so that the observed distribution of the numbers is

$$\hat{p}^{(j)} = \begin{cases} 1/10, & j = 100, 101, 102, \dots, 109 \\ 0, & \text{otherwise} \end{cases} \tag{2.19}$$

for $j = 0,\ 1,\ 2,\ 3,\ \dots$. The P-values, each computed via 4,000,000 simulations, are

- Kolmogorov-Smirnov: .0075

- root-mean-square distance: .998
- χ^2: .999
- G^2 (the log–likelihood-ratio): .999
- Freeman-Tukey (the Hellinger distance): .998

For definitions and further discussion of the χ^2, G^2, and Freeman-Tukey statistics, see Section 1.2. The Kolmogorov-Smirnov statistic is far more powerful for this example, in which the bins have a natural ordering (in this example the bins are the nonnegative integers).

Figure 2.1 plots the model probabilities $p_0^{(0)}$, $p_0^{(1)}$, $p_0^{(2)}$, … defined in (2.18) along with the observed proportions $\hat{p}^{(0)}$, $\hat{p}^{(1)}$, $\hat{p}^{(2)}$, … defined in (2.19). Figure 2.2 plots the model probabilities $p_0^{(0)}$, $p_0^{(1)}$, $p_0^{(2)}$, … along with analogues of the proportions $\hat{p}^{(0)}$, $\hat{p}^{(1)}$, $\hat{p}^{(2)}$, … for a simulation generating 10 i.i.d. draws according to the model.

Figure 2.3 plots the cumulative model probabilities $p_0^{(0)}$, $p_0^{(0)} + p_0^{(1)}$, $p_0^{(0)} + p_0^{(1)} + p_0^{(2)}$, … along with the cumulative observed proportions $\hat{p}^{(0)}$, $\hat{p}^{(0)} + \hat{p}^{(1)}$, $\hat{p}^{(0)} + \hat{p}^{(1)} + \hat{p}^{(2)}$, …. Figure 2.4 plots the cumulative model probabilities $p_0^{(0)}$, $p_0^{(0)} + p_0^{(1)}$, $p_0^{(0)} + p_0^{(1)} + p_0^{(2)}$, … along with analogues of the cumulative proportions $\hat{p}^{(0)}$, $\hat{p}^{(0)} + \hat{p}^{(1)}$, $\hat{p}^{(0)} + \hat{p}^{(1)} + \hat{p}^{(2)}$, … for the simulation generating 10 i.i.d. draws according to the model.

2.5.3 A test of Hardy-Weinberg equilibrium

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 (1992), namely to arise via random sampling from the model

$$p_0^{(j,k)}(\theta_1, \theta_2, \dots, \theta_9) = \begin{cases} 2 \cdot \theta_j \cdot \theta_k, & j > k \\ (\theta_k)^2, & j = k \end{cases} \tag{2.20}$$

for $j, k = 1, 2, \dots, 9$ with $j \geq k$, under the constraint that

$$\sum_{j=1}^{9} \theta_j = 1, \tag{2.21}$$

where the parameters θ_1, θ_2, …, θ_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 $j, k = 1, 2, \dots, 9$ with $j \geq k$, 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 θ_1, θ_2, …, θ_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 (2.20); if a goodness-of-fit statistic rejects the model with high confidence, then we can be confident that mating has not been suitably random.

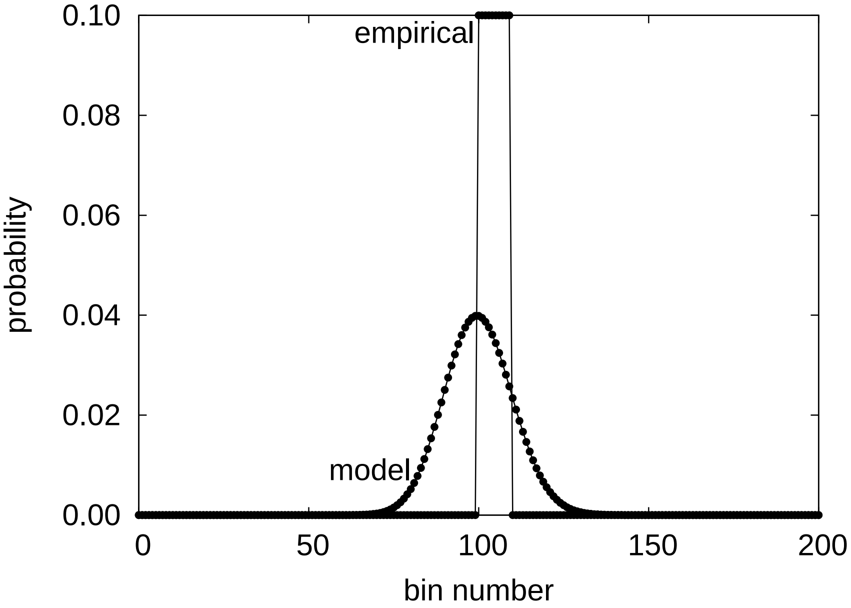


Figure 2.1: Proportions associated with the bins for the observations

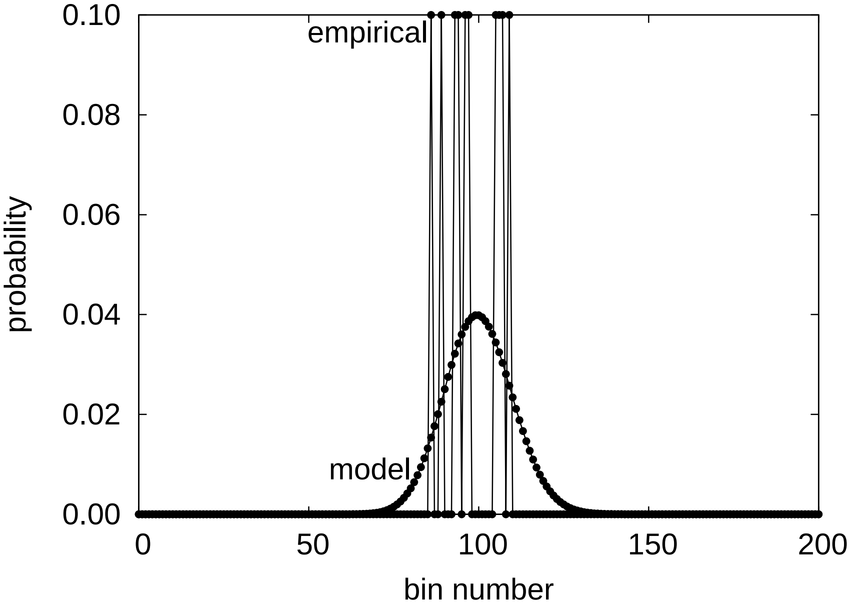


Figure 2.2: Proportions associated with the bins for a simulation

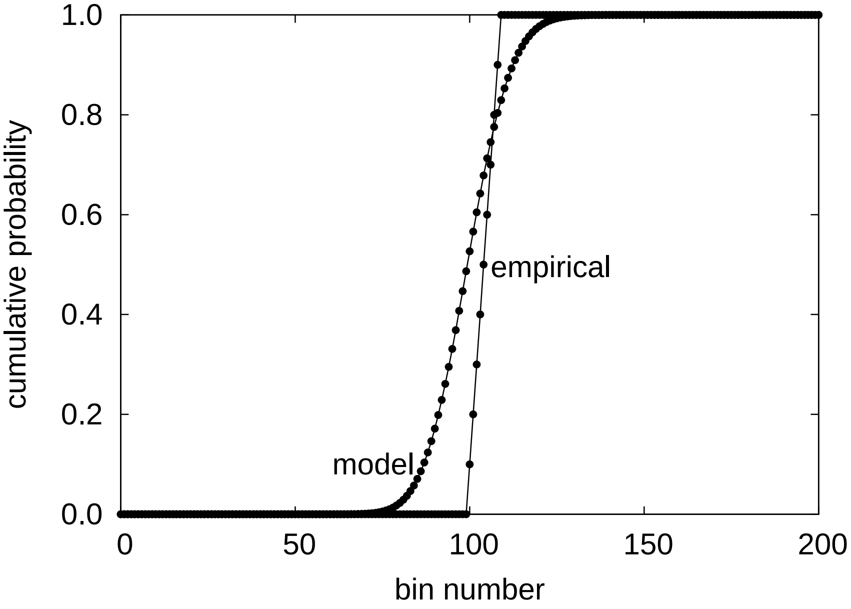


Figure 2.3: Cumulative proportions associated with the bins for the observations

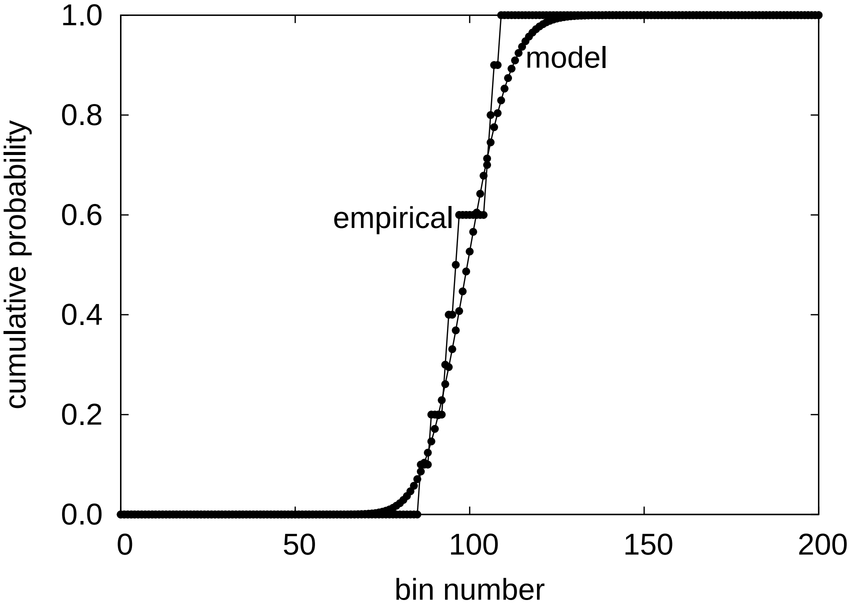


Figure 2.4: Cumulative proportions associated with the bins for the simulation from Figure 2.2

Table 2.1 provides data on $n = 8297$ individuals; we duplicated Figure 3 of Guo and Thompson (1992) to obtain Table 2.1. Figure 2.5 plots the associated P-values, each computed via 90,000 Monte-Carlo simulations. The Kolmogorov-Smirnov statistic depends on the ordering of the bins; for the first trial $t = 1$ in Figure 2.5, the order of the bins is the lexicographical ordering, namely $(1, 1)$, $(2, 1)$, $(2, 2)$, $(3, 1)$, $(3, 2)$, $(3, 3)$, ..., $(9, 9)$. The nine trials $t = 2$, 3, ..., 10 displayed in Figure 2.5 use pseudorandom orderings of the bins. Please note that the root-mean-square distance does not depend on the ordering.

Generally, a more powerful statistic produces lower P-values. In Figure 2.5, the P-values for the Kolmogorov-Smirnov statistic are sometimes lower, sometimes higher than the P-values for the root-mean-square distance. There is no particularly natural ordering of the bins for Figure 2.5; Figure 2.5 displays 10 different orderings corresponding to 10 different trials. Figure 2.5 demonstrates that the root-mean-square distance is more reliable than the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order) for the bins.

Remark 2.5.2. The P-values for classical goodness-of-fit statistics are substantially higher; the classical statistics are less powerful for this example. The P-values, each computed via 4,000,000 Monte-Carlo simulations, are

- root-mean-square distance: .039
- χ^2: .693
- G^2 (the log–likelihood-ratio): .600
- Freeman-Tukey (the Hellinger distance): .562

For definitions and further discussion of the χ^2, G^2, and Freeman-Tukey statistics, see Section 1.2. Like the root-mean-square distance, the χ^2, G^2, and Freeman-Tukey statistics are all invariant to the ordering of the bins.

2.5.4 A test of uniformity

Table 2.2 duplicates Table 1 of Gilchrist (2010), giving the colors of the $n = 62$ pieces of candy in a 2.17 ounce bag. Figure 2.6 plots the P-values for Table 2.2 to be consistent up to expected random fluctuations with Table 2.3, the model of uniform proportions. We computed each P-value via 4,000,000 Monte-Carlo simulations. The Kolmogorov-Smirnov statistic depends on the ordering of the bins; the ten trials $t = 1$, 2, ..., 10 displayed in Figure 2.6 use pseudorandom orderings of the bins. The root-mean-square distance does not depend on the ordering.

Generally, a more powerful statistic produces lower P-values. In Figure 2.6, the P-values for the Kolmogorov-Smirnov statistic are sometimes lower, sometimes higher than the P-values for the root-mean-square distance. There is no particularly natural ordering of the bins for Table 2.3; Figure 2.6 displays 10 different pseudorandom orderings corresponding to 10 different trials. Figure 2.6 illustrates that the root-mean-square distance is more reliable than the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order) for the bins.

Table 2.1: Frequencies of pairs of Rhesus haplotypes

		k								
	$j \backslash k$	1	2	3	4	5	6	7	8	9
	1	1236								
	2	120	3							
	3	18	0	0						
	4	982	55	7	249					
j	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

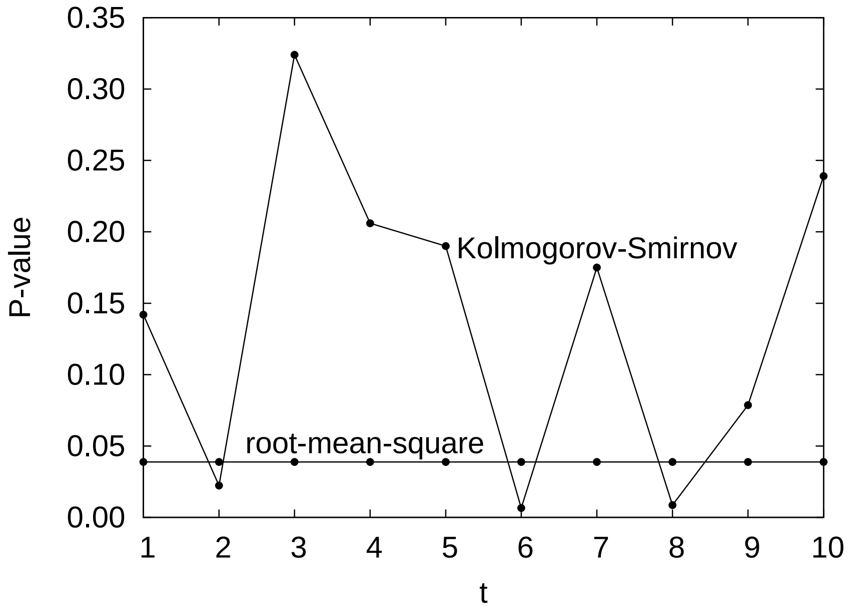


Figure 2.5: P-values for Table 2.1 to be consistent with formula (2.20)

Table 2.2: Observed frequencies of colors of candies in a 2.17 ounce bag

color	red	orange	yellow	green	violet
number	15	9	14	11	13

Table 2.3: Expected frequencies of colors of candies in a 2.17 ounce bag

color	red	orange	yellow	green	violet
number	12.4	12.4	12.4	12.4	12.4

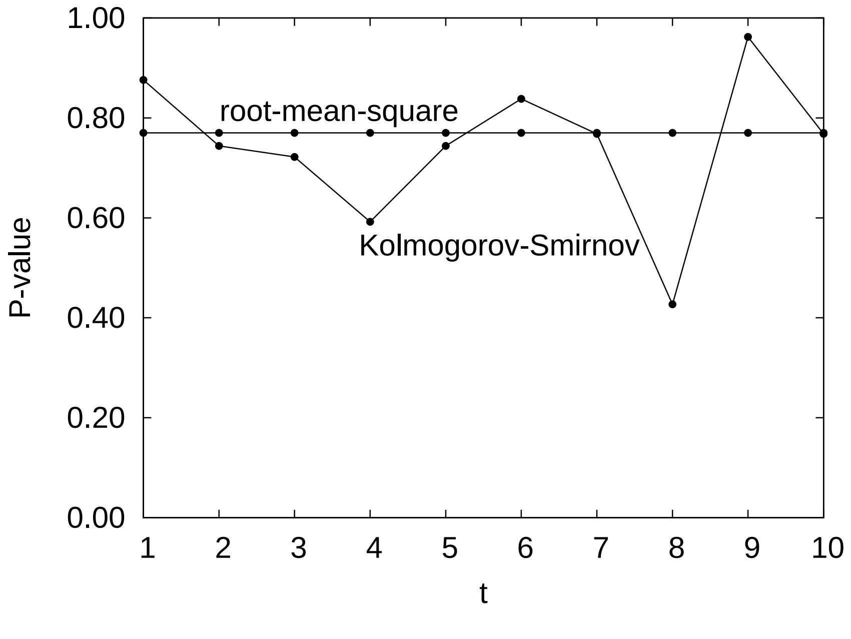


Figure 2.6: P-values for Table 2.2 to be consistent with the model displayed in Table 2.3

Remark 2.5.3. Table 2.2 provides a possible means for ordering the bins. However, such an ordering will depend on the observed data. Using a data-dependent ordering can profoundly alter the nature of the goodness-of-fit statistic; see Remark 2.3.2.

Remark 2.5.4. Like the root-mean-square distance, many classical goodness-of-fit statistics are invariant to the ordering of the bins. The following are P-values, each computed via 4,000,000 Monte-Carlo simulations:

- root-mean-square distance: .770
- χ^2: .770
- G^2 (the log–likelihood-ratio): .766
- Freeman-Tukey (the Hellinger distance): .755

For definitions and further discussion of the χ^2, G^2, and Freeman-Tukey statistics, see Section 1.2. For this example, the root-mean-square distance and the χ^2 statistic produce exactly the same P-values: for the model of homogeneous proportions, displayed in Table 2.3, the root-mean-square distance is directly proportional to the square root of the χ^2 statistic, and hence the root-mean-square distance is a strictly increasing function of χ^2.

Chapter 3

Hardy-Weinberg proportions

Compared with classical statistics for detecting deviations from Hardy-Weinberg equilibrium, the root-mean-square discrepancy can be significantly more powerful. Unlike χ^2, the log–likelihood-ratio, and Fisher's exact tests, which are sensitive to relative discrepancies between genotypic frequencies, the root-mean-square is sensitive to absolute discrepancies. This can increase statistical power, as illustrated in the present chapter both through asymptotic analysis and for benchmark datasets.

3.1 Introduction

Hardy (1908) and Weinberg (1908) independently derived equations corroborating the theory of Mendelian inheritance, proving that in a large population of randomly mating individuals, the proportions of alleles and genotypes at a locus stay unchanged in the absence of specific disturbing influences. Today, Hardy-Weinberg equilibrium (HWE) is a common hypothesis used in scientific domains ranging from botany (for example, by Weising (2005)) to genetic epidemiology (for example, by Sham (2001) and Khoury et al. (2004)). Statistical tests of deviation from HWE are fundamental for validating the hypothesis of HWE. Traditionally, χ^2 or an asymptotically equivalent variant such as the log–likelihood-ratio was used for this assessment. Before computers became readily available, the asymptotic χ^2 approximation for the statistics used in these tests was critical. However, with the now widespread availability of computers, exact tests can be computed effortlessly, opening the door to more powerful goodness-of-fit tests.

The seminal paper of Guo and Thompson (1992) campaigned for an exact test of HWE based on the likelihood function. While their work renewed interest in conditional exact tests for Hardy-Weinberg equilibrium, including contributions by Raymond and Rousset (1995), Diaconis and Sturmfels (1998), and Wigginton et al. (2005), likelihood-based tests have also been subject to criticism, and there is little evidence that such tests are more powerful than other exact tests, such as those based on likelihood *ratios* (as highlighted, for example, by Engels (2009)) or those based on the root-mean-square.

As discussed below, the classical datasets from Guo and Thompson (1992) indicate that goodness-of-fit tests based on the root-mean-square distance can be more powerful than classical tests at detecting deviations from Hardy-Weinberg equilibrium, sometimes by an order of magnitude. In particular, at the 5% significance level, the root-mean-square test rejects the hypothesis of HWE for the dataset in Example 1 of this chapter while the classic tests do not, and at the 1% significance level rejects the hypothesis of HWE for the dataset in Example 2 while the classic tests do not.

These results are not a chance anomaly. Upon further analysis of the datasets, we find that the classical tests, tuned to detect relative discrepancies, can be blind to overwhelmingly large discrepancies among common genotypes that get drowned out by expected statistical fluctuations in rare genotypes. The root-mean-square, on the other hand, detects deviations in absolute discrepancies, and easily detects large discrepancies in common genotypes.

To make these observations precise, we show that, in the asymptotic limit for which the number of draws and number of alleles tend to infinity together, the root-mean-square statistic has asymptotic power one while the classic statistics have asymptotic power zero against a family of alternatives involving one common allele, several rare alleles, and an excess of observed common genotypes. We observe numerically that this asymptotic limit is reached very quickly, on datasets involving no more than 10 alleles and 30 draws. Chapters 1 and 8 discuss the classical asymptotic limit, for which the number of draws tends to infinity but the number of alleles remains fixed; as shown in Chapters 1 and 8, the root-mean-square statistic is often more powerful in this asymptotic limit as well as others.

None of the statistics we consider produces a test that is uniformly more powerful than any other, but while each statistic focuses power on its own class of alternatives, the root-mean-square statistic is more relevant for many deviations of interest in practice. At the very least, the root-mean-square and the classical statistics focus on complementary classes of alternatives, and their combined P-values provide more information than the individual P-values.

The results of our analysis are consistent with the numerous experiments reported in Chapter 1, which exhibit the often greater power of the root-mean-square in detecting meaningful discrepancies for nonuniform model distributions. Chapter 4 provides several examples for which the root-mean-square is more powerful than the classical statistics when testing for homogeneity in contingency tables. We should remark that the root-mean-square can be interpreted as an analog for nominal data of the discrete Kolmogorov-Smirnov statistic; see Chapter 2, for example.

The rest of the present chapter has the following structure: Section 3.2 recalls the set-up and motivation for testing Hardy-Weinberg equilibrium. Section 3.3 defines the relevant test statistics. Section 3.4 compares the performance of these statistics on the classic datasets from Guo and Thompson (1992). Section 3.5 numerically evaluates their statistical power and size for several datasets with prescribed deviations from HWE. Section 3.6 provides an asymptotic analysis of the various statistics to highlight the limited power of the classical statistics relative to the root-mean-square statistic in distinguishing important classes of deviations from Hardy-Weinberg equilibrium. Section 3.7 provides concluding remarks.

3.2 Set-up and motivation

Recall that a "gene" refers to a segment of DNA at a particular location ("locus") on a chromosome. The gene may assume one of several discrete variations, and these variants are known as “alleles.” An individual carries two alleles for each autosomal gene — one allele selected at random from the mother’s pair of alleles, and one allele selected at random from the father’s pair of alleles. These two alleles, considered as an unordered pair, constitute the individual’s genotype. A gene having r alleles $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_r$ has $r(r+1)/2$ possible genotypes; these genotypes are naturally indexed over a lower-triangular array such as those displayed in Figure 3.1.

A population is said to be in Hardy-Weinberg equilibrium (HWE) with respect to a given system of alleles if the proportion of individuals in the population with two distinct alleles is twice the product of the allele proportions, and the proportion of individuals in the population with two copies of the same allele is the square of that allele’s proportion in the population. That is, if $p_{j,k}$ is the proportion of genotype $\{\mathrm{A}_j, \mathrm{A}_k\}$ in the population, and if θ_k is the proportion of allele A_k in the population, then the system is in equilibrium if

$$p_{j,k} = p_{j,k}(\theta_j, \theta_k) = \begin{cases} 2 \cdot \theta_j \cdot \theta_k, & j > k \\ (\theta_k)^2, & j = k \end{cases} \tag{3.1}$$

for all integers j and k such that $1 \le k \le j \le r$. A large population of genotypes that is in Hardy-Weinberg equilibrium will remain in Hardy-Weinberg equilibrium, assuming that mating is random and that there are no disturbing forces — no selection, no mutation, no migration, and so on. Also, Hardy-Weinberg equilibrium is “neutral”: if any assumptions are violated in a particular generation, perturbing the population, then in the next generation the population will be in a new equilibrium (assuming the assumptions are back in effect), with the new equilibrium specified by the new allele proportions. Hardy-Weinberg equilibrium is thus a robust and reliable certificate that a population is not evolving with respect to the gene of interest, and the detection of deviations from Hardy-Weinberg equilibrium is crucial in many genetic analyses.

3.3 Testing for deviations

In practice, genotyping all individuals in a population is not usually feasible, and inference about the distribution of genotypes in the population must be based on random sampling. If a population of genotypes $\{\mathrm{A}_j, \mathrm{A}_k\}$ with underlying genotypic proportions $(p_{j,k})$ is large enough, a random sample of n genotypes X_1, X_2, …, X_n from the population can be regarded as a sequence of independent and identically distributed draws from the multinomial distribution specified by probabilities

$$\mathrm{Prob}\Big\{X_i = \{\mathrm{A}_j, \mathrm{A}_k\}\Big\} = p_{j,k} \tag{3.2}$$

for $i = 1, 2, \ldots, n$ and for all integers j and k such that $1 \le k \le j \le r$. If genotype $\{\mathrm{A}_j, \mathrm{A}_k\}$ is observed $n_{j,k}$ times in the sample of n genotypes, then the number n_j of instances of allele A_j among the $2n$ observed alleles is

$$n_j = \sum_{k=1}^{j} n_{j,k} + \sum_{k=j}^{r} n_{k,j} \tag{3.3}$$

for $j = 1, 2, \ldots, r$. In order to gauge the consistency of the sample counts $(n_{j,k})$ with Hardy-Weinberg equilibrium, we must first specify the $r-1$ free parameters $\theta_1, \theta_2, \ldots, \theta_{r-1}$ corresponding to the underlying allele proportions in the HWE model (3.1). Intuitively, the observed proportions of alleles, $n_1/(2n)$, $n_2/(2n)$, $\ldots$, $n_{r-1}/(2n)$, are the best estimates for $\theta_1, \theta_2, \ldots, \theta_{r-1}$; indeed, it is not difficult to verify that the observed proportions of alleles are the maximum-likelihood estimates for the underlying parameters $\theta_1, \theta_2, \ldots, \theta_{r-1}$ in the family of HWE equilibrium equations (3.1). These parameter specifications give rise to the *model* counts of genotypes under Hardy-Weinberg equilibrium,

$$m_{j,k} = (2 - \delta_{j,k})(n_j \cdot n_k)/(4n) \tag{3.4}$$

for all integers j and k such that $1 \le k \le j \le r$, where $\delta_{j,k}$ is the Kronecker delta,

$$\delta_{j,k} = \begin{cases} 0, & j \neq k \\ 1, & j = k \end{cases} \tag{3.5}$$

for $j, k = 1, 2, \ldots, r$.

The observed counts of genotypes $(n_{j,k})$ in a random sample from a population in Hardy-Weinberg equilibrium usually do not deviate too much from their associated model counts $(m_{j,k})$. Suitable statistics — such as the P-values discussed in the coming subsection — help distinguish inevitable fluctuations due to finite sample sizes and finite populations from model mismatch. Without additional prior information, a goodness-of-fit test serves as an omnibus litmus test to gauge consistency of the data with the Hardy-Weinberg equilibrium model. Ideally, the goodness-of-fit test should be sensitive to the full range of interesting alternative distributions; more realistically, several different goodness-of-fit tests can be used jointly, each sensitive to its own class of alternatives. If such an essentially nonparametric test indicates deviation from equilibrium, different parametric tests can then be used to elucidate particular effects of the deviation such as directions of disequilibrium or level of inbreeding. Several parametric Bayesian methods can be appropriate, and we refer the reader to the discussions of Chen and Thomson (1999), Shoemaker et al. (1998), Ayres and Balding (1998), Lauretto et al. (2009), Li and Graubard (2009), and Consonni et al. (2011). In this chapter, we will focus on nonparametric (or nearly nonparametric) tests of fit, but we emphasize that goodness-of-fit tests should be combined with Bayesian approaches and other types of evidence for and against the equilibrium hypothesis.

3.3.1 Goodness-of-fit testing

A goodness-of-fit test compares the model and empirical distributions using one of many possible divergences. Three classic measures of discrepancy, all special cases of Cressie-Read power-divergences, are χ^2,

$$\chi^2 = \sum_{1 \le k \le j \le r} \frac{\left(n_{j,k} - m_{j,k}\right)^2}{m_{j,k}}, \tag{3.6}$$

the log–likelihood-ratio

$$g^2 = 2 \sum_{1 \le k \le j \le r} n_{j,k} \ln\left(\frac{n_{j,k}}{m_{j,k}}\right), \tag{3.7}$$

and the Hellinger/Freeman-Tukey distance

$$h^2 = 4 \sum_{1 \le k \le j \le r} \left(\sqrt{n_{j,k}} - \sqrt{m_{j,k}}\right)^2. \tag{3.8}$$

The end result of a goodness-of-fit test is the "P-value," the probability of observing a discrepancy between model and sample proportions of genotypes at least as extreme as the measured discrepancy, under the null hypothesis of i.i.d. draws from the model. If a goodness-of-fit test returns a sufficiently small P-value — .01 or .001, for example — then we can be highly confident that the model assumptions do not hold. A more powerful measure of discrepancy for a given data set will produce a smaller P-value if the null hypothesis is not consistent with the data. We remark that there are subtleties involved with the definition and interpretation of P-values, as discussed, for example, in the appendix.

In the present chapter, we consider two types of commonly-used P-values, the "plain" P-value and the "fully conditional" P-value. Other possibilities include Bayesian P-values, as discussed by Gelman (2003), and other formulations.

To compute the plain P-value, we repeatedly simulate n i.i.d. draws from the model multinomial distribution $(m_{j,k}/n)$. For each simulation, say the ith, we compute the genotypic counts $N_{j,k}^{(i)}$, the allelic counts $N_j^{(i)} = \left(\sum_{k=1}^{j} N_{j,k}^{(i)} + \sum_{k=j}^{r} N_{k,j}^{(i)}\right)$, the allelic proportions $\Theta_j^{(i)} = N_j^{(i)}/(2n)$, and the equilibrium model counts associated with the simulated sample, $M_{j,k}^{(i)} = (2 - \delta_{j,k}) N_j^{(i)} N_k^{(i)}/(4n)$. The plain P-value is the fraction of times the discrepancy between the simulated counts $(N_{j,k}^{(i)})$ and their model counts $(M_{j,k}^{(i)})$ is at least as large as the measured discrepancy between the observed counts $(n_{j,k})$ and their model counts $(m_{j,k})$.

The fully conditional P-value imposes additional restrictions on the probability space associated with the null hypothesis. In contrast with the set-up for plain P-values, the observed counts of alleles, n_1, n_2, $\ldots$, n_r, are treated as known quantities in the model when calculating the fully conditional P-value, remaining fixed upon hypothetical repetition of the experiment. More specifically, we repeatedly simulate n i.i.d. draws from the hypergeometric distribution that results from conditioning the multinomial model distribution $(m_{j,k}/n)$ on the observed allele counts, $N_1 = n_1$, $N_2 = n_2$, $\ldots$, $N_r = n_r$. Guo and Thompson (1992) provided an efficient means for performing such a simulation: apply a random permutation to the sequence

$$\Bigl(\underbrace{\overbrace{\mathrm{A}_1, \mathrm{A}_1, \ldots, \mathrm{A}_1}^{n_1}, \overbrace{\mathrm{A}_2, \mathrm{A}_2, \ldots, \mathrm{A}_2}^{n_2}, \ldots, \overbrace{\mathrm{A}_r, \mathrm{A}_r, \ldots, \mathrm{A}_r}^{n_r}}_{2n}\Bigr) \tag{3.9}$$

and identify successive pairs $\{\mathrm{A}_{2j}, \mathrm{A}_{2j+1}\}$. The fully conditional P-value is the fraction of times the discrepancy between the simulated counts $(N_{j,k}^{(i)})$ and the model counts $(m_{j,k})$ is at least as large as the measured discrepancy.

Section 3.8 below provides pseudocode for calculating plain and fully conditional P-values.

Remark 3.3.1. Note that P-values computed by repeated Monte-Carlo simulation are in fact exact: given any specified precision ε, Hoeffding's inequality guarantees with 99.9% confidence that the P-value obtained using ℓ simulations will be within ε of the P-value P

obtained using infinitely many simulations, as long as $\ell \geq 4/\varepsilon^2$. In all our experiments, we used $\ell = 16{,}000{,}000$ Monte-Carlo simulations, so that the reported three digits of precision in our P-values are correct with at least 99.9% confidence.

Remark 3.3.2. The parameters θ_1, θ_2, …, θ_r in the family of HWE distributions (3.1) are known as nuisance parameters. The fully conditional test we describe is a variant of the conditional test proposed by Fisher (1925) and studied by Mehta and Patel (1983) for dealing with nuisance parameters in the context of contingency-table analysis, and amounts to conditioning on a minimally sufficient statistic for estimating the nuisance parameters. Conditioning in the context of HWE testing dates back to the works of Levene (1949) and Haldane (1954), but was not considered feasible for large data sets until Guo and Thompson (1992) derived the aforementioned method for efficiently simulating draws from the conditional distribution. Note that while conditioning on the counts of alleles in the observed population does effectively eliminate the nuisance parameters from the null hypothesis, it also imposes additional assumptions on the experiment that are not necessarily reflective of reality. In small samples drawn from a large population, the allele counts are not known a priori and estimates of the allele counts can change upon repetition of the experiment. The plain P-value takes this into account, unlike the fully conditional P-value; for details, see the appendix.

3.3.2 The negative log-likelihood statistic

A popular alternative to testing Hardy-Weinberg equilibrium with the power-divergence discrepancies χ^2, g^2, and h^2 is to use a discrepancy based directly on the likelihood function for the multinomial distribution,

$$\begin{aligned}\mathcal{L} &= \operatorname{Prob}\Big\{N_{1,1} = n_{1,1},\ N_{2,1} = n_{2,1},\ \ldots, N_{r,r} = n_{r,r}\Big\} \\ &= \frac{n!}{n_{1,1}! \cdot n_{2,1}! \cdots n_{r,r}! \cdot n^n} m_{1,1}^{n_{1,1}} \cdot m_{2,1}^{n_{2,1}} \cdots m_{r,r}^{n_{r,r}}. \qquad (3.10)\end{aligned}$$

Because the likelihood has an excessively large dynamic range, the negative of the logarithm of the likelihood, $L = -\ln(\mathcal{L})$, is instead used as the test statistic. Because the logarithm is nondecreasing over its domain, the negative likelihood function and negative log-likelihood function produce the same P-value. The negative log-likelihood function looks similar to the log–likelihood-ratio function g^2 defined in (3.7), but there is an important distinction: the log–likelihood-ratio, which sums the logarithms of *ratios* between observed and expected counts, is a proper statistical divergence. The negative log-likelihood function is not a divergence, and this results in several undesirable properties that have led many, including Gibbons and Pratt (1975), Radlow and Alf (1975), and Engels (2009), to criticize its use.

The negative log-likelihood function does have something in common with the power-divergence discrepancies: under the null hypothesis, the negative log-likelihood statistic L and the power-divergence statistics X^2, G^2, and H^2 are all asymptotically equivalent, with the latter three converging in distribution to a χ^2 distribution with $r(r-1)/2$ degrees of freedom as the number of draws n tends to infinity and the number of alleles remains fixed; Brownlee (1965) (among others) provides a proof. Using a statistic with an easily calculated (or tabulated) asymptotic approximation was necessary before computers became widely available. Now, however, the exact (non-asymptotic) P-values for these statistics or any other measure of discrepancy are easy to compute via Monte-Carlo simulations.

3.3.3 The root-mean-square statistic

A natural measure of discrepancy for goodness-of-fit testing which has not received as much attention in the literature is the root-mean-square distance,

$$f = \sqrt{\frac{2}{r(r+1)} \sum_{1 \le k \le j \le r} (n_{j,k} - m_{j,k})^2}. \tag{3.11}$$

Please note that the root-mean-square f is closely related to the Frobenius distance of Chapter 4. The square of the root-mean-square distance is proportional to χ^2 from (3.6) when the model distribution is uniform, but takes on a very different character when the model distribution diverges from uniformity. In practice, multiallelic distributions of genotypes are often very nonuniform, with a few common alleles and many rare alleles. In contrast to the classical statistics, the asymptotic distribution for the root-mean-square statistic F in the limit of infinitely many draws, while completely well-defined and efficiently calculated on a modern computer, depends on the model distribution (see Chapters 6 and 7). This has likely contributed to its underrepresentation in the literature, as much of the classical statistical methodology was canonized before computers became readily accessible. Using the pseudocode provided in Section 3.8, P-values for the root-mean-square statistic are now just as easy to compute as P-values for any of the classical statistics.

3.4 Numerical results

In this section, we compare the performances of the root-mean-square and the classical statistics for detecting deviations from Hardy-Weinberg equilibrium. We evaluate the various statistics on the three benchmark datasets from Guo and Thompson (1992). Figure 3.1 displays the three datasets (Examples 1, 2, and 3) as lower-triangular arrays of counts. For each array, the boldface entry in each cell corresponds to the number $n_{j,k}$ of observed counts of genotype $\{\mathrm{A}_j, \mathrm{A}_k\}$ in the sample, and the second entry in each cell corresponds to the expected number $m_{j,k}$ of counts under HWE.

For each example, and for each of the five statistics X^2, G^2, H^2, L, and F, we calculate both the plain and fully conditional P-values using 16,000,000 Monte-Carlo simulations for each calculation. Recall that a small P-value P lets us infer, with $100(1-P)\%$ confidence, that the draws are not i.i.d. or that the draws are inconsistent with the HWE model.

The results of the analyses of Examples 1, 2, and 3, displayed in Table 3.1, suggest that for both plain and fully conditional exact tests of goodness-of-fit, the root-mean-square can be significantly more powerful than the classic statistics in detecting deviations. In particular, the root-mean-square test rejects the null hypothesis of HWE at the 5% significance level in Example 1 while the other statistics do not, and rejects the null hypothesis of HWE at the 1% significance level in Example 2 while the other statistics do not.

Figure 3.2 contains boxplots for relative root-mean-square discrepancies and relative χ^2 discrepancies simulated under the plain Hardy-Weinberg equilibrium null hypothesis corresponding to the dataset from Example 2. Each boxplot displays the median, upper and lower quartiles, and whiskers reaching from the 1st to 99th percentiles for the observed and model proportions. The boxplots are for simulated data, whereas the large open circles indicate the observed data. As seen in the two boxplots, the normalization by expected proportion in the summands of χ^2 from (3.6) is reflected in the larger contribution of relative discrepancies to the reported P-values; in contrast, the equal weighting of the summands of the root-mean-square from (3.11) is reflected in the larger contribution of absolute discrepancies to the reported root-mean-square P-values. In particular, both the χ^2 and root-mean-square tests report a statistically significant deviation in the 5th-largest index, corresponding to the 982 observed counts versus 1057.6 expected counts of genotype $\{\mathrm{A}_4, \mathrm{A}_1\}$ in Example 2. However, the P-value reported by the root-mean-square test is an order of magnitude smaller than the P-value reported by the χ^2 test. In the χ^2 summation, the expected relative deviations in the rare genotypes mask the statistically significant deviation in the 5th-largest index.

The P-values for the dataset from Example 1 admit a similar explanation.

In contrast, the small P-value for the χ^2 test in Example 3 is due to the single observed count of genotype $\{\mathrm{A}_6, \mathrm{A}_6\}$. Indeed, by removing this single count from the dataset and re-running the χ^2 goodness-of-fit test on the remaining $n = 29$ draws, the χ^2 statistic X^2 returns a P-value of .207 ± .001 (meaning that χ^2 no longer finds the discrepancy to be significant).

3.5 Numerical power studies

In this section, we compare the power and size for the statistics X^2, G^2, H^2, L, and F in detecting deviations from HWE, considering populations with increased homozygosity (as due to inbreeding), populations with increased heterozygosity, and populations of genotypes undergoing selection (Chen and Thomson, 1999; Ayres and Balding, 1998; Lauretto et al., 2009). (Recall that power is one minus the probability of a false negative, i.e., of an error of type II; size is the probability of a false positive, i.e., of an error of type I.) The results in Table 3.2 support the assertion that the root-mean-square and the classic statistics focus their power on complementary classes of alternatives. We consider four parameter specifications:

Alternative 1: $r = 10$, $n = 50$, and $\theta_1 = \theta_2 = 1/3$, and $\theta_j = 1/24$ for $3 \le j \le 10$

Alternative 2: $r = 10$, $n = 100$, and $\theta_1 = \theta_2 = 1/3$, and $\theta_j = 1/24$ for $3 \le j \le 10$

Alternative 3: $r = 10$, $n = 200$, and $\theta_1 = \theta_2 = 1/3$, and $\theta_j = 1/24$ for $3 \le j \le 10$

Alternative 4: $r = 20$, $n = 200$, and $\theta_j = a/j$ for $1 \le j \le 20$, where $a = (\sum_{j=1}^{20} 1/j)^{-1}$

3.5.1 Deviations due to selection

When there is selection for or against a particular allele or genotype in the population, the result is an excess or deficiency of genotypes carrying the particular allele or pair of alleles compared to what would be expected under HWE. To account for selection, one introduces fitness parameters $w_{j,k} > 0$ into the HWE equations:

$$p_{j,k} = \begin{cases} 2w_{j,k} \cdot \theta_j \cdot \theta_k / v, & 1 \le k < j \le r \\ w_{k,k} \cdot (\theta_k)^2 / v, & 1 \le k = j \le r, \end{cases} \tag{3.12}$$

where v is a normalization constant. We consider the scenario where the common allele A_1 is undergoing selection, so that genotypes carrying allele A_1 have higher fitness in the population:

$$w_{j,k} = \begin{cases} 3/2, & k = 1 \\ 1, & k > 1. \end{cases} \tag{3.13}$$

Table 3.2 lists the power and size of the various statistical tests in detecting deviations from HWE due to selection for common alleles. In all examples, the root-mean-square statistic appears to be more powerful than the classic statistics while maintaining the correct size. We will provide theoretical justification for these observations through an asymptotic analysis in Section 3.6.

3.5.2 Deviations due to inbreeding

We now consider genotypic distributions parameterized by an inbreeding coefficient, c, which describes how likely members of the population with similar genetic make-up are to mate with each other:

$$p_{j,k} = \begin{cases} 2(1-c) \cdot \theta_j \cdot \theta_k, & 1 \le k < j \le r \\ (\theta_k)^2 + c \cdot \theta_k \cdot (1-\theta_k), & 1 \le k = j \le r. \end{cases} \tag{3.14}$$

Hardy-Weinberg equilibrium corresponds to $c = 0$. A negative value for c corresponds to a deficiency of homozygotes, while a positive value for c corresponds to an excess of homozygotes. Table 3.2 displays the power of the various tests against alternatives of the form (3.14), with $c = +1/10$. The root-mean-square statistic appears to be less powerful than the classic statistics in detecting deviations due to inbreeding. Moreover, it is often desired to estimate the inbreeding coefficient c, for example because of its role in quantifying the behavior of marker-trait association tests in non-HWE populations; as shown, for example, by Rohlfs and Weir (2008), the χ^2 test statistic is equal to $n\hat{c}$, where $\hat{c}$ is the maximum likelihood estimator.

3.6 An asymptotic power analysis

In this section, we prove that the root-mean-square has asymptotic power one while the classical statistics have asymptotic power zero for a representative family of datasets. As a model for the setting where the number of draws and number of genotypes are of the same magnitude, we consider a limit in which the number of alleles and number of draws tend to infinity together. Note that the asymptotic χ^2 approximation of the classical statistics is not valid in this limit.

We consider a gene having $r + 1$ alleles, with one common allele and r rare alleles. The Common Allele dataset involves $n = 3r$ observed genotypes, distributed as indicated in Table 3.3.

	A_1	A_2	A_3	A_4
A_1	**0** .6722			
A_2	**3** 3.667	**1** 5		
A_3	**5** 3.667	**18** 10	**1** 5	
A_4	**3** 2.322	**7** 6.333	**5** 6.333	**2** 2.006

(a) **Example 1:** $n = 45$.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
A_1	**1236** 1206.9								
A_2	**120** 121.67	**3** 3.0662							
A_3	**18** 17.926	**0** .90352	**0** .06656						
A_4	**982** 1057.6	**55** 53.308	**7** 7.8541	**249** 231.70					
A_5	**32** 28.605	**1** 1.4418	**0** .21243	**12** 12.533	**0** .16949				
A_6	**2582** 2556.2	**132** 128.84	**20** 18.982	**1162** 1120.0	**29** 30.291	**1312** 1353.4			
A_7	**6** 5.3396	**0** .26913	**0** .03965	**4** 2.3395	**0** .06328	**4** 5.6543	**0** .00591		
A_8	**2** .76281	**0** .03845	**0** .00566	**0** .33422	**0** .00904	**0** .80776	**0** .00169	**0** .00012	
A_9	**115** 127.01	**5** 6.4015	**2** .94317	**53** 55.647	**1** 1.5051	**149** 134.49	**0** .28094	**0** .04014	**4** 3.3412

(b) **Example 2:** $n = 8297$.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	**3** 1.875							
A_2	**4** 3.5	**2** 1.633						
A_3	**2** 2.75	**2** 2.567	**2** 1.01					
A_4	**3** 3.0	**3** 2.8	**2** 2.2	**1** 1.2				
A_5	**0** .5	**1** .467	**0** .367	**0** .4	**0** .033			
A_6	**0** .5	**0** .467	**0** .367	**0** .4	**0** .067	**1** .033		
A_7	**0** .25	**0** .233	**1** .183	**0** .2	**0** .033	**0** .033	**0** .0083	
A_8	**0** .75	**0** .7	**0** .55	**2** .6	**1** .1	**0** .1	**0** .050	**0** .075

(c) **Example 3:** $n = 30$.

Figure 3.1: The three datasets from Guo and Thompson (1992). Observed counts are bold; model counts are given below the observed counts.

Table 3.1: P-values with 99.9% confidence intervals for X^2, the log–likelihood-ratio statistic G^2, the Hellinger distance H^2, the negative log-likelihood statistic L, and the root-mean-square F, for the observed genotypic counts in Examples 1, 2, and 3 to be consistent with the Hardy-Weinberg equilibrium model.

Example 1

statistic	plain P-value	fully conditional P-value
X^2	.693 ± .001	.709 ± .001
G^2	.600 ± .001	.630 ± .001
H^2	.562 ± .001	.602 ± .001
L	.648 ± .001	.714 ± .001
F	*.039* ± .001	*.039* ± .001

Example 2

statistic	plain P-value	fully conditional P-value
X^2	.020 ± .001	.020 ± .001
G^2	.013 ± .001	.013 ± .001
H^2	.027 ± .001	.025 ± .001
L	.016 ± .001	.018 ± .001
F	*.002* ± .001	*.002* ± .001

Example 3

statistic	plain P-value	fully conditional P-value
X^2	*.015* ± .001	*.026* ± .001
G^2	.181 ± .001	.276 ± .001
H^2	.307 ± .001	.449 ± .001
L	.155 ± .001	.207 ± .001
F	.885 ± .001	.917 ± .001

Table 3.2: Statistical power and size of the various tests of HWE against deviations due to selection, i.e., deviations of the form (3.12) with parameters as specified in Alternatives 1–4 and fitness parameters (3.13), and against deviations due to inbreeding, i.e., deviations of the form (3.14) with parameters as specified in Alternatives 1–4 and inbreeding parameter $c = 1/10$. Power and size are at the 5% significance level, computed using 5000 simulations from the alternative distribution and expected distribution, respectively, with 5000 Monte-Carlo trials per simulation.

Deviations due to selection for common allele — plain P-value								
	Alternative 1		Alternative 2		Alternative 3		Alternative 4	
statistic	power	size	power	size	power	size	power	size
X^2	.06	.06	.04	.05	.04	.04	<.01	.06
G^2	.07	.08	.07	.06	.07	.06	.01	.08
H^2	.07	.07	.08	.06	.08	.05	.01	.07
L	.03	.04	.03	.04	.04	.04	<.01	.03
F	*.09*	.05	*.13*	.05	*.19*	.05	*.23*	.05

Deviations due to common-allele selection — fully conditional P-value								
	Alternative 1		Alternative 2		Alternative 3		Alternative 4	
statistic	power	size	power	size	power	size	power	size
X^2	.04	.05	.03	.04	.05	.06	.03	.05
G^2	.05	.05	.04	.04	.06	.06	.04	.05
H^2	.04	.05	.05	.05	.07	.06	.04	.05
L	.04	.05	.03	.04	.03	.06	.02	.05
F	*.09*	.05	*.11*	.05	*.13*	.06	*.15*	.05

Deviations due to inbreeding — plain P-value								
	Alternative 1		Alternative 2		Alternative 3		Alternative 4	
statistic	power	size	power	size	power	size	power	size
X^2	.20	.06	.34	.05	.60	.04	.64	.06
G^2	*.25*	.08	.29	.06	.48	.06	.64	.08
H^2	.19	.07	.18	.06	.28	.05	.42	.07
L	.23	.04	*.39*	.04	*.63*	.04	*.70*	.03
F	.11	.05	.16	.05	.26	.05	.29	.05

Deviations due to inbreeding — fully conditional P-value								
	Alternative 1		Alternative 2		Alternative 3		Alternative 4	
statistic	power	size	power	size	power	size	power	size
X^2	.21	.05	.35	.04	.61	.06	.68	.05
G^2	.18	.05	.26	.04	.48	.06	.56	.05
H^2	.14	.05	.16	.05	.30	.06	.36	.05
L	*.25*	.05	*.37*	.04	*.63*	.06	*.74*	.05
F	.12	.05	.15	.05	.27	.06	.32	.05

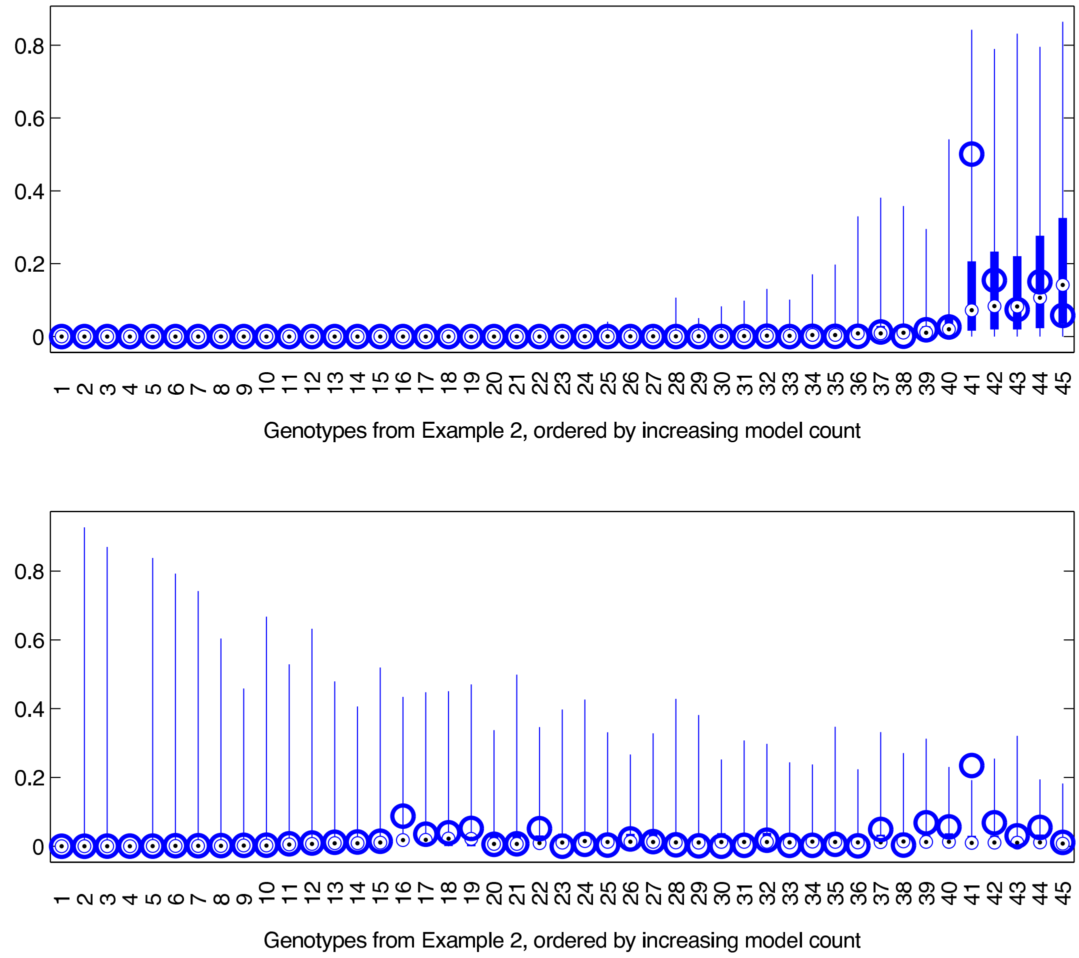


Figure 3.2: Expected versus observed relative root-mean-square discrepancies (top) and expected versus observed relative χ^2-discrepancies (bottom) for the data in Example 2. The root-mean-square is tuned to detect large absolute discrepancies between observed and expected proportions, while χ^2 (and asymptotically equivalent statistics such as the log–likelihood-ratio) is tuned to detect large relative discrepancies between observed and expected proportions.

Table 3.3: Common Allele dataset

$n = 3r$ observed genotypes		
$n_{1,1} = r$ of type $\{A_1, A_1\}$,	$n_{1,1}/n = 1/3$	
$n_{j,1} = 2$ of type $\{A_j, A_1\}$,	$n_{j,1}/n = 2/(3r)$,	$2 \le j \le r+1$
$n_{j,k} = 0$ of type $\{A_j, A_k\}$,	$n_{j,k}/n = 0$,	$2 \le k \le j \le r+1$
$n_1 = 4r$ alleles of type A_1,	$n_1/(2n) = 2/3$	
$n_j = 2$ alleles of type A_j,	$n_j/(2n) = 1/(3r)$,	$2 \le j \le r+1$

The maximum-likelihood model counts for this dataset are

$$\begin{cases} m_{1,1} = 4r/3, & \\ m_{j,1} = 4/3, & 2 \le j \le r+1, \\ m_{j,j} = 1/(3r), & 2 \le j \le r+1, \\ m_{j,k} = 2/(3r), & 2 \le k < j \le r+1. \end{cases} \tag{3.15}$$

To see that the Common Allele dataset becomes increasingly inconsistent with the Hardy-Weinberg model as r increases, observe that under the null hypothesis, we would expect in a sample of $n = 3r$ genotypes to see $r/3$ genotypes containing only rare alleles. The Common Allele dataset however contains *no* genotypes consisting of only rare alleles. In spite of this inconsistency, the plain P-values for each of the four classic statistics X^2, G^2, H^2, and L converge to 1 as $r \to \infty$, indicating zero asymptotic power. In contrast, the P-value for the root-mean-square statistic converges to 0.

Theorem 3.6.1. *In the limit $r \to \infty$, the plain P-values (as computed via Algorithm 1 of Section 3.8) given by X^2, the log–likelihood-ratio G^2, the Hellinger distance H^2, and the negative log-likelihood L for the Common Allele dataset to be consistent with the Hardy-Weinberg equilibrium model all converge to 1, while the plain P-value for the root-mean-square converges to 0.*

The crux of the proof is that, as r increases, relative fluctuations in the rare genotypes simulated under HWE become sufficiently large that the sum of relative discrepancies expected under the null hypothesis exceeds the sum of the observed relative discrepancies. However, the sum of absolute fluctuations expected under the HWE model remains bounded below the sum of the observed absolute discrepancies.

We now sketch a proof of Theorem 3.6.1. In the proof of Theorem 3.6.1, we will use the notation $U(r) \gtrsim V(r)$ to indicate that there exists some positive real number c such that $U(r) \ge c \cdot V(r)$ with probability approaching 1 as $r \to \infty$. We use the notation $U(r) \lesssim V(r)$ accordingly. We write $U(r) \sim V(r)$ to mean that $U(r)/V(r) \to 1$ as $r \to \infty$. We write $U(r) \to v$ to mean that $U(r)$ converges in distribution to a unit mass concentrated at the point v, as $r \to \infty$.

Sketch of proof of Theorem 3.6.1. We will be using the notation for computing plain P-values in Algorithm 1 of Section 3.8, along with the Common Allele dataset in Table 3.3 and its maximum-likelihood HWE model counts (3.15). We will refer to A_1 as the "common" allele and to $\{\mathrm{A}_1, \mathrm{A}_1\}$ as the common genotype; we will refer to the remaining r alleles as "rare," to genotypes of the form $\{\mathrm{A}_j, \mathrm{A}_1\}$, $2 \le j \le r+1$, as "rare observed" genotypes, and to genotypes of the form $\{\mathrm{A}_j, \mathrm{A}_k\}$, $2 \le k \le j \le r+1$, as "unobserved" genotypes.

1. Since the model proportion $\theta_1 = 2/3$ remains constant as r increases, but the number of draws $n = 3r$ tends to ∞, the law of large numbers implies $\Theta_1 \to \theta_1 = 2/3$, and so $M_{1,1}/n \to m_{1,1}/n = 4/9$ and $\sum_{j=2}^{r+1} \Theta_j = 1 - \Theta_1 \to 1/3$. Thus, asymptotically, 2/3 of the alleles and 4/9 of the genotypes simulated from the model will be common.

2. The law of large numbers similarly yields that $\sum_{j=2}^{r+1} M_{j,1}/n \to \sum_{j=2}^{r+1} m_{j,1}/n = 4/9$ and $\sum_{2\le k\le j\le r+1} M_{j,k}/n \to \sum_{2\le k\le j\le r+1} m_{j,k}/n = 1/9$. Thus, asymptotically, 4/9 of the draws simulated from the model will be rare observed genotypes, while 1/9 of the simulated draws will be unobserved genotypes.

3. With probability approaching 1 as $r \to \infty$, every one of the roughly $n/9 = r/3$ simulated draws from the pool of $r(r+1)/2$ types of unobserved genotypes will have a different genotype from the others. At this point, roughly $r/3$ of the simulated proportions $N_{j,k}/n,\ 2 \le k \le j \le r+1$, will equal $1/(3r)$, while the others will be 0.

4. The coupon collector's problem (described, for example, by Motwani and Raghavan (1995)) implies that with probability approaching 1 as $r \to \infty$, among the roughly $2r$ simulated draws from the pool of r rare alleles, no rare allele will be drawn more than $\log(r)$ times (fixing the base of the logarithm at any real number greater than 1 that does not depend on r), and at least $3r/4$ among the r rare alleles will be drawn at least twice.

In particular, the last point above implies that, with probability approaching 1 as $r \to \infty$, all of the simulated rare proportions $\Theta_j = \Theta_j(r),\ 2 \le j \le r+1$, will satisfy

$$\Theta_j(r) \le \log(r)/r, \tag{3.16}$$

and, for at least $3r/4$ among the r simulated rare proportions,

$$1/(3r) \le \Theta_j(r) \le \log(r)/r. \tag{3.17}$$

1. *The P-value for the root-mean-square goes to 0 when $r \to \infty$:* The measured sum-square discrepancy $\tilde{f}^2 = \frac{(r+1)(r+2)}{2n^2} f^2$ between the observed proportions $n_{j,k}/n$ and the model proportions $m_{j,k}/n$ is

$$\begin{aligned} \tilde{f}^2 &= \left(\frac{n_{1,1}}{n} - \frac{m_{1,1}}{n}\right)^2 + \sum_{j=2}^{r+1} \left(\frac{n_{j,1}}{n} - \frac{m_{j,1}}{n}\right)^2 + \sum_{2 \le k \le j \le r+1} \left(\frac{m_{j,k}}{n}\right)^2 \\ &= \left(\frac{1}{9}\right)^2 + \frac{4}{81r} + \frac{1}{81r^3} + \frac{2(r-1)}{81r^3}. \end{aligned} \tag{3.18}$$

As $r \to \infty$,

$$\tilde{f} \to \frac{1}{9}. \tag{3.19}$$

If we instead consider the sum-square statistic $\tilde{F}^2 = \frac{(r+1)(r+2)}{2n^2} F^2$ resulting from drawing $n = 3r$ genotypes i.i.d. from the model distribution (3.15), points 1, 3, and 4 above together with (3.16) give

$$\begin{aligned} \tilde{F}^2 &\lesssim \frac{(N_{1,1} - 4r/3)^2}{9r^2} + \sum_{j=2}^{r+1} \left(\frac{4\log r}{3r}\right)^2 \\ &+ \sum_{2 \le k \le j \le r+1 : N_{j,k}=1} \left(\frac{1}{3r}\right)^2 + \sum_{2 \le k \le j \le r+1 : N_{j,k}=0} \left(\frac{\log r}{r}\right)^4 \\ &\overset{d}{\sim} \frac{Z^2}{27r/4} + \frac{16(\log r)^2}{9r} + \frac{r}{3}\frac{1}{9r^2} + \left(\frac{r(r+1)}{2} - \frac{r}{3}\right)\left(\frac{\log r}{r}\right)^4, \end{aligned} \tag{3.20}$$

where (on account of the central limit theorem) $Z = (N_{1,1} - 4r/3)/\sqrt{4r/3}$ converges in distribution to a standard normal distribution as $r \to \infty$. Therefore, as $r \to \infty$,

$$\tilde{F} \to 0. \tag{3.21}$$

Combining (3.19) and (3.21) shows that the P-value for the root-mean-square statistic, $P = \text{Prob}\{F \geq f\} = \text{Prob}\{\tilde{F} \geq \tilde{f}\}$, goes to 0 as $r \to \infty$.

2. *The P-value for X^2 goes to 1 when $r \to \infty$:* Just as with the measured sum-square discrepancy $\tilde{f}$, the measured normalized χ^2 discrepancy $\tilde{\chi}^2 = \chi^2/n$ converges to some finite positive real number as $r \to \infty$. Alternatively, if we simulate $n = 3r$ genotypes from the model distribution and (following point 3 above) consider only those roughly $r/3$ summands in the normalized χ^2 statistic $\tilde{X}^2 = X^2/n$ corresponding to the unobserved genotypes with one simulated draw, we obtain from (3.16) that

$$\begin{aligned} \tilde{X}^2 &\gtrsim \frac{r}{3} \min_{2 \leq k \leq j \leq r+1 : N_{j,k}=1} \left(\frac{N_{j,k}}{n} - \frac{M_{j,k}}{n} \right)^2 \Big/ \left(\frac{M_{j,k}}{n} \right) \\ &\gtrsim \frac{r}{3} \left(\frac{1}{3r} - \left(\frac{\log r}{r} \right)^2 \right)^2 \Big/ \left(\frac{\log r}{r} \right)^2 . \end{aligned} \tag{3.22}$$

From this it follows that $\tilde{X}^2 \gtrsim \frac{r}{(\log r)^2} \to \infty$, and so the P-value for the χ^2 statistic, $P = \text{Prob}\{X^2 \geq \chi^2\} = \text{Prob}\{\tilde{X}^2 \geq \tilde{\chi}^2\}$, goes to 1 as $r \to \infty$.

3. *The P-values for the log–likelihood-ratio G^2 and negative log-likelihood L go to 1 when $r \to \infty$* by an argument analogous to that used for the χ^2 P-value.

4. *The P-value for the Hellinger statistic H^2 goes to 1 when $r \to \infty$:* We have to be a bit more careful with the analysis of the Hellinger discrepancy $\tilde{h}^2 = h^2/(4n)$. The observed discrepancy is

$$\begin{aligned} \tilde{h}^2 &= \frac{(\sqrt{3}-2)^2}{9} + \sum_{j=2}^{r+1} \left(\sqrt{\frac{2}{3r}} - \sqrt{\frac{4}{9r}} \right)^2 + \sum_{2 \leq k < j \leq r+1} \frac{2}{9r^2} + \sum_{j=2}^{r+1} \frac{1}{9r^2} \\ &= \frac{(\sqrt{3}-2)^2}{9} + \frac{10 - 4\sqrt{6}}{9} + \frac{1}{9} \\ &= .14 \ldots . \end{aligned} \tag{3.23}$$

Alternatively, suppose we simulate $n = 3r$ genotypes from the model distribution and consider r sufficiently large. Every estimated rare allele proportion will be bounded: $\Theta_j \leq \log(r)/r$, as stated in (3.16). Furthermore, by (3.17), at least 3/4 of these proportions will satisfy $\Theta_j \geq 1/(3r)$, ensuring that at least $(3/4)^2 r^2/2 - r$ among the $r(r+1)/2$ simulated proportions for the unobserved genotypes satisfy $M_{j,k}/n \geq 2/(9r^2)$.

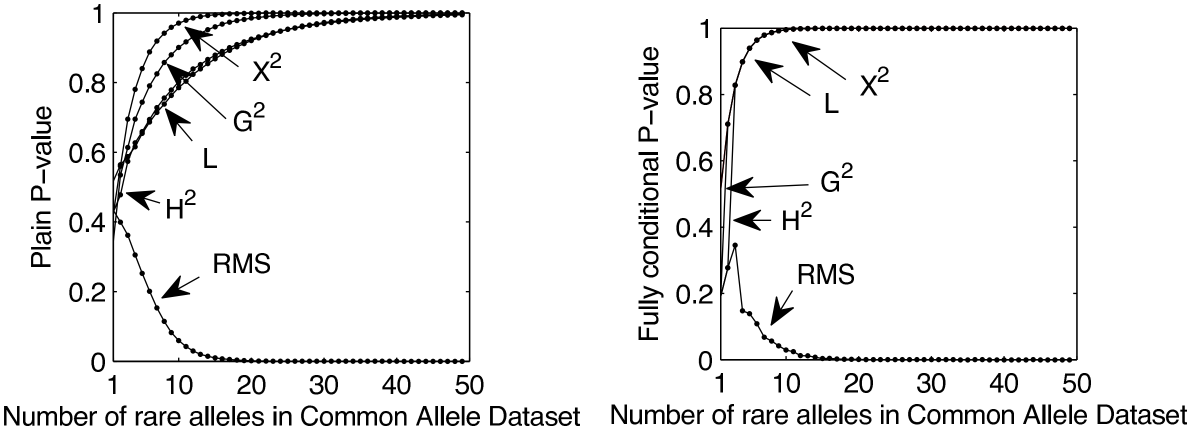


Figure 3.3: P-values (accurate to three digits with 99% confidence) for X^2, the log–likelihood-ratio statistic G^2, the Hellinger statistic H^2, the negative log-likelihood statistic L, and the root-mean-square F, for the observed genotypic counts in the Common Allele dataset to be consistent with the Hardy-Weinberg equilibrium model (3.4), as a function of the number of rare alleles r.

Then, for sufficiently large r,

$$\begin{aligned}
\tilde{H}^2 &\geq \sum_{2\leq j\leq k\leq r+1} \left(\sqrt{N_{j,k}/n} - \sqrt{M_{j,k}/n}\right)^2 \\
&\geq \#\{j,k : N_{j,k} = 1\} \left(\frac{1}{\sqrt{3r}} - \frac{\log(r)}{r}\right)^2 \\
&+ \left(\left(\frac{3}{4}\right)^2 \frac{r^2}{2} - r - \#\{j,k : N_{j,k} = 1\}\right) \left(\frac{2}{9r^2}\right) \\
&\sim \frac{r}{3}\left(\frac{1}{\sqrt{3r}} - \frac{\log(r)}{r}\right)^2 + \left(\left(\frac{3}{4}\right)^2 \frac{r^2}{2} - r - \frac{r}{3}\right) \left(\frac{2}{9r^2}\right) \\
&\to .17\ldots. \qquad (3.24)
\end{aligned}$$

Combining (3.23) and (3.24), we conclude that the P-value for the Hellinger distance, $P = \text{Prob}\{H^2 \geq h^2\} = \text{Prob}\{\tilde{H}^2 \geq \tilde{h}^2\}$, goes to 1 as $r \to \infty$.

Figure 3.3 shows that the convergence of the classic P-values to 1, and of the root-mean-square P-value to 0, occurs very quickly. This convergence is demonstrated for both the plain and fully conditional P-values, even though Theorem 3.6.1 applies directly only to the plain P-values.

To conclude this section, we remark that the particular distribution of draws in the Common Allele dataset was rather arbitrary, and that a similar asymptotic analysis holds for many other datasets. For example, we could have considered a dataset involving two, three, or four common alleles, or one common allele and three fairly common alleles, and so on.

3.7 Conclusion

On two of the three benchmark datasets from Guo and Thompson (1992), the root-mean-square test rejects the hypothesis of Hardy-Weinberg equilibrium at standard significance levels while the classic tests fail to reject the null hypothesis. These numerical results, along with the asymptotic power analysis of Section 3.6, suggest that the root-mean-square statistic can be better than the classic statistics at detecting many deviations of interest in practice. At the very least, the root-mean-square statistic and the classic statistics focus on complementary classes of deviations from Hardy-Weinberg equilibrium, and their combined P-values provide a more informative test than either P-value used on its own.

3.8 Addendum: Pseudocodes for computing exact P-values

Algorithm 1: Computing the plain P-value

Input: observed genotype counts $(n_{j,k})$ and allele counts (n_j), number of Monte-Carlo simulations ℓ, and test statistic D ($D = X^2$, G^2, H^2, $\dots$)
Output: plain P-value associated with the measure of discrepancy $D(n_{j,k}, m_{j,k})$

```
Procedure:
Compute maximum-likelihood model counts $m_{j,k} = (2 - \delta_{j,k})\, n_j\, n_k/(4n)$
Measure the discrepancy $d = D(n_{j,k}, m_{j,k})$
for $i = 1$ to $\ell$
    - Draw $n$ genotypes $X_1^{(i)}, \dots, X_q^{(i)}, \dots, X_n^{(i)}$ i.i.d. from the multinomial model distribution $(m_{j,k}/n)$
    - Aggregate simulated genotype counts $N_{j,k}^{(i)} = \#\{q : X_q^{(i)} = \{A_j, A_k\}\}$
    - Aggregate simulated allele counts $N_j^{(i)} = \big(\sum_{k=1}^{j} N_{j,k}^{(i)} + \sum_{k=j}^{r} N_{k,j}^{(i)}\big)$ and proportions $\Theta_j^{(i)} = N_j^{(i)}/(2n)$
    - Compute maximum-likelihood counts $M_{j,k}^{(i)} = (2 - \delta_{j,k})\, N_j^{(i)} N_k^{(i)}/(4n)$
    - Evaluate simulated discrepancy $D^{(i)} = D(N_{j,k}^{(i)}, M_{j,k}^{(i)})$
endfor
return plain P-value, $P = \#\{i : D^{(i)} \geq d\}/\ell$
```

Algorithm 2: Computing the fully conditional P-value

Input: observed genotype counts $(n_{j,k})$ and allele counts (n_j), number of Monte-Carlo simulations ℓ, and test statistic D ($D = X^2$, G^2, H^2, $\dots$)
Output: fully conditional P-value associated with the measure of discrepancy $D(n_{j,k}, m_{j,k})$

```
Procedure:
Compute maximum-likelihood model counts $m_{j,k} = (2 - \delta_{j,k})\, n_j\, n_k/(4n)$
Measure the discrepancy $d = D(n_{j,k}, m_{j,k})$
for $i = 1$ to $\ell$
    - Apply a random permutation to the sequence of alleles as in (3.9) to obtain $n$ genotypes $X_1^{(i)}, \dots, X_q^{(i)}, \dots, X_n^{(i)}$ with fixed allele counts $n_j$
    - Aggregate simulated genotype counts $N_{j,k}^{(i)} = \#\{q : X_q^{(i)} = \{A_j, A_k\}\}$
    - Evaluate simulated discrepancy $D^{(i)} = D(N_{j,k}^{(i)}, m_{j,k})$
endfor
return fully conditional P-value, $P = \#\{i : D^{(i)} \geq d\}/\ell$
```

Chapter 4

Homogeneity in contingency-tables/crosstabs

The model for homogeneity of proportions in a two-way contingency-table/cross-tabulation is the same as the model of independence, except that the probabilistic process generating the data is viewed as fixing the column totals (but not the row totals). When gauging the consistency of observed data with the assumption of independence, the previous chapters illustrate that the Euclidean/Frobenius/Hilbert-Schmidt distance is often far more powerful than the classical statistics such as χ^2, the log–likelihood-ratio “G^2,” the Freeman-Tukey/Hellinger distance, and other members of the Cressie-Read power-divergence family. This chapter indicates that the Euclidean/Frobenius/Hilbert-Schmidt distance can be more powerful for gauging the consistency of observed data with the assumption of homogeneity, too.

4.1 Recap

The statistical analysis of categorical data is commonly formulated in the framework of contingency-tables/cross-tabulations; Table 4.1 provides a typical two-way example (see, for instance, Chapter 4 of Andersen, 1990, for a comprehensive treatment). A common task is to ascertain whether the given data (displayed in Table 4.1) is consistent up to expected statistical fluctuations with the model for homogeneity of proportions (displayed in Table 4.2). When considering homogeneity, we assume that the probabilistic process generating the given data fixes the column totals (but not the row totals) by construction. Therefore, to gauge whether the given data displayed in Table 4.1 is consistent with the assumed homogeneity displayed in Table 4.2, we perform the following three steps:

1. We generate s sets of draws, with the kth set consisting of $n_{\cdot,k}$ independent and identically distributed draws from the probability distribution $(p_1, p_2, \dots, p_r)$, where $p_1 = n_{1,\cdot}/n$, $p_2 = n_{2,\cdot}/n$, …, $p_r = n_{r,\cdot}/n$. Please note that $p_j = (n_{j,\cdot} \cdot n_{\cdot,k}/n)/n_{\cdot,k}$ for $j = 1$, 2, …, r; these are homogeneous proportions (since p_1, p_2, …, p_r are the same for every column index k).

2. For each of the s sets of draws — say the kth set — we define $N_{j,k}$ to be the number of draws falling in the jth row, for $j = 1$, 2, …, r.

Table 4.1: A typical two-way contingency-table/cross-tab ($n_{j,k}$ is a nonnegative integer with $j = 1, 2, \ldots, r, \quad k = 1, 2, \ldots, s; \quad n_{j,\bullet} = \sum_{k=1}^{s} n_{j,k}$ is a row total with $j = 1, 2, \ldots, r$; $n_{\bullet,k} = \sum_{j=1}^{r} n_{j,k}$ is a column total with $k = 1, 2, \ldots, s$; and $n_{\bullet,\bullet} = \sum_{j=1}^{r} \sum_{k=1}^{s} n_{j,k} = n$ is the grand total)

	1	2	$\cdots$	s	
1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,s}$	$n_{1,\bullet}$
2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,s}$	$n_{2,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$n_{r,1}$	$n_{r,2}$	$\cdots$	$n_{r,s}$	$n_{r,\bullet}$
	$n_{\bullet,1}$	$n_{\bullet,2}$	$\cdots$	$n_{\bullet,s}$	$n_{\bullet,\bullet}$

Table 4.2: The model for homogeneity of proportions ($n_{1,\bullet}$, $n_{2,\bullet}$, $\ldots$, $n_{r,\bullet}$ are the row totals; $n_{\bullet,1}$, $n_{\bullet,2}$, $\ldots$, $n_{\bullet,s}$ are the column totals; and $n_{\bullet,\bullet} = n$ is the grand total)

	1	2	$\cdots$	s	
1	$n_{1,\bullet} \cdot n_{\bullet,1}/n$	$n_{1,\bullet} \cdot n_{\bullet,2}/n$	$\cdots$	$n_{1,\bullet} \cdot n_{\bullet,s}/n$	$n_{1,\bullet}$
2	$n_{2,\bullet} \cdot n_{\bullet,1}/n$	$n_{2,\bullet} \cdot n_{\bullet,2}/n$	$\cdots$	$n_{2,\bullet} \cdot n_{\bullet,s}/n$	$n_{2,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$n_{r,\bullet} \cdot n_{\bullet,1}/n$	$n_{r,\bullet} \cdot n_{\bullet,2}/n$	$\cdots$	$n_{r,\bullet} \cdot n_{\bullet,s}/n$	$n_{r,\bullet}$
	$n_{\bullet,1}$	$n_{\bullet,2}$	$\cdots$	$n_{\bullet,s}$	$n_{\bullet,\bullet}$

3. We calculate the probability P that the discrepancy between the simulated counts $N_{j,k}$ and the model $N_{j,\bullet} \cdot N_{\bullet,k}/n$ is greater than or equal to the discrepancy between the observed counts $n_{j,k}$ and the assumed $n_{j,\bullet} \cdot n_{\bullet,k}/n$. When calculating this probability, we view $N_{j,k}$ and $N_{j,\bullet}$ as random, while viewing all other numbers as fixed. Please note that, by construction, $N_{\bullet,k} = n_{\bullet,k}$ for $k = 1, 2, \ldots, s$.

The number P defined in Step 3 is known as the (exact) P-value. Given the P-value P, we can have $100(1 - P)\%$ confidence that the observed draws are not consistent with assuming the homogeneity displayed in Table 4.2. See the appendix for further discussion of P-values and their interpretation; the appendix details subtleties involved in the definition and interpretation of these P-values, which differ from some other common types of "P-values." We now continue to Section 4.2, which quantifies the discrepancies specifying the P-value P in Step 3 above.

4.2 Definitions of the divergences

The definition above of the P-value P requires a metric for measuring the discrepancies. The canonical choices are χ^2 and the log–likelihood-ratio G^2:

$$\chi^2 = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(n_{j,k} - (n_{j,\bullet} \cdot n_{\bullet,k}/n))^2}{n_{j,\bullet} \cdot n_{\bullet,k}/n} \tag{4.1}$$

$$X^2 = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(N_{j,k} - (N_{j,\bullet} \cdot N_{\bullet,k}/n))^2}{N_{j,\bullet} \cdot N_{\bullet,k}/n} \tag{4.2}$$

$$P_{\chi^2} = \mathrm{Prob}\{X^2 \geq \chi^2\} \tag{4.3}$$

$$g^2 = 2 \sum_{j=1}^{r} \sum_{k=1}^{s} n_{j,k} \cdot \ln\left(\frac{n_{j,k}}{n_{j,\bullet} \cdot n_{\bullet,k}/n}\right) \tag{4.4}$$

$$G^2 = 2 \sum_{j=1}^{r} \sum_{k=1}^{s} N_{j,k} \cdot \ln\left(\frac{N_{j,k}}{N_{j,\bullet} \cdot N_{\bullet,k}/n}\right) \tag{4.5}$$

$$P_{g^2} = \mathrm{Prob}\{G^2 \geq g^2\} \tag{4.6}$$

Other possibilities include the Hellinger (or Freeman-Tukey) distance and the Frobenius (or Hilbert-Schmidt or Euclidean) distance:

$$h^2 = 4 \sum_{j=1}^{r} \sum_{k=1}^{s} (\sqrt{n_{j,k}} - \sqrt{n_{j,\bullet} \cdot n_{\bullet,k}/n})^2 \tag{4.7}$$

$$H^2 = 4 \sum_{j=1}^{r} \sum_{k=1}^{s} (\sqrt{N_{j,k}} - \sqrt{N_{j,\bullet} \cdot N_{\bullet,k}/n})^2 \tag{4.8}$$

$$P_{h^2} = \mathrm{Prob}\{H^2 \geq h^2\} \tag{4.9}$$

$$f^2 = \sum_{j=1}^{r} \sum_{k=1}^{s} (n_{j,k} - (n_{j,\bullet} \cdot n_{\bullet,k}/n))^2 \tag{4.10}$$

$$F^2 = \sum_{j=1}^{r} \sum_{k=1}^{s} (N_{j,k} - (N_{j,\bullet} \cdot N_{\bullet,k}/n))^2 \tag{4.11}$$

$$P_{f^2} = \mathrm{Prob}\{F^2 \geq f^2\} \tag{4.12}$$

When taking probabilities in (4.3), (4.6), (4.9), and (4.12), we view the uppercase X^2, G^2, H^2, and F^2 as random variables, while viewing the lowercase χ^2, g^2, h^2, and f^2 as fixed numbers.

As discussed, for example, by Rao (2002), X^2, G^2, and H^2 all converge to the same distribution in the limit of large numbers of draws — X^2, G^2, and H^2 are the best-known members of the Cressie-Read power-divergence family. F^2 is not a member of the Cressie-Read power-divergence family and does not necessarily converge to the same distribution as X^2, G^2, and H^2. Chapters 1 and 3 illustrate the many advantages of F^2 when neither the row totals nor the column totals are fixed; the present chapter illustrates the advantages when the column totals are fixed. However, F^2 is not uniformly more powerful than the classical statistics. We recommend using both F^2 and a classical statistic such as G^2.

In the sequel, Section 4.3 summarizes an algorithm for computing the P-values defined above. Section 4.4 analyzes several data sets.

4.3 Computation of P-values

The definitions of the P-values in (4.3), (4.6), (4.9), and (4.12) involve the probabilities of certain events. In the present chapter, we compute these probabilities via Monte-Carlo simulations with guaranteed error bounds. Specifically, we conduct a large number ℓ of simulations; in each simulation — say the ith — we perform the following steps (using the data of Table 4.1):

1. We generate s sets of draws, with the kth set consisting of $n_{\bullet,k}$ independent and identically distributed draws from the probability distribution $(p_1, p_2, \dots, p_r)$, where $p_1 = n_{1,\bullet}/n$, $p_2 = n_{2,\bullet}/n$, …, $p_r = n_{r,\bullet}/n$. Please note that $p_j = (n_{j,\bullet} \cdot n_{\bullet,k}/n)/n_{\bullet,k}$ for $j = 1$, 2, …, r; these are homogeneous proportions (since p_1, p_2, …, p_r are the same for every column index k). Furthermore, the underlying distribution of the draws does not depend on i.

2. For each of the s sets of draws — say the kth set — we define $n^{(i)}_{j,k}$ to be the number of draws falling in the jth row, for $j = 1$, 2, …, r.

3. We calculate the discrepancy $f^2_{(i)}$ between the simulated counts $n^{(i)}_{j,k}$ and the model $n^{(i)}_{j,\bullet} \cdot n^{(i)}_{\bullet,k}/n$, that is,
$$f^2_{(i)} = \sum_{j=1}^{r} \sum_{k=1}^{s} (n^{(i)}_{j,k} - (n^{(i)}_{j,\bullet} \cdot n^{(i)}_{\bullet,k}/n))^2. \tag{4.13}$$

An estimate of the P-value P_{f^2} is the fraction of $f^2_{(1)}$, $f^2_{(2)}$, …, $f^2_{(\ell)}$ which are greater than or equal to f^2 defined in (4.10). As discussed in Remark 1.3.2, the standard error of the estimate is $\sqrt{P_{f^2}(1 - P_{f^2})/\ell}$, where ℓ is the number of simulations.

Needless to say, we can compute the P-values for χ^2, g^2, and h^2 via similar procedures, with the same error bounds.

Remark 4.3.1. For all computations reported in the present chapter, we generated random numbers via the C programming language procedure given on page 9 of Marsaglia (2003), implementing the recommended complementary multiply with carry.

4.4 Data analysis

To compare the performance of the various metrics for measuring the discrepancies between observed and simulated data, we analyze several data sets. Using the procedure of Section 4.3, we conduct $\ell = 4{,}000{,}000$ Monte-Carlo simulations per P-value, for each of the examples presented below. The standard error of the resulting estimate for the P-value P is then $\sqrt{P(1-P)}/2000$; see Remark 1.3.2. Before reporting the P-values associated with the data sets, we make two remarks concerning their interpretation:

Remark 4.4.1. A significance test can only indicate that observed data *cannot* be reasonably assumed to have arisen from the model of homogeneous proportions; a significance test cannot prove that the observed data *can* be reasonably assumed to have arisen from the model of homogeneity. Thus, aside from considerations of multiple testing, if any statistic strongly signals that the data cannot be reasonably assumed to have arisen from the model of homogeneity, then we must reject (or at least question) the model — irrespective of any large P-values for other statistics. For instance, if the P-value for the Frobenius distance f^2 is very small, then we should not accept the model of homogeneity, not even if the P-values for χ^2, the log–likelihood-ratio g^2, and the Freeman-Tukey/Hellinger-distance h^2 are large.

Remark 4.4.2. 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 (1992); unlike the log–likelihood-ratio 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.

The 11×2 Table 4.3 displays the data for our first example, which has 22 entries in all. Table 4.4 displays the model of homogeneous proportions for Table 4.3. The P-values for Table 4.3 for the assumption that Table 4.4 gives the correct underlying distribution are

χ^2 (X^2):	.0868
log–likelihood-ratio (G^2):	.0906
Freeman-Tukey/Hellinger (H^2):	.0959
negative log-likelihood:	.0905
Frobenius (F^2):	.00838

Please note that the P-value for the Frobenius distance is over an order of magnitude smaller than the P-values for the classical statistics.

The 7×3 Table 4.7 displays the data for our second example, which has 21 entries in all. Table 4.8 displays the model of homogeneous proportions for Table 4.7. The P-values for Table 4.7 for the assumption that Table 4.8 gives the correct underlying distribution are

χ^2 (X^2):	.145
log–likelihood-ratio (G^2):	.292
Freeman-Tukey/Hellinger (H^2):	.493
negative log-likelihood:	.132
Frobenius (F^2):	.0286

Please note that the P-value for the Frobenius distance is over four times smaller than the P-values for the classical statistics.

The 9×2 Table 4.11 displays the data for our third example, which has 18 entries in all. Table 4.12 displays the model of homogeneous proportions for Table 4.11. The P-values for Table 4.11 for the assumption that Table 4.12 gives the correct underlying distribution are

χ^2 (X^2):	.123
log–likelihood-ratio (G^2):	.138
Freeman-Tukey/Hellinger (H^2):	.157
negative log-likelihood:	.114
Frobenius (F^2):	.0344

Please note that the P-value for the Frobenius distance is over three times smaller than the P-values for the classical statistics.

The 5×3 Table 4.15 displays the data for our final example, which has 15 entries in all. Table 4.16 displays the model of homogeneous proportions for Table 4.15. The P-values for Table 4.15 for the assumption that Table 4.16 gives the correct underlying distribution are

χ^2 (X^2):	.276
log–likelihood-ratio (G^2):	.171
Freeman-Tukey/Hellinger (H^2):	.0794
negative log-likelihood:	.235
Frobenius (F^2):	.199

In this example, none of the statistics produces a very small P-value; the smallest arises from the Freeman-Tukey/Hellinger distance in this case.

Remark 4.4.3. Appropriate binning (or rebinning) to uniformize the frequencies associated with the entries in the contingency-tables/cross-tabulations can mitigate the problem with the classical statistics. Yet rebinning is a black art that is liable to improperly influence the result of a significance test, and the usual data-dependent rebinning calls for Monte-Carlo simulations to calculate P-values accurately anyways. Rebinning always requires careful extra work. A principal advantage of the Frobenius distance is that it does not require any rebinning; indeed, the Frobenius distance is most powerful without any rebinning. Note also that optimally rebinning data such as that displayed in Table 4.3 can be very challenging.

Remark 4.4.4. The Frobenius distance is significantly more powerful than the classical statistics for gauging the consistency of observed data with the assumption of homogeneity in many of the examples of the present chapter. This may or may not be typical of most applications; actually, we suspect that our last example — in which all the statistics perform similarly — is the most representative. Even so, these examples illustrate that there are important circumstances in which the Frobenius distance is much more powerful than the classical statistics.

Table 4.3: Results of polls in June 1983 for Danish parliamentary elections, from Chapter 4 of Andersen (1990)

Party	Poll 1		Poll 2	
A	416	(33.1%)	268	(38.9%)
B	45	(3.6%)	22	(3.2%)
C	338	(26.9%)	160	(23.2%)
E	13	(1.0%)	6	(0.9%)
F	131	(10.4%)	66	(9.6%)
K	18	(1.4%)	10	(1.5%)
M	47	(3.7%)	16	(2.3%)
Q	20	(1.6%)	8	(1.2%)
V	129	(10.3%)	92	(13.4%)
Y	22	(1.8%)	9	(1.3%)
Z	76	(6.1%)	32	(4.6%)
All	1255	(100.0%)	689	(100.0%)

Table 4.4: The model of homogeneous proportions for Table 4.3

Party	Poll 1		Poll 2	
A	441.6	(35.2%)	242.4	(35.2%)
B	43.3	(3.4%)	23.7	(3.4%)
C	321.5	(25.6%)	176.5	(25.6%)
E	12.3	(1.0%)	6.7	(1.0%)
F	127.2	(10.1%)	69.8	(10.1%)
K	18.1	(1.4%)	9.9	(1.4%)
M	40.7	(3.2%)	22.3	(3.2%)
Q	18.1	(1.4%)	9.9	(1.4%)
V	142.7	(11.4%)	78.3	(11.4%)
Y	20.0	(1.6%)	11.0	(1.6%)
Z	69.7	(5.6%)	38.3	(5.6%)
All	1255.0	(100.0%)	689.0	(100.0%)

Table 4.5: Differences between the entries of Table 4.3 and the corresponding entries of Table 4.4

Party	Poll 1	Poll 2
A	−25.6	25.6
B	1.7	−1.7
C	16.5	−16.5
E	0.7	−0.7
F	3.8	−3.8
K	−0.1	0.1
M	6.3	−6.3
Q	1.9	−1.9
V	−13.7	13.7
Y	2.0	−2.0
Z	6.3	−6.3
All	0.0	0.0

Table 4.6: The entries of Table 4.5 divided by the square roots of the corresponding entries of Table 4.4

Party	Poll 1	Poll 2
A	−1.2	1.6
B	0.3	−0.4
C	0.9	−1.2
E	0.2	−0.3
F	0.3	−0.5
K	−0.0	0.0
M	1.0	−1.3
Q	0.5	−0.6
V	−1.1	1.5
Y	0.4	−0.6
Z	0.8	−1.0
All	0.0	0.0

Table 4.7: Reasons for (or absence of) premature termination of the treatment of maniacal patients in three groups from Bowden et al. (1994) (the three groups are those treated with divalproex, those treated with lithium, and those "treated" with a placebo)

Reason	Divalproex		Lithium		Placebo	
Lack of efficacy	21	(30.4%)	12	(33.3%)	38	(51.4%)
Intolerance	4	(5.8%)	4	(11.1%)	2	(2.7%)
Recovered	3	(4.3%)	2	(5.6%)	2	(2.7%)
Noncompliance	1	(1.4%)	1	(2.8%)	3	(4.1%)
Another illness	0	(0.0%)	1	(2.8%)	0	(0.0%)
Administration	4	(5.8%)	2	(5.6%)	2	(2.7%)
Not terminated	36	(52.2%)	14	(38.9%)	27	(36.5%)
All	69	(100.0%)	36	(100.0%)	74	(100.0%)

Table 4.8: The model of homogeneous proportions for Table 4.7

Reason	Divalproex		Lithium		Placebo	
Lack of efficacy	27.4	(39.7%)	14.3	(39.7%)	29.4	(39.7%)
Intolerance	3.9	(5.6%)	2.0	(5.6%)	4.1	(5.6%)
Recovered	2.7	(3.9%)	1.4	(3.9%)	2.9	(3.9%)
Noncompliance	1.9	(2.8%)	1.0	(2.8%)	2.1	(2.8%)
Another illness	0.4	(0.6%)	0.2	(0.6%)	0.4	(0.6%)
Administration	3.1	(4.5%)	1.6	(4.5%)	3.3	(4.5%)
Not terminated	29.7	(43.0%)	15.5	(43.0%)	31.8	(43.0%)
All	69.0	(100.0%)	36.0	(100.0%)	74.0	(100.0%)

Table 4.9: Differences between the entries of Table 4.7 and the corresponding entries of Table 4.8

Reason	Divalproex	Lithium	Placebo
Lack of efficacy	−6.4	−2.3	8.6
Intolerance	0.1	2.0	−2.1
Recovered	0.3	0.6	−0.9
Noncompliance	−0.9	0.0	0.9
Another illness	−0.4	0.8	−0.4
Administration	0.9	0.4	−1.3
Not terminated	6.3	−1.5	−4.8
All	0.0	0.0	0.0

Table 4.10: The entries of Table 4.9 divided by the square roots of the corresponding entries of Table 4.8

Reason	Divalproex	Lithium	Placebo
Lack of efficacy	−1.2	−0.6	1.6
Intolerance	0.1	1.4	−1.0
Recovered	0.2	0.5	−0.5
Noncompliance	−0.7	0.0	0.6
Another illness	−0.6	1.8	−0.6
Administration	0.5	0.3	−0.7
Not terminated	1.2	−0.4	−0.9
All	0.0	0.0	0.0

Table 4.11: Results for the 2012 Republican U.S. presidential nomination, from a CBS News poll of November 6–10, 2011 (released November 11, 2011) and from a Pew Research Center poll of November 9–11, 2011 (released November 17, 2011), as reconstructed from percentages rounded to the nearest whole numbers (the original counts were not reported) for Republican primary voters

Candidate	CBS		Pew	
Michele Bachmann	15	(4.6%)	21	(5.1%)
Herman Cain	69	(21.2%)	103	(25.0%)
Newt Gingrich	57	(17.5%)	66	(16.0%)
Jon Huntsman	4	(1.2%)	4	(1.0%)
Ron Paul	19	(5.8%)	33	(8.0%)
Rick Perry	31	(9.5%)	37	(9.0%)
Mitt Romney	57	(17.5%)	91	(22.1%)
Rick Santorum	8	(2.5%)	8	(1.9%)
Do not know	65	(20.0%)	49	(11.9%)
All	325	(100.0%)	412	(100.0%)

Table 4.12: The model of homogeneous proportions for Table 4.11

Candidate	CBS		Pew	
Michele Bachmann	15.9	(4.9%)	20.1	(4.9%)
Herman Cain	75.8	(23.3%)	96.2	(23.3%)
Newt Gingrich	54.2	(16.7%)	68.8	(16.7%)
Jon Huntsman	3.5	(1.1%)	4.5	(1.1%)
Ron Paul	22.9	(7.1%)	29.1	(7.1%)
Rick Perry	30.0	(9.2%)	38.0	(9.2%)
Mitt Romney	65.3	(20.1%)	82.7	(20.1%)
Rick Santorum	7.1	(2.2%)	8.9	(2.2%)
Do not know	50.3	(15.5%)	63.7	(15.5%)
All	325.0	(100.0%)	412.0	(100.0%)

Table 4.13: Differences between the entries of Table 4.11 and the corresponding entries of Table 4.12

Candidate	CBS	Pew
Michele Bachmann	−0.9	0.9
Herman Cain	−6.8	6.8
Newt Gingrich	2.8	−2.8
Jon Huntsman	0.5	−0.5
Ron Paul	−3.9	3.9
Rick Perry	1.0	−1.0
Mitt Romney	−8.3	8.3
Rick Santorum	0.9	−0.9
Do not know	14.7	−14.7
All	0.0	0.0

Table 4.14: The entries of Table 4.13 divided by the square roots of the corresponding entries of Table 4.12

Candidate	CBS	Pew
Michele Bachmann	−0.2	0.2
Herman Cain	−0.8	0.7
Newt Gingrich	0.4	−0.3
Jon Huntsman	0.3	−0.2
Ron Paul	−0.8	0.7
Rick Perry	0.2	−0.2
Mitt Romney	−1.0	0.9
Rick Santorum	0.4	−0.3
Do not know	2.1	−1.8
All	0.0	0.0

Table 4.15: Reactions to prior treatment with lithium (when treated before with lithium) of maniacal patients in three groups from Bowden et al. (1994) (the three groups are those treated with divalproex, those treated with lithium, and those "treated" with a placebo)

Reaction	Divalproex		Lithium		Placebo	
Effective and tolerated	22	(31.9%)	16	(44.4%)	19	(25.7%)
Effective but not tolerated	7	(10.1%)	0	(0.0%)	6	(8.1%)
Ineffective but tolerated	19	(27.5%)	11	(30.6%)	31	(41.9%)
Ineffective and not tolerated	6	(8.7%)	4	(11.1%)	5	(6.8%)
No prior lithium treatment	15	(21.7%)	5	(13.9%)	13	(17.6%)
All	69	(100.0%)	36	(100.0%)	74	(100.0%)

Table 4.16: The model of homogeneous proportions for Table 4.15

Reaction	Divalproex		Lithium		Placebo	
Effective and tolerated	22.0	(31.8%)	11.5	(31.8%)	23.6	(31.8%)
Effective but not tolerated	5.0	(7.3%)	2.6	(7.3%)	5.4	(7.3%)
Ineffective but tolerated	23.5	(34.1%)	12.3	(34.1%)	25.2	(34.1%)
Ineffective and not tolerated	5.8	(8.4%)	3.0	(8.4%)	6.2	(8.4%)
No prior lithium treatment	12.7	(18.4%)	6.6	(18.4%)	13.6	(18.4%)
All	69.0	(100.0%)	36.0	(100.0%)	74.0	(100.0%)

Table 4.17: Differences between the entries of Table 4.15 and the corresponding entries of Table 4.16

Reaction	Divalproex	Lithium	Placebo
Effective and tolerated	0.0	4.5	−4.6
Effective but not tolerated	2.0	−2.6	0.6
Ineffective but tolerated	−4.5	−1.3	5.8
Ineffective and not tolerated	0.2	1.0	−1.2
No prior lithium treatment	2.3	−1.6	−0.6
All	0.0	0.0	0.0

Table 4.18: The entries of Table 4.17 divided by the square roots of the corresponding entries of Table 4.16

Reaction	Divalproex	Lithium	Placebo
Effective and tolerated	0.0	1.3	−0.9
Effective but not tolerated	0.9	−1.6	0.3
Ineffective but tolerated	−0.9	−0.4	1.2
Ineffective and not tolerated	0.1	0.6	−0.5
No prior lithium treatment	0.6	−0.6	−0.2
All	0.0	0.0	0.0

Chapter 5

Goodness of fit for generalized linear models

Previous chapters illustrate that the root-mean-square discrepancy can be more powerful than canonical measures of discrepancy (such as the deviance or log–likelihood-ratio) when testing the goodness-of-fit for distributional profile (such as tests of Hardy-Weinberg proportions, and of homogeneity, independence, and symmetry in contingency-tables/cross-tabulations). The present chapter illustrates that the root-mean-square can also be more powerful in testing goodness-of-fit for generalized linear models such as that for the standard Poisson regression, when the values of the dependent variables are not too small. When the values of the dependent variables are very small, the discrete Kolmogorov-Smirnov approach is preferable; furthermore, in such circumstances a suitable extension of the Kolmogorov-Smirnov approach can increase statistical power by orders of magnitude, as this chapter illustrates via the generalized linear model associated with standard logistic regression.

5.1 Poisson regression

To test the goodness of fit for a Poisson regression with n nonnegative integers y_1, y_2, $\dots$, y_n and the corresponding values of the independent variables, $x_{1,1}$, $x_{1,2}$, $\dots$, $x_{m,n}$, the null hypothesis is

$$H_0 : y_1, y_2, \dots, y_n \text{ are draws from independent Poisson distributions} \\ \text{with means } \hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_n, \text{ respectively,} \quad (5.1)$$

where

$$\ln(\hat{\mu}_k) = \hat{\beta}^{(0)} + \sum_{j=1}^{m} \hat{\beta}^{(j)} x_{j,k} \quad (5.2)$$

for $k = 1$, 2, $\dots$, n, with the vector β being a nuisance parameter and $\hat{\beta}$ its maximum-likelihood estimate. (Interpreting this type of null hypothesis involves subtleties explicated in the appendix.) This section illustrates testing the goodness of fit for a Poisson regression via an example. The values of the observations y_1, y_2, $\dots$, y_n for this example are not so small; if instead the counts were low, then the more sophisticated methods of Section 5.2 would be more appropriate.

We fit the data in Table 5.1 on $n = 44$ rivers from Sepkoski and Rex (1974). The dependent variable for the regression is the number of species of mussels. The $m = 9$ independent variables are the area of the drainage basin, the nitrate concentration, the hydronium concentration, the amount of residue left by dissolved solids after evaporation (a proxy for the concentration of the solids in the original solution), all their logarithms (including both the area and its logarithm improved the fit somewhat, as did including both the concentrations and their logarithms), and the number of stepping stones "sv" defined by Sepkoski and Rex (1974). For each of the 44 rivers, we modeled the logarithm of the expected number of species as being a constant plus a linear combination of the independent variables (assuming that the number of species follows a Poisson distribution), with the constant and the coefficients in the linear combination being the same for all rivers. This is standard Poisson regression; we used "glmfit" in Matlab for all calculations. The P-values computed via 4 million Monte-Carlo simulations (as in Section 1.3) are

- deviance/log–likelihood-ratio (G^2): .0216
- Freeman-Tukey/Hellinger distance: .0543
- χ^2: .0059
- root-mean-square: .0028

Here the root-mean-square is the most powerful.

Remark 5.1.1. The deviance (or log–likelihood-ratio) is the following divergence of the observations from the Poisson-regression model of (5.1) and (5.2):

$$g^2 = 2\sum_{k=1}^{n} y_k \ln(y_k/\hat{\mu}_k). \tag{5.3}$$

The Freeman-Tukey or Hellinger distance is

$$h^2 = 4\sum_{k=1}^{n}(\sqrt{y_k} - \sqrt{\hat{\mu}_k})^2. \tag{5.4}$$

The χ^2 divergence is

$$\chi^2 = \sum_{k=1}^{n}(y_k - \hat{\mu}_k)^2/\hat{\mu}_k. \tag{5.5}$$

The root-mean-square distance is

$$x = \sqrt{\sum_{k=1}^{n}(y_k - \hat{\mu}_k)^2/n}. \tag{5.6}$$

Table 5.1: Number of species of mussels, area of the drainage basin in square miles, nitrate concentration in parts per million, hydronium concentration — in gram-ions — times 10^7 (pH is the negative of the base-10 logarithm of the hydronium concentration in gram-ions), amount of residue left by dissolved solids after evaporation, and number of stepping stones "sv" from Sepkoski and Rex (1974) for 44 rivers

species	area	nitrate	hydronium	residues	steps
9	8440	0.8	4.0	57	21
8	5960	0.4	3.2	31	20
7	3510	0.6	2.5	65	19
6	1730	0.8	2.5	33	18
11	5020	2.6	6.3	78	17
8	425	8.4	20	120	14
7	1480	3.5	3.2	86	13
11	11500	1.3	2.5	75	12
11	2050	2.1	1.0	138	11
14	13500	0.8	.63	101	10
13	11100	8.7	3.2	122	9
11	27900	3.4	1.3	164	12
14	14400	3.7	0.5	156	11
9	2560	1.1	2.0	47	10
7	2790	0.7	2.0	59	9
12	9850	1.2	1.6	102	8
9	4870	1.2	1.6	99	7
10	9680	4.1	1.3	85	7
12	4150	1.8	1.0	82	6
15	5490	2.6	1.6	76	5
20	9000	1.3	2.0	78	4
10	2100	0.7	10	30	3
13	14400	0.5	3.2	57	3
21	17300	0.5	3.2	57	2
8	3440	0.8	1.6	53	1
21	10800	0.6	5.0	49	0
12	4360	0.1	1.6	55	1
18	14700	0.6	1.0	45	2
2	3720	0.2	32	29	3
4	1870	0.5	20	64	4
10	8800	0.9	1.0	391	5
7	3190	0.9	1.3	88	5
9	2260	0.8	0.5	257	5
6	657	0.2	1.6	105	5
3	521	1.6	3.2	520	5
8	719	0.6	0.4	171	5
9	2120	0.3	0.2	229	5
2	349	1.0	.25	461	6
13	10100	1.0	0.4	133	7
20	3010	0.8	1.6	81	8
33	19700	0.7	1.6	55	9
18	4770	1.1	1.0	57	10
11	1380	0.2	3.2	44	11
23	4270	0.1	2.0	61	11

Remark 5.1.2. A popular model related to Poisson regression is the regression of a nonnegative integer-valued random variable Y on a positive real-valued X, with Y distributed according to a Poisson distribution whose mean is proportional to X (that is, the conditional distribution of Y given X is Poisson with mean proportional to X). More precisely, given nonnegative integers y_1, y_2, …, y_s corresponding to distinct positive real numbers x_1, x_2, …, x_s, we test the significance of assuming

$$H_0 : y_1, y_2, \dots, y_s \text{ are independent draws from the Poisson distributions} \\ \text{with means } \hat{\theta} \cdot x_1,\ \hat{\theta} \cdot x_2,\ \dots,\ \hat{\theta} \cdot x_s, \text{ respectively,} \tag{5.7}$$

where $\hat{\theta}$ is the maximum-likelihood estimate for the proportionality constant,

$$\hat{\theta} = \frac{\sum_{k=1}^{s} y_k}{\sum_{k=1}^{s} x_k}. \tag{5.8}$$

This model is equivalent to tests for homogeneity of proportions in a two-row two-way contingency-table/cross-tabulation for which the entries in the first row are y_1, y_2, …, y_s, and the entries in the second row are an extremely large multiple γ of x_1, x_2, …, x_s. The equivalence follows from the fact that the first entry of column k in such a table is distributed according to the binomial distribution for a large number $\gamma \cdot x_k + y_k \approx \gamma \cdot x_k$ of Bernoulli trials, with the probability of success in each trial approaching $\mathbf{E}\, Y_k/(\gamma \cdot x_k)$ asymptotically in the limit of large γ, where Y_k is the random variable Y conditioned on X taking the value x_k. Such a binomial distribution for Y_k converges to the Poisson distribution with mean $\mathbf{E}\, Y_k$ in the limit of large γ. Homogeneity is the assumption that

$$\mathbf{E}\, Y_j = \frac{(y_j + \gamma x_j) \sum_{k=1}^{s} y_k}{\sum_{k=1}^{s} (y_k + \gamma x_k)} \tag{5.9}$$

for $j = 1$, 2, …, s, which simplifies in the limit of large γ via formula (5.8) to

$$\mathbf{E}\, Y_j = \hat{\theta} \cdot x_j \tag{5.10}$$

for $j = 1$, 2, …, s, that is, the mean of Y is proportional to X.

This type of regression is of interest in testing for the homogeneity of disease incidence or prevalence (assessing whether the incidence or prevalence depends on geographic or ethnic factors, for example). Further information about tests for homogeneity is available, for example, in Chapter 4. In particular, the tests of Chapter 4 are appropriate even when the diseased populations tabulated in the first row are not necessarily much smaller than the unafflicted populations tabulated in the second row; in general the distribution of an entry in the first row is binomial and becomes Poisson only when the total populations are much greater than the afflicted populations.

5.2 Logistic regression

Subtleties arise in testing goodness-of-fit for regressions (including logistic regression) in which the values of the dependent variables predominantly are small nonnegative integers (so-called “low counts”). Explicitly accounting for all applicable independent variables, even when the model being tested does not, is critical. This can increase statistical power by orders of magnitude. The connection with variable and model selection is important.

5.2.1 Introduction

Testing goodness-of-fit for logistic regression has received a remarkable amount of attention, with contributions from Hosmer and Lemeshow (1980), Beran and Millar (1991), Su and Wei (1991), Royston (1992), Stute and Zhu (2002), Weitzen et al. (2004), Pan and Lin (2005), and Allison (2012), among many others; see, for example, the references of Hosmer and Lemeshow (2000). The discussion in the present section is closely related. However, even when computing exact P-values as in earlier works, as summarized in the addendum, Section 5.2.4, omissions in the standard tests can reduce the standards' statistical power by orders of magnitude. The present, introductory subsection addresses the omissions. For details, see Section 5.2.2 or the concluding remarks (Section 5.2.3).

A representative use of logistic regression is to model the absence (0) or presence (1) of coronary heart disease — resulting in a "dependent variable" for coronary heart disease that is binary/dichotomous (meaning that the observed values are zeros and ones). The regression predicts the dependent variable via "independent variables" such as age, cholesterol level, diastolic and systolic blood pressure, and others listed in Section 5.2.2.4 below.

To be precise, the null hypothesis for a logistic regression of a binary/dichotomous random variable Y_k on real numbers $x_{j,k}$ (with $k = 1, 2, \dots, n$, and $j = 1, 2, \dots, m \ll n$), given observations $y_1, y_2, \dots, y_n$ of $Y_1, Y_2, \dots, Y_n$, respectively, is

$$H_0 : y_1, y_2, \dots, y_n \text{ are draws from independent Bernoulli distributions} \\ \text{with means } \hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_n, \text{ respectively,} \quad (5.11)$$

where

$$\operatorname{logit}(\hat{\mu}_k) = \hat{\beta}^{(0)} + \sum_{j=1}^{\ell} \hat{\beta}^{(j)} x_{j,k} \quad (5.12)$$

for $k = 1, 2, \dots, n$; here the vector β is a nuisance parameter, $\hat{\beta}$ is its maximum-likelihood estimate, ℓ is a nonnegative integer no greater than m, and

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right) \quad (5.13)$$

is the logarithm of the odds $p/(1-p)$; needless to say, altering the base of the logarithm just rescales β. (Interpreting this type of null hypothesis involves subtleties explicated in the appendix.) In this general formulation, the number ℓ of terms in the sum from (5.12) may be less than the total number m of independent variables (while m itself should be significantly less than n in order to avoid overfitting); nevertheless, accounting for all applicable independent variables when testing goodness-of-fit is very important, as discussed shortly and further detailed below in Section 5.2.2.1.

Goodness of fit gauges the consistency of the given observed values with the model assumed under the null hypothesis. The observed values $y_1, y_2, \dots, y_n$ are zeros and ones, so probing their distribution requires aggregating or accumulating these very low counts, just as in the case of draws from continuous probability densities (the continuous case requires use of cumulative distribution functions, probability density estimation, Neyman smooth tests, or something similarly aggregative). A cumulative measure of the distance between the observed data and the model assumed under the null hypothesis is the discrete Kolmogorov-Smirnov statistic detailed in Chapter 2, namely,

$$d = \max_{1 \le j \le n} \left| \sum_{k=1}^{j} y_{\sigma_k} - \sum_{k=1}^{j} \hat{\mu}_{\sigma_k} \right| = \max_{1 \le j \le n} \left| \sum_{k=1}^{j} r_{\sigma_k} \right|, \tag{5.14}$$

where the ordering σ is a permutation of the integers 1, 2, …, n, and the residual r_k is

$$r_k = y_k - \hat{\mu}_k \tag{5.15}$$

for $k = 1, 2, \dots, n$. Please note that $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_n$ are from (5.12); (5.12) is similar to (5.16) below, but (5.16) influences only the permutation σ.

Critically, the ordering σ in (5.14) should take into account all applicable independent variables, not just those included in the model of (5.11) and (5.12). If a suitable ordering for the observations is not known a priori, there is a good substitute, namely to choose a permutation σ of the integers 1, 2, …, n such that $\tilde{\mu}_{\sigma_1} \le \tilde{\mu}_{\sigma_2} \le \dots \le \tilde{\mu}_{\sigma_n}$, replacing (5.12) with a fit to all the data via

$$\operatorname{logit}(\tilde{\mu}_k) = \tilde{\beta}^{(0)} + \sum_{j=1}^{m} \tilde{\beta}^{(j)} x_{j,k} \tag{5.16}$$

for $k = 1, 2, \dots, n$ (notice that the sum here runs up to m, not just ℓ), where $\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_n$ are the estimated means of the postulated independent Bernoulli distributions. That is, this natural ordering sorts based on the estimated means $\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_n$ when estimated using all applicable independent variables, not using only the ℓ terms appearing in the sum from (5.12). This ordering incorporates information from all values $x_{1,1}, x_{1,2}, \dots, x_{m,n}$, as these values influence the estimated values $\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_n$ in (5.16).

Goodness of fit is gauged via the "P-value." As reviewed in the addendum, Section 5.2.4, the P-value associated with d defined in (5.14) is the proportion of Monte-Carlo simulations for which the distance d is greater than or equal to the distance d for the original, observed data, in the limit of a large number of simulations, each simulated according to (5.11) and (5.12). If a P-value is very small, then we can be confident that the model, given in (5.11) and (5.12), is not consistent with the data, even when allowing for the expected statistical fluctuations.

The next subsection illustrates that the above approach can be very powerful. This introductory subsection will now conclude with incidental remarks about other approaches:

Remark 5.2.1. If we replace the natural ordering (discussed above) with a permutation σ satisfying $r_{\sigma_1} \le r_{\sigma_2} \le \dots \le r_{\sigma_n}$, then (5.14) simplifies to

$$d = \frac{1}{2} \sum_{k=1}^{n} |r_k|. \tag{5.17}$$

Maximizing (5.14) over all permutations σ of the integers 1, 2, …, n also yields (5.17). The simplification to (5.17) is due to Hoeffding (1965) (see the top of page 396 in Hoeffding's article); this follows from the fact that the sum of the residuals is 0, that is, $\sum_{k=1}^{n} r_k = 0$ (the sum of the residuals is 0 on account of the constant term $\hat{\beta}^{(0)}$ in (5.12)).

Remark 5.2.2. An often appealing alternative to the Kolmogorov-Smirnov statistic defined in (5.14) is the Kuiper statistic

$$d = \max_{1 \le j \le n} \sum_{k=1}^{j} r_{\sigma_k} - \min_{1 \le j \le n} \sum_{k=1}^{j} r_{\sigma_k}. \tag{5.18}$$

Under the ordering for which the permutation σ satisfies $r_{\sigma_1} \le r_{\sigma_2} \le \dots \le r_{\sigma_n}$, (5.18) simplifies to (5.14) and hence to (5.17), since $\sum_{k=1}^{n} r_{\sigma_k} = \sum_{k=1}^{n} r_k = 0$, while this choice of ordering ensures $\sum_{k=1}^{j} r_{\sigma_k} \le 0$ for all $j = 1, 2, \dots, n$. In fact, (5.18) simplifies to (5.14) and hence to (5.17), for any permutation σ such that $r_{\sigma_{k+1}} \le r_{\sigma_{k+2}} \le \dots \le r_{\sigma_n} \le r_{\sigma_1} \le r_{\sigma_2} \le \dots \le r_{\sigma_k}$ for some positive integer k. Indeed, the value of (5.18) for such a permutation is the same as for a permutation satisfying $r_{\sigma_1} \le r_{\sigma_2} \le \dots \le r_{\sigma_n}$; this invariance is the principal appeal of the Kuiper statistic, as discussed by Stephens (1970) and Section 14.3.4 of Press et al. (2007).

Remark 5.2.3. In this remark, we consider the case when the vectors $(x_{1,k}, x_{2,k}, \dots, x_{\ell,k})$ are different for different values of k. This is the case for the above example of modeling the absence or presence of coronary heart disease, provided that no two subjects have exactly the same age, cholesterol level, diastolic and systolic blood pressure, and so on. The deviance (also known as the log–likelihood ratio or "G^2") is then

$$g^2 = -2 \sum_{k=1}^{n} \Big(y_k \ln(\hat{\mu}_k) + (1 - y_k) \ln(1 - \hat{\mu}_k) \Big) = -2 \sum_{y_k = 0} \ln(1 - \hat{\mu}_k) - 2 \sum_{y_k = 1} \ln(\hat{\mu}_k), \tag{5.19}$$

where $\hat{\mu}_k$ is the maximum-likelihood estimate of the mean for the Bernoulli distribution producing y_k under the model; $\hat{\mu}_k$ is defined in (5.12). If $\hat{\mu}_k$ is small whenever $y_k = 0$ and $1 - \hat{\mu}_k$ is small whenever $y_k = 1$ (that is, the absolute residual $|r_k| = |y_k - \hat{\mu}_k|$ is small for $k = 1, 2, \dots, n$), then

$$- \sum_{y_k = 0} \ln(1 - \hat{\mu}_k) - \sum_{y_k = 1} \ln(\hat{\mu}_k) \approx \sum_{y_k = 0} \hat{\mu}_k + \sum_{y_k = 1} (1 - \hat{\mu}_k). \tag{5.20}$$

Moreover,

$$\sum_{y_k = 0} \hat{\mu}_k + \sum_{y_k = 1} (1 - \hat{\mu}_k) = - \sum_{y_k = 0} r_k + \sum_{y_k = 1} r_k = \sum_{y_k = 0} |r_k| + \sum_{y_k = 1} |r_k| = \sum_{k=1}^{n} |r_k|, \tag{5.21}$$

where $r_k = y_k - \hat{\mu}_k$ is the residual. Combining (5.19), (5.20), and (5.21) yields that

$$g^2 \approx 2 \sum_{k=1}^{n} |r_k| \tag{5.22}$$

if the absolute residuals $|r_1|$, $|r_2|$, $\dots$, $|r_n|$ are small. Please note that the right-hand side of (5.22) is exactly four times the right-hand side of (5.17). The deviance g^2 is not terribly helpful for gauging goodness-of-fit with logistic regression, as the deviance is a deterministic function of the estimated means of the postulated Bernoulli distributions — the deviance depends on the observed data only through dependence on the estimated means. In fact, McCullagh and Nelder (1989) and others have shown that

$$g^2 = 2\sum_{k=1}^{n} \hat{\mu}_k \ln\left(\frac{1-\hat{\mu}_k}{\hat{\mu}_k}\right) - 2\sum_{k=1}^{n} \ln(1-\hat{\mu}_k) \tag{5.23}$$

when the vectors $(x_{1,k}, x_{2,k}, \dots, x_{\ell,k})$, with $k = 1, 2, \dots, n$, are distinct; please notice that the observations y_1, y_2, …, y_n do not appear explicitly in the right-hand side of (5.23) (though of course the estimated means $\hat{\mu}_1$, $\hat{\mu}_2$, …, $\hat{\mu}_n$ do depend on the observations y_1, y_2, …, y_n).

Remark 5.2.4. The tests of Hosmer and Lemeshow (1980, 2000) replace the Kolmogorov-Smirnov statistic d defined in (5.14) with χ^2 for the model of binomial distributions corresponding to quantiles (typically the deciles) for the estimated means of the postulated Bernoulli distributions. For example, if n is divisible by 10 (for notational convenience), then we could replace d defined in (5.14) with

$$h_\ell = \sum_{k=1}^{10} \frac{(n_k - \hat{\eta}_k)^2}{\hat{\eta}_k(1 - 10\hat{\eta}_k/n)}, \tag{5.24}$$

where n_k is the sum of the observed values y_1, y_2, …, y_n from the kth group (that is, n_k is the number of values from the kth group that are equal to 1), grouping according to the deciles of the estimated means $\hat{\mu}_1$, $\hat{\mu}_2$, …, $\hat{\mu}_n$, and where $\hat{\eta}_k$ is the sum of the estimated means $\hat{\mu}_1$, $\hat{\mu}_2$, …, $\hat{\mu}_n$ in the kth decile (that is, in the kth group). Notice that the denominator in (5.24) would be the expected value of the numerator if n_k were drawn from a binomial distribution whose maximal possible value is $n/10$ and whose mean is $\hat{\eta}_k$ (conditional on knowing $\hat{\eta}_k$). This approach of Hosmer and Lemeshow (1980, 2000) is similar to the cumulative (Kolmogorov-Smirnov) approach detailed in the present section, but involves explicit binning. Allison (2012) criticizes the binning and provides references to many other critiques.

5.2.2 Examples

5.2.2.1 Overview

This subsection introduces two thought experiments, as well as their illustration below via the computer-assisted analysis of three real data sets. The statistic detailed in Section 5.2.1 — Kolmogorov-Smirnov ordered based on $\tilde{\mu}$ — far outperforms the standard statistic described in Remark 5.2.4 — Hosmer-Lemeshow ordered based on $\hat{\mu}$. Whereas $\hat{\mu}$ considers only the ℓ summands in (5.12), $\tilde{\mu}$ takes into account all m independent variables in (5.16).

Accounting for all applicable independent variables when testing goodness-of-fit is critical; using only the independent variables included in (5.12) is not always sufficient. Consider, for example, the case when $\ell = 0$ in (5.12), that is, when the sum in (5.12) is absent. The resulting logistic regression amounts to regressing y_1, y_2, …, y_n against a constant; the estimated means in (5.12) are then all the same, that is, $\hat{\mu}_1 = \hat{\mu}_2 = \cdots = \hat{\mu}_n$. Unless nearly all y_1, y_2, …, y_n are equal, the fit cannot possibly be good, yet no standard test for goodness of fit can detect the poor fit. Indeed, any ordering for the observations must be random with the standard tests, since the estimated means are all the same, $\hat{\mu}_1 = \hat{\mu}_2 = \cdots = \hat{\mu}_n$. The standard tests consider only the independent variables appearing in (5.12) — surely not enough when $\ell = 0$. Without substantial information from independent variables to inform the ordering and the associated accumulation or aggregation, the data cannot possibly invalidate the tested model. The values of y_1, y_2, …, y_n (each 0 or 1) are not informative if their ordering is totally random. The data can invalidate the tested model only when there are variables that can inform the ordering, possibly including variables not included in the tested model.

Similarly, if $\ell = 1$ and the values $x_{1,1}$, $x_{1,2}$, …, $x_{1,n}$ are drawn independently and with identical distributions (i.i.d.) from the uniform distribution over $(0, 1)$, with no relation to y_1, y_2, …, y_n, then the fit via (5.12) cannot possibly be good. Yet, standard tests for goodness of fit (which consider only those independent variables appearing in (5.12), not any "extras" from the rest of the data set) cannot detect the poor fit. Moreover, such problems are not limited to thought experiments, as Hosmer et al. (1997), Weitzen et al. (2004), and the remainder of the present subsection illustrate via the analysis of real data.

The following examples analyze several data sets, referring to the test statistic detailed in Section 5.2.1 as Kolmogorov-Smirnov ordered based on $\tilde{\mu}$, and we recommend Kolmogorov-Smirnov ordered based on $\tilde{\mu}$ for general-purpose use; the standard statistic (detailed in Remark 5.2.4) is Hosmer-Lemeshow grouped based on $\hat{\mu}$ (not $\tilde{\mu}$), which performs relatively poorly in our tests. The other statistics mentioned in the tables below (namely, G^2, Freeman-Tukey, χ^2, and the Euclidean distance or root-mean-square) are not really relevant to testing goodness-of-fit for logistic regression, as they do not depend on the ordering; the tables include their P-values only for completeness, illustrating their low statistical power. As always, if any of the P-values is very small, then we can have confidence that the model in (5.11) and (5.12) does not yield a good fit; that is, we can have confidence that the observed data is not consistent (up to the expected statistical fluctuations) with assuming (5.11) and (5.12).

We used Matlab's "glmfit" for all calculations, incorporating the addendum, Section 5.2.4, to compute all P-values; Remark 1.3.2 discusses the resulting accuracy. We calculated each P-value via 4,000,000 Monte-Carlo simulations; running so many simulations is not really necessary.

5.2.2.2 Data from Finney (1947)

Table 5.3 displays the P-values for several goodness-of-fit tests applied to the classic data set of Finney (1947). Table 5.2 displays the data set, which consists of $n = 39$ observations of a dependent variable together with 2 independent variables.

Table 5.3 reports on four experiments. First, we set $\ell = 0$, i.e., omit the sum in (5.12) entirely, while retaining $m = 2$ in (5.16). Second, we set $\ell = 1$ and generate a new, additional independent variable, drawing $x_{1,1}$, $x_{1,2}$, …, $x_{1,n}$ i.i.d. from the uniform distribution over $(0, 1)$. Third, we discard the extra random independent variable, retaining both original independent variables, with $\ell = 2$ and $m = 2$. Fourth, we again include the extra random independent variable (making $m = 2+1$), but do not include it in the tested model (so $\ell = 2$).

For the Hosmer-Lemeshow statistics of Remark 5.2.4, we consider different groupings. In the first, with 3 groups in all, each group contains 13 observations. In the second, with 5 groups in all, the initial 4 groups contain 8 observations each, and the last contains 7.

5.2.2.3 Data from Hosmer and Lemeshow (2000)

Table 5.4 displays the P-values for several goodness-of-fit tests applied to the "UIS" data set of Hosmer and Lemeshow (2000). This data set consists of $n = 575$ observations of a dependent variable "dfree" together with 11 independent variables "age," "beck," "ndrgfp1," "ndrgfp2," "ivhx_2," "ivhx_3," "race," "treat," "site," "ageXndrgfp1," and "raceXsite" (these include transformations and products of the original variables "dfree," "age," "beck," "ndrugtx," "ivhx," "race," "treat," and "site," as detailed in Chapter 4 of Hosmer and Lemeshow (2000)). For the Hosmer-Lemeshow statistics of Remark 5.2.4, we aggregate the data into 10 groups, with the initial 9 containing 58 observations each, and the last containing 53.

Table 5.4 reports on five experiments. First, we set $\ell = 0$, that is to say, omit the sum in (5.12) entirely, while retaining $m = 11$ in (5.16). Second, we set $\ell = 1$ and generate a new, additional independent variable, drawing $x_{1,1}$, $x_{1,2}$, ..., $x_{1,n}$ i.i.d. from the uniform distribution over $(0, 1)$. Third, we discard the extra random independent variable, retaining the original $m = 11$, and set $\ell = 9$, taking the independent variables in the regression to be "age," "beck," "ndrgfp1," "ndrgfp2," "ivhx_2," "ivhx_3," "race," "treat," and "site" (these are all those not involving products of other variables). Fourth, we include all 11 original independent variables, with both $\ell = 11$ and $m = 11$. Fifth, we again include the extra random independent variable (making $m = 11 + 1$), but do not include it in the tested model (so $\ell = 11$).

5.2.2.4 Data from Kleinbaum and Klein (2010)

Table 5.5 displays the P-values for several goodness-of-fit tests applied to the "Evans County" data set of Kleinbaum and Klein (2010). This data set consists of $n = 609$ observations of a dependent variable "chd" together with 10 independent variables "age," "cat," "chl," "dbp," "ecg," "hpt," "sbp," "smk," "catXchl," and "catXhpt" (these include products of the original variables "chd," "age," "cat," "chl," "dbp," "ecg," "hpt," "sbp," and "smk," as in model "EC4" from Chapter 9 of Kleinbaum and Klein (2010)). For the Hosmer-Lemeshow statistics of Remark 5.2.4, we aggregate the data into 10 groups, with the initial 9 containing 61 observations each, and the last containing 60.

Table 5.5 reports the results of five experiments. First, we set $\ell = 0$, i.e., omit the sum in (5.12) entirely, while retaining $m = 10$ in (5.16). Second, we set $\ell = 1$ and generate a new, additional independent variable, drawing $x_{1,1}$, $x_{1,2}$, ..., $x_{1,n}$ i.i.d. from the uniform distribution over $(0, 1)$. Third, we discard the extra random independent variable, retaining the original $m = 10$, and set $\ell = 6$, taking the independent variables in the regression to be "age," "cat," "chl," "ecg," "hpt," and "smk" — those included for model "EC3" in Chapter 9 of Kleinbaum and Klein (2010). Fourth, we include all 10 original independent variables, with both $\ell = 10$ and $m = 10$. Fifth, we again include the extra random independent variable (making $m = 10 + 1$), but do not include it in the tested model (so $\ell = 10$).

Table 5.2: $x_{1,k}$, $x_{2,k}$, and y_k for $k = 1, 2, \ldots, 39$, from Finney (1947); Finney (1947) refers to y as "response"

x_1	x_2	y
1.57	0.92	1
1.54	1.04	1
1.10	1.40	1
0.88	1.18	1
0.90	1.51	1
0.85	1.54	1
0.78	0.88	0
1.04	1.23	0
0.95	0.88	0
0.95	0.65	0
0.90	0.76	0
0.74	1.44	0
0.78	1.48	0
1.15	1.37	1
0.88	1.57	1
1.36	1.21	1
1.51	1.20	1
0.93	1.15	1
1.23	1.03	0
1.26	1.26	1
0.60	1.30	0
0.98	1.13	0
1.13	1.13	0
1.18	1.13	0
1.20	1.25	1
0.78	1.18	0
1.26	1.18	1
0.98	1.28	0
1.28	0.98	1
1.20	0.60	0
1.43	0.88	1
1.37	0.48	0
1.04	1.26	0
1.04	1.34	1
1.08	1.30	1
0.90	1.52	1
0.98	1.28	0
0.88	1.28	0
1.11	1.21	1

Table 5.3: P-values for the data set of Finney (1947) (see Table 5.2); $n = 39$

	$\ell = 0$;[1] $m = 2$	$\ell = 1$;[2] $m = 2{+}1$	$\ell = 2$;[3] $m = 2$	$\ell = 2$;[4] $m = 2{+}1$
Kolmogorov-Smirnov (ordering based on $\tilde{\mu}$)	.0000003†	.0000003†	.0075	.039
Kolmogorov-Smirnov (ordering based on $\hat{\mu}$)	.249	.452	.0075	.0075
Kolmogorov-Smirnov (ordering based on r)	.134	.053	.355	.355
G^2 (the deviance, log–likelihood-ratio, ...)	.248	.053	.324	.324
Freeman-Tukey (Hellinger distance)	.517	.239	.250	.250
χ^2 (sum of the squares of Pearson residuals)	.385	.405	.182	.182
Euclidean distance or root-mean-square	.236	.053	.393	.393
Hosmer-Lemeshow (with 3 groups from $\tilde{\mu}$)	.582	.248	.107	.065
Hosmer-Lemeshow (with 3 groups from $\hat{\mu}$)	.590	.298	.107	.107
Hosmer-Lemeshow (with 5 groups from $\tilde{\mu}$)	.658	.324	.787	.695
Hosmer-Lemeshow (with 5 groups from $\hat{\mu}$)	.597	.201	.787	.787

[1] i.e., omitting the sum in (5.12) entirely, while retaining the original $m = 2$

[2] i.e., with an extra independent variable, with $x_{1,1}$, $x_{1,2}$, ..., $x_{1,n}$ drawn i.i.d. from $U(0, 1)$

[3] i.e., including both original independent variables

[4] i.e., with an extra independent variable, with $x_{3,1}$, $x_{3,2}$, ..., $x_{3,n}$ drawn i.i.d. from $U(0, 1)$, but without including the extra variable in the model being tested (while still including both original independent variables)

† in these cases, only 1 simulation (out of 4,000,000) produced a Kolmogorov-Smirnov statistic at least as large as that for the original, observed data

Table 5.4: P-values for the "UIS" data set of Hosmer and Lemeshow (2000); $n = 575$

	$\ell = 0;$[1] $m = 11$	$\ell = 1;$[2] $m = 11{+}1$	$\ell = 9;$[3] $m = 11$	$\ell = 11;$[4] $m = 11$	$\ell = 11;$[5] $m = 11{+}1$
Kolmogorov-Smirnov (ordering based on $\tilde{\mu}$)	.00001	.00001	.0049	.736	.861
Kolmogorov-Smirnov (ordering based on $\hat{\mu}$)	.609	.276	.115	.736	.736
Kolmogorov-Smirnov (ordering based on r)	.486	.484	.334	.319	.319
G^2 (the deviance, log–likelihood-ratio, ...)	.516	.482	.343	.311	.311
Freeman-Tukey (Hellinger distance)	.516	.478	.314	.286	.286
χ^2 (sum of the squares of Pearson residuals)	.473	.734	.740	.300	.300
Euclidean distance or root-mean-square	.507	.484	.317	.319	.319
Hosmer-Lemeshow (for deciles of $\tilde{\mu}$)	.023	.017	.991	.673	.594
Hosmer-Lemeshow (for deciles of $\hat{\mu}$)	.688	.249	.781	.673	.673

[1] i.e., omitting the sum in (5.12) entirely, while retaining the original $m = 11$

[2] i.e., with an extra independent variable, with $x_{1,1}$, $x_{1,2}$, ..., $x_{1,n}$ drawn i.i.d. from $U(0, 1)$

[3] i.e., taking the independent variables in the regression to be "age," "beck," "ndrgfp1," "ndrgfp2," "ivhx_2," "ivhx_3," "race," "treat," and "site" (these are all those not involving products of other variables), while retaining the original $m = 11$

[4] i.e., including all 11 original independent variables

[5] i.e., with an extra independent variable, with $x_{12,1}, x_{12,2}, \ldots, x_{12,n}$ drawn i.i.d. from $U(0, 1)$, but without including the extra variable in the model being tested (while still including all 11 original independent variables)

Table 5.5: P-values for the "Evans County" data set of Kleinbaum and Klein (2010); $n = 609$

	$\ell = 0;^1$ $m = 10$	$\ell = 1;^2$ $m = 10+1$	$\ell = 6;^3$ $m = 10$	$\ell = 10;^4$ $m = 10$	$\ell = 10;^5$ $m = 10+1$
Kolmogorov-Smirnov (ordering based on $\tilde{\mu}$)	$\le .0000003^\dagger$	$\le .0000003^\dagger$	$\le .0000003^\dagger$	.193	.328
Kolmogorov-Smirnov (ordering based on $\hat{\mu}$)	.357	.905	.738	.193	.193
Kolmogorov-Smirnov (ordering based on r)	.471	.474	.431	.418	.418
G^2 (the deviance, log–likelihood-ratio, . . .)	.519	.472	.412	.357	.357
Freeman-Tukey (Hellinger distance)	.485	.472	.404	.405	.405
χ^2 (sum of the squares of Pearson residuals)	.354	.427	.759	.010	.010
Euclidean distance or root-mean-square	.514	.474	.431	.451	.451
Hosmer-Lemeshow (for deciles of $\tilde{\mu}$)	$\le .0000003^\dagger$	.000002	.995	.237	.186
Hosmer-Lemeshow (for deciles of $\hat{\mu}$)	.651	.585	.822	.237	.237

[1] i.e., omitting the sum in (5.12) entirely, while retaining the original $m = 10$

[2] i.e., with an extra independent variable, with $x_{1,1}$, $x_{1,2}$, . . . , $x_{1,n}$ drawn i.i.d. from $U(0, 1)$

[3] i.e., taking the independent variables in the regression to be "age," "cat," "chl," "ecg," "hpt," and "smk" — those included for model "EC3" in Chapter 9 of Kleinbaum and Klein (2010) — while retaining the original $m = 10$

[4] i.e., including all 10 original independent variables

[5] i.e., with an extra independent variable, with $x_{11,1}$, $x_{11,2}$, . . . , $x_{11,n}$ drawn i.i.d. from $U(0, 1)$, but without including the extra variable in the model being tested (while still including all 10 original independent variables)

[†] in these cases, no simulations (out of 4,000,000) produced a test statistic at least as large as that for the original, observed data

5.2.3 Conclusion

The discrete Kolmogorov-Smirnov test with an ordering based on all applicable independent variables produces P-values that are orders of magnitude better than those for the standards (such as the usual Hosmer-Lemeshow test) in many circumstances for which the model clearly fits very poorly. In particular, this happens if the model omits significant explanatory variables that are in the given data. The Kolmogorov-Smirnov approach is not the only possibility, but the above examples (both the thought experiments and the real data analyses) argue strongly in favor of aggregating based on all applicable independent variables, not based on just those incorporated into the model being tested. In fact, this is a strong argument relevant to testing goodness-of-fit for any regression with low counts, including the simplest logistic regression that is the focus of the present section.

5.2.4 Addendum: Computation of P-values

This subsection briefly describes Monte-Carlo simulations yielding estimates for P-values; the standard errors of the estimates are inversely proportional to the square root of the number of simulations conducted. The P-values being estimated are exact for any number n of observations and also have desirable properties in the limit that n is large, as detailed in the appendix. To calculate a P-value, we first estimate β from the given observations, obtaining $\hat{\beta}$ in (5.12) and $\tilde{\beta}$ in (5.16), and then calculate the test statistic (such as d in (5.14)). We next run many simulations. To conduct a single simulation, we perform the following three-step procedure:

1. we generate n independent draws according to (5.11) and (5.12),

2. we fit the parameter β from the data generated in Step 1, both using all m variables in (5.16) and using only those ℓ in (5.12), obtaining new estimates $\tilde{\tilde{\beta}}$ and $\hat{\hat{\beta}}$, respectively, and

3. we calculate the test statistic (such as the discrete Kolmogorov-Smirnov distance d) using the new y_1, y_2, …, y_n generated in Step 1 and new estimates $\hat{\hat{\mu}}_1$, $\hat{\hat{\mu}}_2$, …, $\hat{\hat{\mu}}_n$, determining the ordering (that is, the permutation σ from Section 5.2.1) for the statistic by sorting $\tilde{\tilde{\mu}}_1$, $\tilde{\tilde{\mu}}_2$, …, $\tilde{\tilde{\mu}}_n$, with

$$\operatorname{logit}(\hat{\hat{\mu}}_k) = \hat{\hat{\beta}}^{(0)} + \sum_{j=1}^{\ell} \hat{\hat{\beta}}^{(j)} x_{j,k} \tag{5.25}$$

$$\operatorname{logit}(\tilde{\tilde{\mu}}_k) = \tilde{\tilde{\beta}}^{(0)} + \sum_{j=1}^{m} \tilde{\tilde{\beta}}^{(j)} x_{j,k} \tag{5.26}$$

   for $k = 1, 2, \dots, n$, where $\tilde{\tilde{\beta}}$ and $\hat{\hat{\beta}}$ are the estimates calculated in Step 2 from the data generated in Step 1.

After conducting many such simulations, we may estimate the P-value as the fraction of the statistics calculated in Step 3 that are greater than or equal to the statistic calculated from the given data. As detailed in Remark 1.3.2, the accuracy of the estimated P-value is inversely proportional to the square root of the number of simulations conducted.

Chapter 6

Asymptotic P-values without nuisance parameters

The present chapter provides efficient black-box algorithms for calculating, in the limit of large numbers of draws, the asymptotic P-values for the root-mean-square statistic discussed in the previous chapters. As observed in the previous chapters, computing the exact P-values via Monte-Carlo simulation is often feasible, too.

6.1 Recap

A basic task in statistics is to ascertain whether a given set of observations does not arise as independent and identically distributed (i.i.d.) draws from a specified probability distribution (this specified distribution is known as the "model"). 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 draws were not taken i.i.d. 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 draws, and then measure the root-mean-square difference between this empirical distribution and the specified model distribution; see, for example, Rao (2002), page 123 of Varadhan et al. (1974), or Section 6.2 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 observations do not arise as i.i.d. draws from the specified probability distribution.

Let us denote by x the value of the root-mean-square for the given draws; let us denote by X the root-mean-square statistic constructed for different draws that definitely are in fact taken i.i.d. from the specified model distribution. Then, the P-value is defined to be the probability that $X \geq x$ (viewing X — but not x — as a random variable).

Unfortunately, the P-values 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 reciprocals of the model probabilities associated with the various bins, obtaining the classic χ^2 statistic; see, for example, Pearson (1900) or Remark 6.2.1 below. However, with the now widespread availability of computers, direct use of the root-mean-square statistic has become feasible and convenient. The present chapter provides efficient black-box algorithms for computing the P-values for any specified model distribution, in the limit of large numbers of draws. Calculating P-values for small numbers of draws via Monte Carlo can also be practical, as illustrated in previous chapters.

The simple statistic described above would seem to be more natural than the standard χ^2 statistic of Pearson (1900), is typically easier to use (since it does not require any rebinning of data), and is more powerful in many circumstances. 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, Chapter 1). The rest of the present chapter has the following structure: Section 6.2 details the statistic discussed above, expressing the P-values for the associated goodness-of-fit test in a form suitable for computation. Section 6.3 discusses the most involved part of the computation of the P-values, computing the cumulative distribution function of the sum of the squares of independent centered Gaussian random variables. Section 6.4 summarizes the method for computing the P-values of the root-mean-square statistic. Section 6.5 applies the method to several examples.

6.2 The sum-of-squares statistic

This section details the root-mean-square discussed briefly in Section 6.1, 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 P-values (in the limit of large numbers of draws): given a value x for the root-mean-square statistic constructed from observations coming from an unknown distribution, the P-value that the observations do not arise as i.i.d. draws from the specified model is the probability that the root-mean-square statistic is greater than or equal to 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 m bins, numbered 1, 2, ..., m, we denote by $p_0^{(1)}$, $p_0^{(2)}$, ..., $p_0^{(m)}$ the probabilities associated with the respective bins under the specified model; of course, $\sum_{j=1}^m p_0^{(j)} = 1$. To obtain a draw conforming to the model, we select at random one of the m bins, with probabilities $p_0^{(1)}$, $p_0^{(2)}$, ..., $p_0^{(m)}$. We perform this selection independently n times. For $j = 1, 2, \ldots, m$, we denote by $Y^{(j)}$ the fraction of times that we choose bin j (that is, $Y^{(j)}$ is the number of times that we choose bin j, divided by n); obviously, $\sum_{j=1}^m Y^{(j)} = 1$. We define X_j to be $\sqrt{n}$ times the difference of $Y^{(j)}$ from its expected value, that is,

$$X_j = \sqrt{n}\,(Y^{(j)} - p_0^{(j)}) \tag{6.1}$$

for $j = 1, 2, \ldots, m$. Finally, we form the statistic

$$X = \sum_{j=1}^{m} X_j^2, \tag{6.2}$$

and now determine its distribution in the limit of large n. X is the square of the root-mean-square statistic $\sqrt{\sum_{j=1}^{m}(nY^{(j)} - np_0^{(j)})^2/n}$. Since the square root is a monotonically increasing function, the P-values are the same whether determined via X or via $\sqrt{X}$; for convenience, we focus on X below.

Remark 6.2.1. The classic χ^2 test for goodness-of-fit of Pearson (1900) replaces (6.2) with the statistic

$$\chi^2 = \sum_{j=1}^{m} \frac{X_j^2}{p_0^{(j)}}, \tag{6.3}$$

where $X_1, X_2, \ldots, X_m$ are the same as in (6.1) and (6.2). χ^2 defined in (6.3) has the advantage that its P-values are the same for every model distribution, independent of the values of $p_0^{(1)}, p_0^{(2)}, \ldots, p_0^{(m)}$, in the limit of large numbers of draws. In contrast, using X defined in (6.2) requires computing its P-values anew for every different model.

The multivariate central limit theorem shows that the joint distribution of $X_1, X_2, \ldots, X_m$ converges in distribution as $n \to \infty$, with the limiting generalized probability density proportional to

$$\exp\left(-\sum_{j=1}^{m} \frac{x_j^2}{2p_0^{(j)}}\right) \; \delta\left(\sum_{j=1}^{m} x_j\right), \tag{6.4}$$

where δ is the Dirac delta; see, for example, Moore and Spruill (1975) or Chapter 25 and Example 15.3 of Kendall et al. (2009). The generalized probability density (6.4) is a centered multivariate Gaussian concentrated on a hyperplane passing through the origin (the hyperplane consists of the points such that $\sum_{j=1}^{m} x_j = 0$); the restriction of the generalized probability density (6.4) to the hyperplane through the origin is also a centered multivariate Gaussian. Thus, the distribution of X defined in (6.2) converges as $n \to \infty$ to the distribution of the sum of the squares of $m - 1$ 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, Moore and Spruill (1975) or Chapter 25 of Kendall et al. (2009). Given these variances, the following section describes an efficient algorithm for computing the probability that the associated sum of squares is greater than or equal to any particular value; this probability is the desired P-value, in the limit of large numbers of draws. See Sections 6.4 and 6.5 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 $1/p_0^{(1)}$, $1/p_0^{(2)}$, $\ldots$, $1/p_0^{(m)}$ from both the left and the right by the projection matrix P whose entries are

$$P_{j,k} = \begin{cases} 1 - \frac{1}{m}, & j = k \\ -\frac{1}{m}, & j \neq k \end{cases} \tag{6.5}$$

for $j, k = 1, 2, \ldots, m$ (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

$$B_{j,k} = \begin{cases} \frac{1}{p_0^{(j)}} - \frac{1}{m}\left(\frac{1}{p_0^{(j)}} + \frac{1}{p_0^{(k)}}\right) + \frac{1}{m^2}\sum_{i=1}^{m}\frac{1}{p_0^{(i)}}, & j = k \\ -\frac{1}{m}\left(\frac{1}{p_0^{(j)}} + \frac{1}{p_0^{(k)}}\right) + \frac{1}{m^2}\sum_{i=1}^{m}\frac{1}{p_0^{(i)}}, & j \neq k \end{cases} \tag{6.6}$$

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

Remark 6.2.2. The $m \times m$ matrix B defined in (6.6) is the sum of a diagonal matrix and a low-rank matrix. The methods of Gu and Eisenstat (1994, 1995) for computing the eigenvalues of such a matrix B require only either $\mathcal{O}(m^2)$ or $\mathcal{O}(m)$ floating-point operations. The $\mathcal{O}(m^2)$ methods of Gu and Eisenstat (1994, 1995) are usually more efficient than the $\mathcal{O}(m)$ method of Gu and Eisenstat (1995), unless m is impractically large.

Remark 6.2.3. It is not hard to accommodate homogeneous linear constraints of the form $\sum_{j=1}^{m} c_j\, x_j = 0$ (where $c_1, c_2, \ldots, c_m$ are real numbers) in addition to the requirement that $\sum_{j=1}^{m} x_j = 0$. Accounting for any additional constraints is entirely analogous to the procedure detailed above for the particular constraint that $\sum_{j=1}^{m} x_j = 0$. 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 Kendall et al. (2009). A detailed treatment is available in Chapter 7.

6.3 Linear combinations of independent χ^2 variates

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 6.3.4 below).

Theorem 6.3.1. *Suppose that m is a positive integer, $X_1, X_2, \ldots, X_m$ are i.i.d. Gaussian random variables of zero mean and unit variance, and $\sigma_1, \sigma_2, \ldots, \sigma_m$ are positive real numbers. Suppose in addition that X is the random variable*

$$X = \sum_{j=1}^{m} |\sigma_j\, X_j|^2. \tag{6.7}$$

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

$$F(x) = \int_0^\infty \operatorname{Im}\left(\frac{e^{1-t}\, e^{it\sqrt{m}}}{\pi\left(t - \frac{1}{1-i\sqrt{m}}\right)\prod_{j=1}^{m}\sqrt{1 - 2(t-1)\sigma_j^2/x + 2it\sigma_j^2\sqrt{m}/x}}\right) dt \tag{6.8}$$

for any positive real number x, and $F(x) = 0$ for any nonpositive real number x. The square roots in (6.8) denote the principal branch, and Im *takes the imaginary part.*

Proof. For any $j = 1, 2, \ldots, m$, the characteristic function of $|X_j|^2$ is

$$\varphi_1(t) = \frac{1}{\sqrt{1 - 2it}}, \tag{6.9}$$

using the principal branch of the square root. By the independence of $X_1, X_2, \ldots, X_m$, the characteristic function of the random variable X defined in (6.7) is therefore

$$\varphi(t) = \prod_{j=1}^{m} \varphi_1(t\sigma_j^2) = \frac{1}{\prod_{j=1}^{m} \sqrt{1 - 2it\sigma_j^2}}. \tag{6.10}$$

The probability density function of X is therefore

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx}\, \varphi(t)\, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itx}}{\prod_{j=1}^{m} \sqrt{1 - 2it\sigma_j^2}}\, dt \tag{6.11}$$

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

$$F(x) = \int_{-\infty}^{x} f(y)\, dy = \frac{1}{2} + \frac{i}{2\pi} \operatorname{PV}\!\int_{-\infty}^{\infty} \frac{e^{-itx}}{t \prod_{j=1}^{m} \sqrt{1 - 2it\sigma_j^2}}\, dt \tag{6.12}$$

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 $F(x)$ 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 (6.12) yields that the cdf is

$$F(x) = \frac{1}{2} + \frac{i}{2\pi} \operatorname{PV}\!\int_{-\infty}^{\infty} \frac{e^{-it}}{t \prod_{j=1}^{m} \sqrt{1 - 2it\sigma_j^2/x}}\, dt \tag{6.13}$$

for any positive real number x, where again PV denotes the principal value. The branch cuts for the integrand in (6.13) are all on the lower half of the imaginary axis.

Though the integration in (6.13) is along $(-\infty, \infty)$, we may shift contours and instead integrate along the rays

$$\left\{(-\sqrt{m} - i)t + i \;:\; t \in (0, \infty)\right\} \tag{6.14}$$

and

$$\left\{(\sqrt{m} - i)t + i \;:\; t \in (0, \infty)\right\}, \tag{6.15}$$

obtaining from (6.13) that

$$\begin{aligned} F(x) = \frac{i}{2\pi} \int_0^{\infty} \Bigg(&\frac{e^{1-t}\, e^{-it\sqrt{m}}}{\left(t - \frac{1}{1+i\sqrt{m}}\right) \prod_{j=1}^{m} \sqrt{1 - 2(t-1)\sigma_j^2/x - 2it\sigma_j^2\sqrt{m}/x}} \\ &- \frac{e^{1-t}\, e^{it\sqrt{m}}}{\left(t - \frac{1}{1-i\sqrt{m}}\right) \prod_{j=1}^{m} \sqrt{1 - 2(t-1)\sigma_j^2/x + 2it\sigma_j^2\sqrt{m}/x}} \Bigg)\, dt \end{aligned} \tag{6.16}$$

for any positive real number x. Combining (6.16) and the definition of “Im” yields (6.8). □

Remark 6.3.2. We chose the contours (6.14) and (6.15) so that the absolute value of the expression under the square root in (6.8) is greater than $\sqrt{m/(m+1)}$. Therefore,

$$\left|\prod_{j=1}^{m} \sqrt{1 - 2(t-1)\sigma_j^2/x + 2it\sigma_j^2\sqrt{m}/x}\right| > \left(\frac{m}{m+1}\right)^{m/4} > \frac{1}{e^{1/4}} \tag{6.17}$$

for any $t \in (0, \infty)$ and any $x \in (0, \infty)$. Thus, the integrand in (6.8) is never large for $t \in (0, \infty)$.

Remark 6.3.3. The integrand in (6.8) decays exponentially fast, at a rate independent of the values of σ_1, σ_2, …, σ_m, and x (see the preceding remark).

Remark 6.3.4. An efficient means of evaluating (6.8) numerically is to employ adaptive Gaussian quadratures; see, for example, Section 4.7 of Press et al. (2007). To attain double-precision accuracy (roughly 15-digit precision), the domain of integration for t in (6.8) 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 6.3.5. For a similar, more general approach, see Rice (1980). For alternative approaches, see Duchesne and de Micheaux (2010). 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 m of bins and on the precision ε of computations, not on the values of σ_1, σ_2, …, σ_m, or x. Indeed, it is easy to see that the numerical evaluation of (6.8) theoretically requires $\mathcal{O}(m \ln^2(\sqrt{m}/\varepsilon))$ quadrature nodes: the denominator of the integrand in (6.8) cannot oscillate more than $m+1$ times (once for each "pole") as t ranges from 0 to ∞, while the numerator of the integrand cannot oscillate more than $\sqrt{m}\,\ln(2\sqrt{m}/\varepsilon)$ times as t ranges from 0 to $\ln(2\sqrt{m}/\varepsilon)$; furthermore, the domain of integration for t in (6.8) need be only $(0,\, \ln(2\sqrt{m}/\varepsilon))$ 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 6.5 below. Also, the observed performance is similar when subtracting the imaginary unit i from the contours (6.14) and (6.15).

6.4 Numerical method

An efficient method for calculating the P-values in the limit of large numbers of draws proceeds as follows. Given draws from any distribution — not necessarily from the specified model — we can form the associated statistic X defined in (6.2) and (6.1); in the limit of large numbers of draws, the P-value that the draws do not arise from the model is then just one minus the cumulative distribution function F in (6.8) evaluated at $x = X$, with σ_j^2 in (6.8) being the reciprocals of the positive eigenvalues of the matrix B defined in (6.6) — after all, $F(x)$ is then the probability that x is greater than the sum of the squares of independent centered Gaussian random variables whose variances are the reciprocals of the positive eigenvalues of B. Remark 6.3.4 above describes an efficient means of evaluating $F(x)$ numerically.

6.5 Numerical examples

This section illustrates the performance of the method of Section 6.4, via numerical examples.

We plot the complementary cumulative distribution function of the square of the root-mean-square statistic whose probability distribution is determined in Section 6.2, in the limit of large numbers of draws. This is the distribution of the statistic X defined in (6.2) when the observations used to form X arise as i.i.d. draws from the same distribution $p_0^{(1)}$, $p_0^{(2)}$, …, $p_0^{(m)}$ used in (6.1) for defining X. In order to evaluate the cumulative distribution function (cdf) F, we apply adaptive Gaussian quadratures to the integral in (6.8) as described in Section 6.3, obtaining σ_j in (6.8) via the algorithm described in Section 6.2.

In applications to goodness-of-fit testing, if the statistic X from (6.2) takes on a value x, then the P-value for the draws to arise from the model distribution is one minus the cdf F in (6.8) evaluated at x. Figures 6.1 and 6.2 plot the P-value $(1-F(x))$ versus x for six example model distributions (examples a, b, c, d, e, f). Table 6.3 provides formulae for the model distributions used in the six examples. Tables 6.1 and 6.2 summarize the computational costs required to attain at least 9-digit absolute accuracy for the plots in Figures 6.1 and 6.2, respectively. Each plot displays $1 - F(x)$ at 100 values for x. Figure 6.2 focuses on the tails of the distributions, corresponding to suitably low P-values.

The following list describes the headings of the tables:

- m is the number of bins in Section 2 ($p_0^{(1)}$, $p_0^{(2)}$, …, $p_0^{(m)}$ are the probabilities of drawing the corresponding bins under the specified model distribution).

- q is the maximum number of quadrature nodes required in any of the 100 evaluations of $1 - F(x)$ displayed in each plot of Figures 6.1 and 6.2.

- t is the total number of seconds required to perform the quadratures for all 100 evaluations of $1 - F(x)$ displayed in each plot of Figures 6.1 and 6.2.

- $p_0^{(j)}$ is the probability associated with bin j ($j = 1, 2, \ldots, m$) in Section 6.2. The constants $C_{(\mathrm{a})}$, $C_{(\mathrm{b})}$, $C_{(\mathrm{c})}$, $C_{(\mathrm{d})}$, $C_{(\mathrm{e})}$, $C_{(\mathrm{f})}$ in Table 6.3 are the positive real numbers chosen such that $\sum_{j=1}^{m} p_0^{(j)} = 1$. 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 (6.6) using the Jacobi algorithm (see, for example, Chapter 8 of Golub and Van Loan (1996)), not taking advantage of Remark 6.2.2; explicitly forming the entries of the matrix B defined in (6.6) can incur a numerical error of at most the machine precision (about 2E–16) times $\max_{1 \le j \le m} p_0^{(j)} / \min_{1 \le j \le m} p_0^{(j)}$, yielding 9-digit accuracy or better for all our examples. Higher precision is possible via the interlacing properties of eigenvalues, following Gu and Eisenstat (1994). 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.

Table 6.1: Values for Figure 6.1

	m	q	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

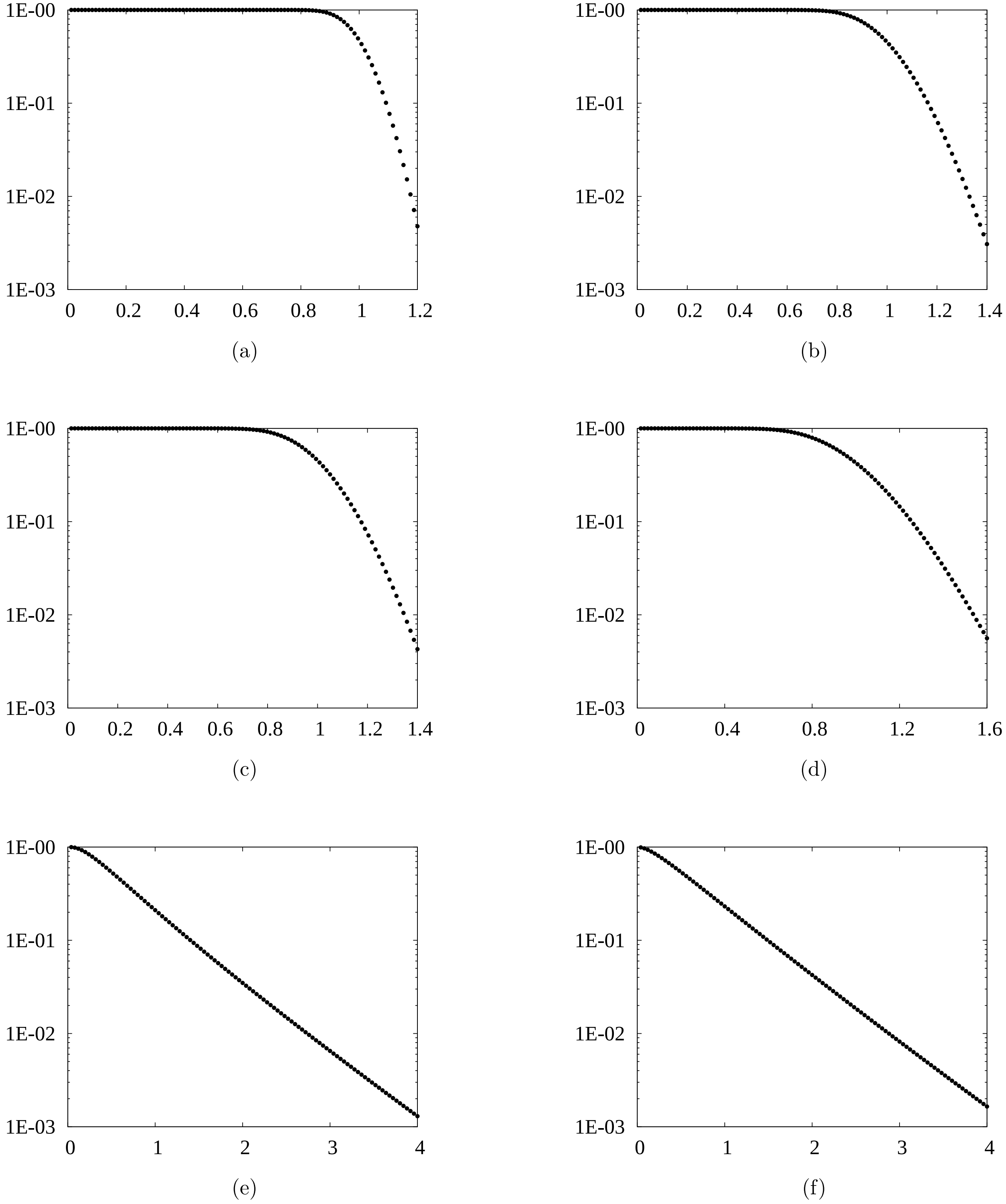
Table 6.2: Values for Figure 6.2

	m	q	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

Remark 6.5.1. It is easy to compute the P-values (in the limit of large numbers of draws) for a distribution having infinitely many bins, but only to any arbitrary accuracy that is greater than the machine precision. Specifically, given an extremely small positive real number ε, we would retain the smallest possible number of bins whose associated probabilities $p_0^{(1)}$, $p_0^{(2)}$, …, $p_0^{(m)}$ satisfy $p_0^{(1)} + p_0^{(2)} + \cdots + p_0^{(m)} \geq 1 - \varepsilon$, and then proceed with the computation as if these finitely many were the only bins. (Needless to say, if the fraction of the experimental draws falling outside the finitely many retained bins is significantly greater than ε, then we can be highly confident that the draws did not arise from the model.)

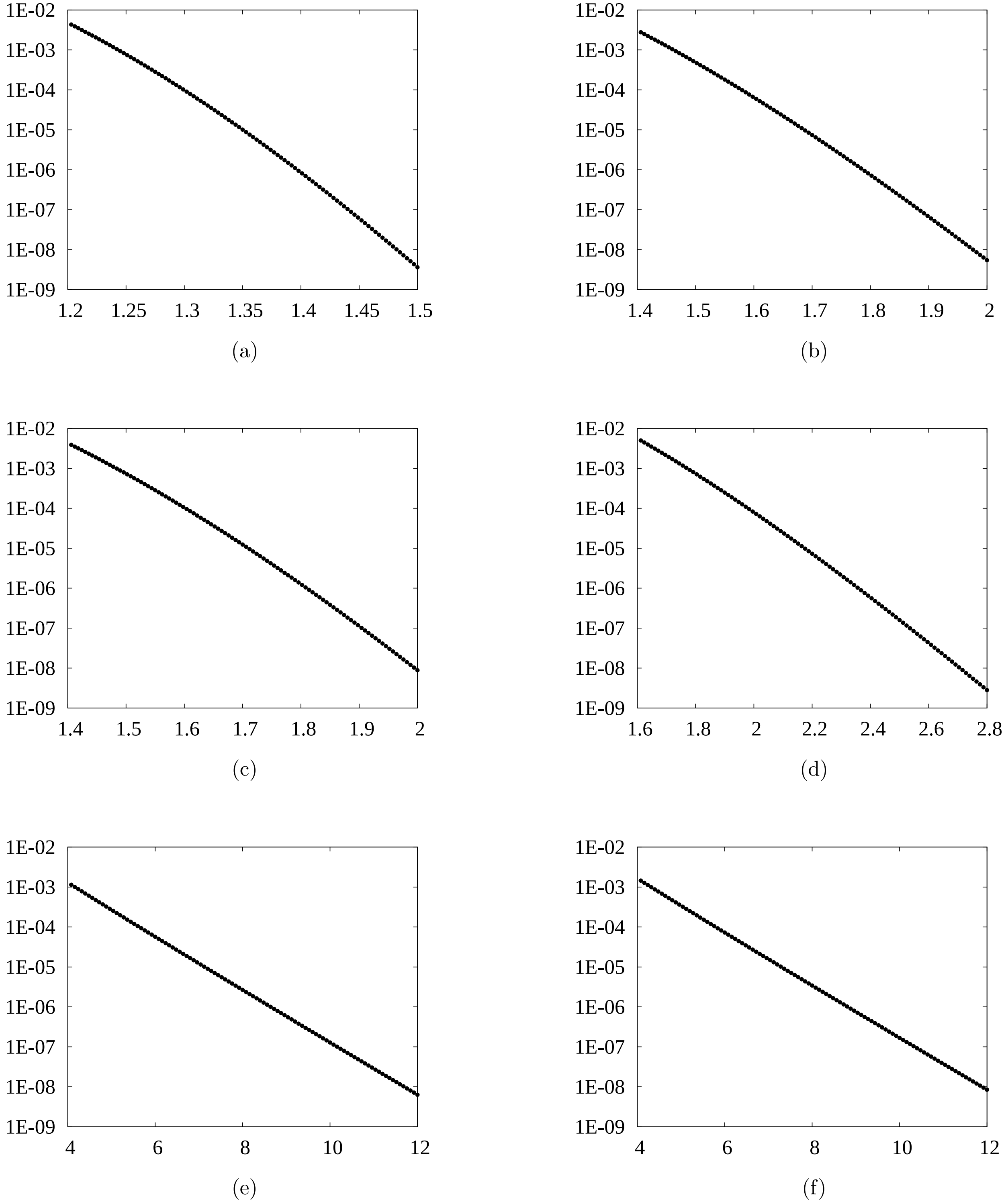
Table 6.3: Values for both Figure 6.1 and Figure 6.2

	m	$p_0^{(j)}$
(a)	500	$C_{(\mathrm{a})} \cdot (300 + j)^{-2}$
(b)	250	$C_{(\mathrm{b})} \cdot (260 - j)^3$
(c)	100	$C_{(\mathrm{c})} \cdot \lfloor (40 + j)/40 \rfloor^{-1/6}$
(d)	50	$C_{(\mathrm{d})} \cdot (1/2 + \ln\lfloor (61 - j)/10 \rfloor)$
(e)	25	$C_{(\mathrm{e})} \cdot \exp(-5j/8)$
(f)	10	$C_{(\mathrm{f})} \cdot \exp(-(j-1)^2/6)$



(a) (b) (c) (d) (e) (f)

Figure 6.1: The vertical axis is $1 - F(x)$ from (6.8); the horizontal axis is x.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6.2: The vertical axis is $1 - F(x)$ from (6.8); the horizontal axis is x.

Chapter 7

Asymptotic P-values with nuisance parameters

This chapter is an extension of the previous; unlike in the previous chapter, the models in the present chapter involve parameter estimation. Once again we provide efficient black-box algorithms for calculating, in the limit of large numbers of draws, the asymptotic P-values for the root-mean-square statistic discussed in the previous chapters. As observed in the previous chapters, computing the exact P-values via Monte Carlo simulation is often feasible, too.

7.1 Recap

A basic task in statistics is to ascertain whether a given set of observations did not arise as independent and identically distributed (i.i.d.) draws from a member of a specified family of probability distributions (the specified family is known as the "model"). 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"). Chapter 6 treats the special case in which the "family" of distributions constituting the model in fact consists of a single, fully specified probability distribution. The present chapter focuses on models parameterized with a single scalar; our techniques extend straightforwardly to any parameterization with multiple scalars (or, equivalently, to any parameterization with a vector).

A natural approach to ascertaining whether a given set of draws does not come from the model uses a root-mean-square statistic. To construct this statistic, we estimate both the parameter and the probability distribution over the bins using the given draws, and then measure the root-mean-square difference between this empirical distribution and the model distribution corresponding to the estimated parameter (for details, see, for example, Rao, 2002; Varadhan et al., 1974, page 123; or Section 7.2 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 did not arise from the model.

To quantify "large" and "confident," let us denote by x the value of the root-mean-square for the given draws; let us denote by X the root-mean-square statistic constructed for different draws that definitely were in fact taken i.i.d. from the model. The P-value is then defined to be the probability that $X \geq x$ (viewing X — but not x — as a random variable).

Unfortunately, the P-values for the simple root-mean-square are different for different models. In order to avoid this seeming inconvenience (at least asymptotically), one may weight the average in the root-mean-square by the reciprocals of the model probabilities associated with the various bins, obtaining the classic χ^2 statistic of Pearson (1900); see Remark 7.2.1 below. However, with the now widespread availability of computers, direct use of the simple root-mean-square statistic has become feasible (and actually turns out to be very convenient). The present chapter provides efficient black-box algorithms for computing the P-values for any model with a smooth parameterization, in the limit of large numbers of draws. Calculating P-values for small numbers of draws via Monte-Carlo simulations can also be practical, as illustrated in previous chapters.

The remainder of the present chapter has the following structure: Section 7.2 details the simple statistic discussed above, expressing the asymptotic P-values for the associated goodness-of-fit test in a form suitable for rapid computation. Section 7.3 summarizes the method for computing the P-values of the root-mean-square statistic. Section 7.4 applies the method to several examples.

7.2 The sum-of-squares statistic

This section details the simple root-mean-square statistic discussed briefly in Section 7.1, determining 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 P-values (in the limit of large numbers of draws): given a value x for the root-mean-square statistic constructed from observations coming from an unknown distribution, and given the value of the maximum-likelihood estimate $\hat{\theta}$ for the parameter of the distribution, the P-value that the observations did not arise as i.i.d. draws from the specified model is the probability that the root-mean-square statistic is greater than or equal to x when constructed from i.i.d. draws that do come from the model distribution associated with the parameter $\hat{\theta}$. (Please note that the definition in (7.2) and (7.3) below of the simple statistic involves the maximum-likelihood estimate $\hat{\theta}$. Maximum likelihood is the canonical method for parameter estimation, and is the focus of the present chapter. See formulae (7.13) and (7.14) below regarding likelihood and maximum-likelihood estimation.)

To begin, we set notation and form the goodness-of-fit statistic X to be analyzed. Given m bins, numbered 1, 2, …, m, we denote by $p_0^{(1)}(\theta)$, $p_0^{(2)}(\theta)$, …, $p_0^{(m)}(\theta)$ the probabilities associated with the respective bins under the specified model, where θ is a real number parameterizing the model; of course,

$$\sum_{j=1}^{m} p_0^{(j)}(\theta) = 1 \tag{7.1}$$

for any parameter θ. In order to obtain a draw conforming to the model for a particular value of θ, we select at random one of the m bins, with probabilities $p_0^{(1)}(\theta)$, $p_0^{(2)}(\theta)$, …,

$p_0^{(m)}(\theta)$. We perform this selection independently n times. For $j = 1, 2, \dots, m$, we denote by $Y^{(j)}$ the fraction of times that we choose bin j (that is, $Y^{(j)}$ is the number of times that we choose bin j, divided by n); obviously, $\sum_{j=1}^m Y^{(j)} = 1$. We define X_j to be $\sqrt{n}$ times the difference of $Y^{(j)}$ from its expected value using the maximum-likelihood estimate $\hat{\theta}$ of the actual parameter θ, that is,

$$X_j = \sqrt{n}\,(Y^{(j)} - p_0^{(j)}(\hat{\theta})) \tag{7.2}$$

for $j = 1, 2, \dots, m$. Finally, we form the statistic

$$X = \sum_{j=1}^{m} X_j^2, \tag{7.3}$$

and now determine its distribution in the limit that the number n of draws is large. The root-mean-square statistic $\sqrt{\sum_{j=1}^m (nY^{(j)} - np_0^{(j)}(\hat{\theta}))^2/n}$ is the square root of X. As the square root is a monotonically increasing function, the P-values are the same whether determined via X or via $\sqrt{X}$; for convenience, we focus on X below.

Remark 7.2.1. The classic χ^2 test for goodness-of-fit of Pearson (1900) replaces (7.3) with the statistic

$$\chi^2 = \sum_{j=1}^{m} \frac{X_j^2}{p_0^{(j)}(\hat{\theta})}, \tag{7.4}$$

where $X_1, X_2, \dots, X_m$ are the same as in (7.2) and (7.3), and $\hat{\theta}$ is the maximum-likelihood estimate of the parameter.

For definiteness, we will be assuming that $p_0^{(1)}, p_0^{(2)}, \dots, p_0^{(m)}$ are differentiable as functions of the parameter θ, that the maximum of the likelihood occurs in the interior of the domain for θ, that the maximum-likelihood estimate $\hat{\theta}$ is almost surely the correct value for the actual parameter θ as $n \to \infty$, and that the variance of $\hat{\theta}$ tends to zero as $n \to \infty$ (thus $\hat{\theta}$ is not "random" in the limit of large numbers of draws). As detailed, for example, by Moore and Spruill (1975) and by Kendall et al. (2009) in a chapter on goodness-of-fit (see also Remark 7.2.3 below), the multivariate central limit theorem then shows that the joint distribution of $X_1, X_2, \dots, X_m$ converges in distribution as $n \to \infty$, with the limiting generalized probability density proportional to

$$\exp\left(-\sum_{j=1}^{m} \frac{x_j^2}{2p_0^{(j)}(\hat{\theta})}\right) \cdot \delta\left(\sum_{j=1}^{m} x_j\right) \cdot \delta\left(\sum_{j=1}^{m} x_j \frac{d}{d\theta} \ln(p_0^{(j)}(\theta))\bigg|_{\theta=\hat{\theta}}\right), \tag{7.5}$$

where δ is the Dirac delta, and $\hat{\theta}$ is the maximum-likelihood estimate of the parameter.

The generalized probability density in (7.5) is a centered multivariate Gaussian distribution concentrated on the intersection of two hyperplanes that both pass through the origin (the intersection of the hyperplanes consists of all the points such that $\sum_{j=1}^m x_j = 0$ and $\sum_{j=1}^m x_j \frac{d}{d\theta} \ln(p_0^{(j)}(\theta))\big|_{\theta=\hat{\theta}} = 0$); the restriction of the generalized probability density (7.5) to the intersection of the hyperplanes is also a centered multivariate Gaussian. Thus, the distribution of X defined in (7.3) converges as $n \to \infty$ to the distribution of the sum of the squares of $m-2$ 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, Chapter 25 of Kendall et al., 2009). Given these variances, Remark 6.3.4 describes an efficient algorithm for computing the probability that the associated sum of squares is greater than or equal to any particular value; this probability is the desired P-value, in the limit of large numbers of draws. For a detailed discussion, see Section 7.3 below.

To compute the variances of the restricted multivariate Gaussian distribution along its principal axes, we perform the following four steps:

1. Form an $m \times 2$ matrix H whose columns both include a vector that is normal to the hyperplane consisting of the points $(x_1, x_2, \dots, x_m)$ such that
$$\sum_{j=1}^{m} x_j = 0, \tag{7.6}$$
and also include a vector that is normal to the hyperplane consisting of the points $(x_1, x_2, \dots, x_m)$ such that
$$\sum_{j=1}^{m} x_j \, \frac{d}{d\theta} \ln(p_0^{(j)}(\theta)) \bigg|_{\theta=\hat{\theta}} = 0, \tag{7.7}$$
where $\hat{\theta}$ is the maximum-likelihood estimate of the parameter. For example, we can take the entries of H to be
$$H_{j,k} = \begin{cases} 1, & k = 1 \\ \frac{d}{d\theta} \ln(p_0^{(j)}(\theta))\big|_{\theta=\hat{\theta}}, & k = 2 \end{cases} \tag{7.8}$$
for $j = 1, 2, \dots, m$ and $k = 1, 2$, where again $\hat{\theta}$ is the maximum-likelihood estimate of the parameter.

2. Form an orthonormal basis for the column space of H, by constructing a pivoted QR decomposition
$$H_{m\times 2} = Q_{m\times 2} \cdot R_{2\times 2} \cdot \Pi_{2\times 2}, \tag{7.9}$$
where the columns of Q are orthonormal, R is upper-triangular, and Π is a permutation matrix. (See, for example, Chapter 5 of Golub and Van Loan, 1996, for details on the construction of such a pivoted QR decomposition.)

3. Form the $m \times m$ diagonal matrix D with the entries
$$D_{j,k} = \begin{cases} 1/p_0^{(j)}(\hat{\theta}), & j = k \\ 0, & j \neq k \end{cases} \tag{7.10}$$
for $j, k = 1, 2, \dots, m$, where $\hat{\theta}$ is the maximum-likelihood estimate of the parameter. Then, multiply D from both the left and the right by the orthogonal projection (namely, $I - QQ^\top$) onto the intersection of the hyperplanes consisting of the points satisfying (7.6) and (7.7), obtaining the $m \times m$ matrix
$$B = (I - QQ^\top)\, D\, (I - QQ^\top), \tag{7.11}$$
where I is the $m \times m$ identity matrix.

4. Find the eigenvalues of the self-adjoint matrix B defined in (7.11). By construction, exactly two of the eigenvalues of B are zeros. The other eigenvalues of B are the reciprocals of the desired variances of the restricted multivariate Gaussian distribution along its principal axes.

Remark 7.2.2. The $m \times m$ matrix B defined in (7.11) is the sum of a diagonal matrix and a low-rank matrix. The methods of Gu and Eisenstat (1994, 1995) for computing the eigenvalues of such a matrix B require only either $\mathcal{O}(m^2)$ or $\mathcal{O}(m)$ floating-point operations. Note that the $\mathcal{O}(m^2)$ methods of Gu and Eisenstat (1994, 1995) are more efficient than the $\mathcal{O}(m)$ procedure of Gu and Eisenstat (1995), unless m is impractically large.

Remark 7.2.3. Under appropriate regularity conditions, it is straightforward to derive the homogeneous linear constraint — analogous to (7.7) — that

$$\sum_{j=1}^{m} X_j \frac{d}{d\theta} \ln(p_0^{(j)}(\theta))\bigg|_{\theta=\hat{\theta}} = 0, \tag{7.12}$$

where $\hat{\theta}$ is the maximum-likelihood estimator. The following is a sketch of the proof of (7.12).

To determine the maximum-likelihood estimate $\hat{\theta}$, we consider the likelihood, namely the multinomial distribution

$$L(y^{(1)}, y^{(2)}, \dots, y^{(m)}, \theta) = n! \prod_{j=1}^{m} \frac{(p_0^{(j)}(\theta))^{ny^{(j)}}}{(ny^{(j)})!}. \tag{7.13}$$

Maximizing (7.13) defines $\hat{\theta}$ via the formula

$$0 = \frac{\partial}{\partial\theta} \ln(L(Y^{(1)}, Y^{(2)}, \dots, Y^{(m)}, \theta))\bigg|_{\theta=\hat{\theta}} = \sum_{j=1}^{m} nY^{(j)} \frac{d}{d\theta} \ln(p_0^{(j)}(\theta))\bigg|_{\theta=\hat{\theta}}. \tag{7.14}$$

It follows from (7.1) that

$$\sum_{j=1}^{m} \frac{d}{d\theta} p_0^{(j)}(\theta) = 0 \tag{7.15}$$

for any parameter θ, in particular for $\theta = \hat{\theta}$. Combining (7.14) and (7.15) yields that

$$\sum_{j=1}^{m} (Y^{(j)} - p_0^{(j)}(\hat{\theta})) \frac{d}{d\theta} \ln(p_0^{(j)}(\theta))\bigg|_{\theta=\hat{\theta}} = 0. \tag{7.16}$$

Combining (7.16) and (7.2) yields (7.12), as desired.

7.3 Numerical method

An efficient method for calculating the P-values in the limit of large numbers of draws proceeds as follows. Given draws from any distribution — not necessarily from the model — we can form the associated statistic X defined in (7.3) and (7.2); in the limit of large numbers of draws, the P-value that the draws do not arise from the model is then just one minus the cumulative distribution function $F(x)$ in (6.8) evaluated at $x = X$, with σ_j^2 in (6.8) obtained via Step 4 of the algorithm of Section 7.2 (after all, $F(x)$ is the probability that x is greater than the sum of the squares of independent centered Gaussian random variables whose variances are given by Step 4 above). Remark 6.3.4 describes an efficient means of evaluating $F(x)$ numerically.

7.4 Numerical examples

This section illustrates the performance of the algorithm of Section 7.3 via several numerical examples.

Figure 7.1 and Table 7.1 correspond to the first example. The model distribution for the first example has 4 bins, with the probabilities indicated in Table 7.4. We will detail the interpretation of the figures and tables shortly.

Figure 7.2 and Table 7.2 correspond to the second example. The model for the second example is the Zipf distribution on 100 bins. The row for Figure/Table 7.2 in Table 7.4 provides a definition of the Zipf distribution.

Figure 7.3 and Table 7.3 correspond to the third example. The model for the third example is the standard Poisson distribution. The row for Figure/Table 7.3 in Table 7.4 provides a definition of the Poisson distribution.

To validate our algorithms, we conduct computational simulations (see Section 6.5 for a complementary approach). In every simulation, we choose the number n of draws to be a very large number, namely $n = 100{,}000$. (The algorithms of the present chapter concern the limit as $n \to \infty$.) Part (a) of the examples uses $\ell = 1{,}000$ simulations; part (b) of the examples uses $\ell = 10{,}000$ simulations. The convergence (as ℓ increases) of the plotted points to the straight line of unit slope through the origin provides numerical validation of our algorithms, for the following reasons.

To create the plots, we run ℓ simulations, each taking $n = 100{,}000$ i.i.d. draws from the model distribution with the specified parameter θ. For each simulation, we compute the statistic X defined in (7.3), forming $Y^{(1)}$, $Y^{(2)}$, …, $Y^{(m)}$ and $\hat{\theta}$ needed in (7.2) and (7.3) using the generated draws. We then compute the asymptotic P-value associated with each of these values for X, as described in Section 7.3, and sort the resulting P-values. These sorted results are the vertical coordinates of the points in the plots; the horizontal coordinates are the equispaced numbers $1/(2\ell)$, $3/(2\ell)$, $5/(2\ell)$, …, $(2\ell - 1)/(2\ell)$.

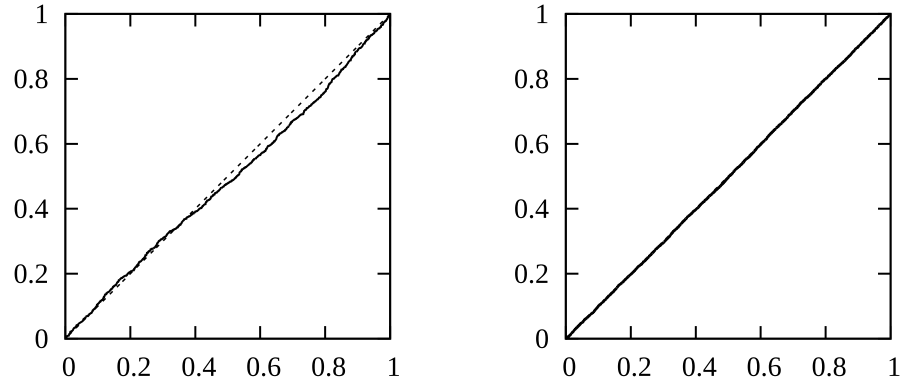
As the number ℓ of simulations increases, and insofar as the number n of draws is very large, the plotted points should converge to the straight line through the origin of slope 1 (and, indeed, our experiments demonstrate this). The dotted line in each plot is the straight line through the origin of slope 1. The trials converge correctly: the root-mean-square statistics for about $\alpha\%$ of the simulations should have P-values of $\alpha\%$ or less, for every $\alpha \in (0, 100)$; in the limit that both the number n of draws and the number ℓ of simulations are large, the computed P-values for exactly $\alpha\%$ of the simulations should be less than or equal to $\alpha\%$ (this follows from the definition of P-values; it also follows from the fact that the P-values for the statistic X are given by its cumulative distribution function F, and from the fact that $F(X)$ is uniformly distributed over $[0, 1]$ for any random variable X distributed according to a continuous cumulative distribution function F).

The following list describes the headings of the tables:

- ℓ is the number of simulations conducted in generating the associated plot.
- θ is the parameter for the model distribution used in generating the i.i.d. draws.
- m is the number of bins in the model (see Remark 7.4.3 with regard to the Poisson distribution of the third example).
- q is the maximum number of quadrature nodes required to evaluate the P-value for any of the ℓ root-mean-square statistics produced by the simulations.
- t is the total number of seconds required to perform the quadratures for evaluating the P-values for all ℓ of the root-mean-square statistics produced by the simulations.
- s is the total number of seconds required to perform all ℓ simulations.
- $p_0^{(j)}(\theta)$ is the probability associated with bin j ($j = 1, 2, \dots, m$), as a function of the parameter θ.
- $\hat{\theta}(Y^{(1)}, Y^{(2)}, \dots, Y^{(m)})$ is the maximum-likelihood estimate of the parameter θ, as a function of the fractions $Y^{(1)}$, $Y^{(2)}$, $\dots$, $Y^{(m)}$ of the draws in the respective bins (a detailed definition of $Y^{(1)}$, $Y^{(2)}$, $\dots$, $Y^{(m)}$ is available in Section 7.2).

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 (7.11) using the Jacobi algorithm (see, for example, Chapter 8 of Golub and Van Loan, 1996), not taking advantage of Remark 7.2.2. We generated the pseudorandom numbers used in the simulations via (Mitchell-Moore-Brent-Knuth) lagged Fibonacci sequences (see, for example, Section 7.1.5 of Press et al., 2007).

Remark 7.4.1. It is easy to compute the P-values (in the limit of large numbers of draws) for a distribution having infinitely many bins, but only to any arbitrary accuracy that is greater than the machine precision. Specifically, given a fully specified model distribution and an extremely small positive real number ε, we would retain the smallest possible number of bins whose associated probabilities $p_0^{(1)}$, $p_0^{(2)}$, $\dots$, $p_0^{(m)}$ satisfy $p_0^{(1)} + p_0^{(2)} + \dots + p_0^{(m)} \geq 1 - \varepsilon$, and then proceed with the computation as if these finitely many were the only bins. When there is a parameter θ being estimated, we observe that the maximum-likelihood estimate $\hat{\theta}$ typically has variance zero and is almost surely correct in the limit of large numbers of draws; thus, as before, we may retain the smallest possible number of bins whose associated probabilities $p_0^{(1)}(\hat{\theta})$, $p_0^{(2)}(\hat{\theta})$, $\dots$, $p_0^{(m)}(\hat{\theta})$ satisfy $p_0^{(1)}(\hat{\theta}) + p_0^{(2)}(\hat{\theta}) + \dots + p_0^{(m)}(\hat{\theta}) \geq 1 - \varepsilon$, and then proceed with the computation as if these finitely many were the only bins. (Needless to say, if the fraction of the experimental draws falling outside the finitely many retained bins is significantly greater than ε, then we can be highly confident that the draws did not arise from the model.)



(a) (b)

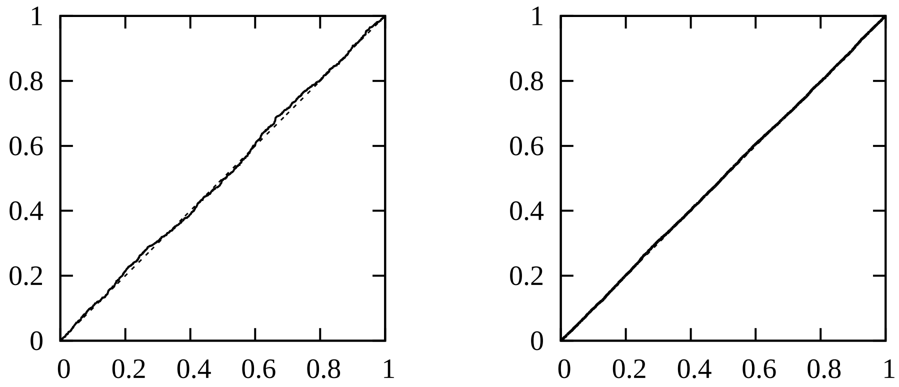
Figure 7.1: 2 × 2 contingency-table/cross-tabulation of Table 7.4

Table 7.1: Values for Figure 7.1

	ℓ	θ	m	q	t	s
(a)	10^3	.03	4	190	.43E0	.44E1
(b)	10^4	.03	4	190	.43E1	.45E2

Remark 7.4.2. For the second example (the Zipf distribution), we computed the maximum-likelihood estimate $\hat{\theta}$ from the data $Y^{(1)}$, $Y^{(2)}$, ..., $Y_0^{(m)}$ by finding the zero of the function $g(\hat{\theta}) = f(\hat{\theta}) - \sum_{j=1}^{m} Y^{(j)} \ln(j) = 0$, where f is the same as in Table 7.4, namely $f(\hat{\theta}) = \left(\sum_{j=1}^{m} j^{-\hat{\theta}} \ln(j)\right) \Big/ \left(\sum_{j=1}^{m} j^{-\hat{\theta}}\right)$. We evaluated the zero $\hat{\theta}$ numerically, via bisection (see, for example, Chapter 9 of Press et al., 2007).

Remark 7.4.3. For the third example (the Poisson distribution), we employed Remark 7.4.1, with $\varepsilon = 10^{-8}$.

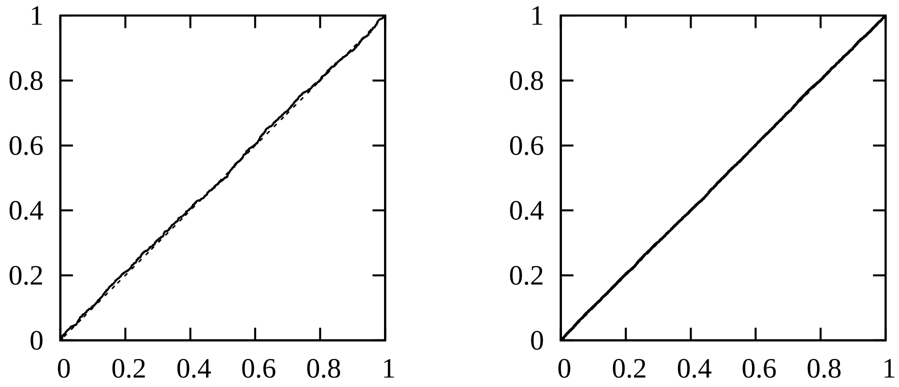


(a) (b)

Figure 7.2: Zipf distribution of Table 7.4

Table 7.2: Values for Figure 7.2

	ℓ	θ	m	q	t	s
(a)	10^3	1	100	350	.92E1	.13E2
(b)	10^4	1	100	390	.11E3	.13E3



(a) (b)

Figure 7.3: Poisson distribution of Table 7.4

Table 7.3: Values for Figure 7.3

	ℓ	θ	m	q	t	s
(a)	10^3	10.3	36	290	.37E1	.86E1
(b)	10^4	10.3	36	330	.37E2	.86E2

Table 7.4: Values for Figures 7.1–7.3 and Tables 7.1–7.3

Fig./Table #	$p_0^{(j)}(\theta)$	$\hat{\theta}(Y^{(1)}, Y^{(2)}, \ldots, Y^{(m)})$
Fig./Table 7.1	$p_0^{(1)} = .04{\cdot}\theta, p_0^{(2)} = .04(1{-}\theta)$ $p_0^{(3)} = .96{\cdot}\theta, p_0^{(4)} = .96(1{-}\theta)$	$\hat{\theta} = Y^{(1)} + Y^{(3)}$
Fig./Table 7.2	$p_0^{(j)} = j^{-\theta} \big/ \sum_{i=1}^{m} i^{-\theta}$	$\hat{\theta} = f^{-1}\big(\sum_{j=1}^{m} Y^{(j)} \ln(j)\big)$, $f(\hat{\theta}) = \Big(\sum_{j=1}^{m} j^{-\hat{\theta}} \ln(j)\Big) \Big/ \Big(\sum_{j=1}^{m} j^{-\hat{\theta}}\Big)$
Fig./Table 7.3	$p_0^{(j)} = e^{-\theta}\theta^{j-1}/(j-1)!$	$\hat{\theta} = \sum_{j=1}^{\infty}(j-1)\, Y^{(j)}$

Chapter 8

Asymptotic power

A natural yet unconventional test for goodness-of-fit measures the discrepancy between the model and empirical distributions via their root-mean-square distance (or, equivalently, via its square). The present chapter characterizes the statistical power of such a test against a family of alternative distributions, in the limit that the number of observations is large, with every alternative departing from the model in the same direction. Specifically, the chapter provides an efficient numerical method for evaluating the cumulative distribution function (cdf) of the square of the root-mean-square distance between the model and empirical distributions under the alternatives, in the limit that the number of observations is large. The chapter illustrates the scheme by plotting the asymptotic power (as a function of the significance level) for several examples.

8.1 Recap

Given n observations, each falling in one of m bins, we would like to test if these observations are consistent with having arisen as independent and identically distributed (i.i.d.) draws from a specified probability distribution p_0 over the m bins (p_0 is known as the "model"). A simple statistic measuring the deviation between p_0 and the observations is the square x_a of the root-sum-square distance between the actually observed distribution of the draws and the expected distribution p_0, that is,

$$x_a = \sum_{j=1}^{m} (y_a^{(j)} - p_0^{(j)})^2, \tag{8.1}$$

where $y_a^{(1)}$, $y_a^{(2)}$, ..., $y_a^{(m)}$ are the proportions of the n observations falling in bins 1, 2, ..., m, respectively.

The "P-value" is then defined to be the probability that $X_0 \geq x_a$, where X_0 would be the same as x_a, but constructed from n draws that definitely are taken i.i.d. from p_0, that is,

$$X_0 = \sum_{j=1}^{m} (Y_0^{(j)} - p_0^{(j)})^2, \tag{8.2}$$

where $Y_0^{(1)}$, $Y_0^{(2)}$, …, $Y_0^{(m)}$ are the proportions of n i.i.d. draws from p_0 falling in bins 1, 2, …, m, respectively. When calculating the P-value — the probability that $X_0 \ge x_a$ — we view X_0 as a random variable while viewing x_a as a fixed number. If the P-value is small, then we can be confident that the observed draws were not taken i.i.d. from the model p_0.

To characterize the statistical power of the P-value based on the root-mean-square, we consider n i.i.d. draws from the alternative distribution

$$p_a = p_0 + a/\sqrt{n}, \tag{8.3}$$

where a is a vector whose m entries satisfy $\sum_{j=1}^{m} a^{(j)} = 0$. We thus need to calculate the distribution of the square X_a of the root-sum-square distance,

$$X_a = \sum_{j=1}^{m} (Y_a^{(j)} - p_0^{(j)})^2, \tag{8.4}$$

where $Y_a^{(1)}$, $Y_a^{(2)}$, …, $Y_a^{(m)}$ are the proportions of n i.i.d. draws from p_a falling in bins 1, 2, …, m, respectively. Section 8.4 below provides an efficient method for calculating the cumulative distribution function (cdf) of $n \cdot X_a$ in the limit that the number n of draws is large. Section 8.5 below then describes how to use such a method to plot the cdf of the P-values; this cdf is the same as the statistical power function of the hypothesis test based on the root-mean-square distance (as a function of the significance level). Presenting this method is the principal purpose of the present chapter, complementing the previous chapters, which compare the root-mean-square distance with classical statistics such as χ^2, the log–likelihood-ratio G^2, and other members of the Cressie-Read power-divergence family.

As reviewed, for example, by Kendall et al. (2009) and Rao (2002), $m{\cdot}n{\cdot}X_a$ defined in (8.4) converges in distribution to a noncentral χ^2 in the limit that the number n of draws is large, when the model p_0 is a uniform distribution. When p_0 is nonuniform, $m \cdot n \cdot X_a$ converges in distribution to the sum of the squares of independent Gaussian random variables in the limit that the number n of draws is large, as shown by Moore and Spruill (1975) and reviewed in Section 8.2 below. Section 8.3 provides integral representations for the cdf of the sum of the squares of independent Gaussian random variables and applies suitable quadratures for their numerical evaluation. Section 8.4 summarizes the numerical method obtained by combining Sections 8.2 and 8.3. Section 8.5 summarizes a scheme for plotting the asymptotic power (as a function of the significance level) using the method of Section 8.4. Section 8.6 illustrates the methods via several numerical examples.

The extension to models with nuisance parameters is straightforward, via the techniques of Chapter 7; the present chapter focuses on the simpler case in which the model p_0 is a single, fully specified probability distribution.

8.2 The sum-of-squares statistic

This section states Theorem 8.2.1, which is a special case of Theorem 4.2 of Moore and Spruill (1975). Before stating the theorem, we need to set up some notation. The set-up amounts to an algorithm for computing the real numbers $\sigma^{(1)}$, $\sigma^{(2)}$, …, $\sigma^{(m-1)}$ and $\zeta^{(1)}$, $\zeta^{(2)}$, …, $\zeta^{(m-1)}$ used in Theorem 8.2.1, where m is an integer greater than 1.

First, we aim to define the positive real numbers $\sigma^{(1)}$, $\sigma^{(2)}$, …, $\sigma^{(m-1)}$, given any $m \times 1$ vector p_0 whose entries are all positive. We define D to be the diagonal $m \times m$ matrix

$$D_{j,k} = \begin{cases} \frac{1}{p_0^{(j)}}, & j = k \\ 0, & j \neq k \end{cases} \tag{8.5}$$

for $j, k = 1, 2, \ldots, m$. We define H to be the $m \times m$ matrix

$$H_{j,k} = \begin{cases} 1 - \frac{1}{m}, & j = k \\ -\frac{1}{m}, & j \neq k \end{cases} \tag{8.6}$$

for $j, k = 1, 2, \ldots, m$. Note that H is an orthogonal projector. We define $B = HDH$, so that B is the self-adjoint $m \times m$ matrix

$$B_{j,k} = \begin{cases} \frac{1}{p_0^{(j)}} - \frac{1}{m}\left(\frac{1}{p_0^{(j)}} + \frac{1}{p_0^{(k)}}\right) + \frac{1}{m^2}\sum_{i=1}^{m} \frac{1}{p_0^{(i)}}, & j = k \\ -\frac{1}{m}\left(\frac{1}{p_0^{(j)}} + \frac{1}{p_0^{(k)}}\right) + \frac{1}{m^2}\sum_{i=1}^{m} \frac{1}{p_0^{(i)}}, & j \neq k \end{cases} \tag{8.7}$$

for $j, k = 1, 2, \ldots, m$. As a self-adjoint matrix whose rank is $m - 1$ (after all, $B = HDH$, H is an orthogonal projector whose rank is $m - 1$, and D is a full-rank diagonal matrix), B given in (8.7) has an eigendecomposition

$$B = Q \Lambda Q^\top, \tag{8.8}$$

where Q is a real unitary $m \times m$ matrix and Λ is a diagonal $m \times m$ matrix such that $\Lambda_{m,m} = 0$. Finally, we define the positive real numbers $\sigma^{(1)}$, $\sigma^{(2)}$, …, $\sigma^{(m-1)}$ via the formula

$$(\sigma^{(j)})^2 = 1/\Lambda_{j,j} \tag{8.9}$$

for $j = 1, 2, \ldots, m - 1$, where $\Lambda_{1,1}$, $\Lambda_{2,2}$, …, $\Lambda_{m,m}$ are the diagonal entries of Λ from the eigendecomposition (8.8).

Next, we define the real numbers $\zeta^{(1)}$, $\zeta^{(2)}$, …, $\zeta^{(m-1)}$, given both p_0 and an $m \times 1$ vector a such that $\sum_{j=1}^{m} a^{(j)} = 0$. We define the $(m - 1) \times 1$ vector

$$\eta = \tilde{Q}^\top a, \tag{8.10}$$

where $\tilde{Q}$ is the leftmost $m \times (m - 1)$ block of Q from the eigendecomposition (8.8), that is, $\tilde{Q}$ is the same as Q after deleting the last column of Q. We can then define the real numbers $\zeta^{(1)}$, $\zeta^{(2)}$, …, $\zeta^{(m-1)}$ via the formula

$$\zeta^{(j)} = \eta^{(j)}/\sigma^{(j)} \tag{8.11}$$

for $j = 1, 2, \ldots, m - 1$, where η is defined in (8.10) and σ is defined in (8.9).

With this notation, we can state the following special case of Theorem 4.2 of Moore and Spruill (1975).

Theorem 8.2.1. *Suppose that m is an integer greater than one, p_0 is a probability distribution over m bins (that is, p_0 is an $m \times 1$ vector whose entries are all positive and $\sum_{j=1}^{m} p_0^{(j)} = 1$), a is an $m \times 1$ vector such that $\sum_{j=1}^{m} a^{(j)} = 0$, and $Y_n^{(1)}$, $Y_n^{(2)}$, …, $Y_n^{(m)}$ are the proportions of draws falling in bins 1, 2, …, m, respectively, out of a total of n i.i.d. draws from the probability distribution*

$$p_a = p_0 + a/\sqrt{n}. \tag{8.12}$$

Suppose further that X_n is the random variable

$$X_n = n \sum_{j=1}^{m} (Y_n^{(j)} - p_0^{(j)})^2. \tag{8.13}$$

Then, X_n converges in distribution to the random variable

$$X_\infty = \sum_{j=1}^{m-1} (\sigma^{(j)})^2 \, (Z^{(j)} + \zeta^{(j)})^2 \tag{8.14}$$

as n becomes large, where $Z^{(1)}$, $Z^{(2)}$, …, $Z^{(m-1)}$ are i.i.d. Gaussian random variables of zero mean and unit variance, $\sigma^{(1)}$, $\sigma^{(2)}$, …, $\sigma^{(m-1)}$ are the positive real numbers defined in (8.9), and $\zeta^{(1)}$, $\zeta^{(2)}$, …, $\zeta^{(m-1)}$ are the real numbers defined in (8.11). The values of $\sigma^{(1)}$, $\sigma^{(2)}$, …, $\sigma^{(m-1)}$ do not depend on the vector a; the values of $\zeta^{(1)}$, $\zeta^{(2)}$, …, $\zeta^{(m-1)}$ do depend on a.

Remark 8.2.2. The $m \times m$ matrix B defined in (8.7) is the sum of a diagonal matrix and a low-rank matrix. The methods of Gu and Eisenstat (1994, 1995) for computing the eigenvalues of such a matrix B and computing the result of applying $Q^\top$ from (8.8) to an arbitrary vector require only either $\mathcal{O}(m^2)$ or $\mathcal{O}(m \log(m))$ floating-point operations. The $\mathcal{O}(m^2)$ methods of Gu and Eisenstat (1994, 1995) are usually more efficient than the $\mathcal{O}(m \log(m))$ method of Gu and Eisenstat (1995), unless m is impractically large.

Remark 8.2.3. In (8.14), X_∞ when the entries of a are all zeros is stochastically strictly less than X_∞ for any nonzero a. Indeed, for any $j = 1, 2, \ldots, m-1$, $(Z^{(j)})^2$ is stochastically strictly less than $(Z^{(j)} + \zeta^{(j)})^2$ if $\zeta^{(j)} \neq 0$, while $\zeta^{(1)} = \zeta^{(2)} = \cdots = \zeta^{(m-1)} = 0$ if and only if the entries of a are all zeros; the fact that $(Z^{(j)})^2$ is stochastically strictly less than $(Z^{(j)} + \zeta^{(j)})^2$ if $\zeta^{(j)} \neq 0$ follows from an easy calculation, as given, for example, in item x on page 615 of Ruben (1974).

8.3 Linear combinations of independent noncentral χ^2 variates

This section describes efficient algorithms for evaluating the cdf of the sum (8.14) of the squares of independent Gaussian random variables. The bibliography of Duchesne and de Micheaux (2010) gives references to possible alternatives to the methods of the present section. Our principal tool is the following theorem, representing the cdf as an integral suitable for evaluation via quadratures (see, for example, Remark 8.3.2 below); the theorem expresses formula 7 of Rice (1980) in the same form as (6.8).

Theorem 8.3.1. *Suppose that ℓ is a positive integer, $Z^{(1)}$, $Z^{(2)}$, …, $Z^{(\ell)}$ are i.i.d. Gaussian random variables of zero mean and unit variance, and $\sigma^{(1)}$, $\sigma^{(2)}$, …, $\sigma^{(\ell)}$ and $\zeta^{(1)}$, $\zeta^{(2)}$, …, $\zeta^{(\ell)}$ are real numbers. Suppose in addition that X is the random variable*

$$X = \sum_{j=1}^{\ell} (\sigma^{(j)})^2 \, (Z^{(j)} + \zeta^{(j)})^2. \tag{8.15}$$

Then, the cdf F of X is

$$F(x) = \int_0^\infty \operatorname{Im}\left(\frac{e^{1-y}\, e^{iy\sqrt{\ell}} \prod_{j=1}^{\ell} e^{(\zeta^{(j)})^2(1-w_j(y))/(2w_j(y))}}{\pi\left(y - \frac{1}{1-i\sqrt{\ell}}\right) \prod_{j=1}^{\ell} \sqrt{w_j(y)}} \right) dy \tag{8.16}$$

for any positive real number x, where

$$w_j(y) = 1 - 2(y-1)(\sigma^{(j)})^2/x + 2iy(\sigma^{(j)})^2\sqrt{\ell}/x, \tag{8.17}$$

and $F(x) = 0$ for any nonpositive real number x. The square roots in (8.16) denote the principal branch, and Im *takes the imaginary part.*

Remark 8.3.2. An efficient means of evaluating (8.16) numerically is to employ adaptive Gaussian quadratures; see, for example, Section 4.7 of Press et al. (2007). Good choices for the lowest orders of the quadratures used in the adaptive Gaussian quadratures are 10 and 21, for double-precision accuracy.

The remainder of the present section (particularly Remark 8.3.5) discusses the numerical stability of the method of Remark 8.3.2 and recalls an alternative integral representation suitable for use when the method of Remark 8.3.2 is not guaranteed to be numerically stable. The following lemma, proven in Remark 6.3.2, ensures that the denominator in (8.16) is not too small.

Lemma 8.3.3. *Suppose that ℓ is a positive integer, and r_1, r_2, …, r_ℓ and y are positive real numbers. Suppose further that (in parallel with formula (8.17) above)*

$$w_j = 1 - r_j(y-1) + r_j iy\sqrt{\ell} \tag{8.18}$$

for $j = 1$, 2, …, ℓ.

Then,

$$\left| \prod_{j=1}^{\ell} \sqrt{w_j} \right| > e^{-1/4}. \tag{8.19}$$

The following lemma ensures that the numerator in (8.16) is not too large, provided that $e^{(\zeta^{(j)})^2/2}$ is not large.

Lemma 8.3.4. *Suppose that r, y, and ℓ are positive real numbers and (in parallel with formulae (8.17) and (8.18) above)*

$$w = 1 - r(y-1) + riy\sqrt{\ell}. \tag{8.20}$$

Then,

$$\left| \frac{1-w}{w} \right| \le \sqrt{1 + \frac{1}{\ell}}. \tag{8.21}$$

Proof. Defining

$$z = \frac{1}{y} \tag{8.22}$$

and

$$c = 1 + \frac{1}{r}, \tag{8.23}$$

we obtain that

$$\frac{1-w}{w} = -\frac{1-z-i\sqrt{\ell}}{1-cz-i\sqrt{\ell}}. \tag{8.24}$$

It follows from (8.24) that

$$\left|\frac{1-w}{w}\right|^2 = \frac{(1-z)^2+\ell}{(1-cz)^2+\ell}. \tag{8.25}$$

It follows from (8.22) that $z \ge 0$ and from (8.23) that $c \ge 1$, and hence

$$cz - 1 \ge z - 1. \tag{8.26}$$

If $z \ge 1$, then (8.26) yields that

$$(cz-1)^2 \ge (z-1)^2, \tag{8.27}$$

which in turn yields that

$$\frac{(1-z)^2+\ell}{(1-cz)^2+\ell} \le \frac{(1-z)^2+\ell}{(1-z)^2+\ell} = 1. \tag{8.28}$$

If $z \le 1$, then (recalling that $z \ge 0$, too)

$$\frac{(1-z)^2+\ell}{(1-cz)^2+\ell} \le \frac{(1-z)^2+\ell}{\ell} \le \frac{1+\ell}{\ell}. \tag{8.29}$$

We see from (8.28) and (8.29) that, in all cases,

$$\frac{(1-z)^2+\ell}{(1-cz)^2+\ell} \le 1 + \frac{1}{\ell}. \tag{8.30}$$

Combining (8.25) and (8.30) yields (8.21). □

Remark 8.3.5. The bound (8.19) shows that the integrand in (8.16) is not too large for any nonnegative y, provided that the numerator of (8.16) is not too large. An upper bound on the numerator follows immediately from (8.21):

$$\left|\prod_{j=1}^{\ell} e^{(\zeta^{(j)})^2(1-w_j(y))/(2w_j(y))}\right| \le \prod_{j=1}^{\ell} e^{(\zeta^{(j)})^2\sqrt{1+1/\ell}/2}. \tag{8.31}$$

For any particular application, we can check that the right-hand side of (8.31) is not too many orders of magnitude in size, guaranteeing that applying quadratures to the integral in (8.16) cannot lead to catastrophic cancellation in floating-point arithmetic. Naturally, it is also possible to check on the magnitude of the integrand in (8.16) during its numerical evaluation, indicating even better numerical stability than guaranteed by our *a priori* estimates. See Theorem 8.3.7 and Remark 8.3.8 below for an alternative integral representation suitable for use when the right-hand side of (8.31) is large.

Remark 8.3.6. The bound in (8.31) is quite pessimistic. Often, in fact, the real part of $(1 - w_j(y))/(2w_j(y))$ is nonpositive, so that

$$\left| e^{(\zeta^{(j)})^2 (1 - w_j(y))/(2w_j(y))} \right| \le 1. \tag{8.32}$$

If the right-hand side of (8.31) is large, then we can use the method of Imhof (1961), Davies (1980), and others, applying numerical quadratures to the integral in the following theorem. Please note that the integrand in the following theorem decays reasonably fast when the right-hand side of (8.31) is large.

Theorem 8.3.7. *Suppose that ℓ is a positive integer, $Z^{(1)}$, $Z^{(2)}$, ..., $Z^{(\ell)}$ are i.i.d. Gaussian random variables of zero mean and unit variance, and $\sigma^{(1)}$, $\sigma^{(2)}$, ..., $\sigma^{(\ell)}$ and $\zeta^{(1)}$, $\zeta^{(2)}$, ..., $\zeta^{(\ell)}$ are real numbers. Suppose in addition that X is the random variable*

$$X = \sum_{j=1}^{\ell} (\sigma^{(j)})^2 \, (Z^{(j)} + \zeta^{(j)})^2. \tag{8.33}$$

Then, the cdf F of X is

$$F(x) = \frac{1}{2} - \int_0^\infty \operatorname{Im}\left(\frac{e^{-iy} \prod_{j=1}^{\ell} e^{(\zeta^{(j)})^2 (1 - v_j(y))/(2v_j(y))}}{\pi y \prod_{j=1}^{\ell} \sqrt{v_j(y)}} \right) dy \tag{8.34}$$

for any positive real number x, where

$$v_j(y) = 1 - 2iy(\sigma^{(j)})^2/x, \tag{8.35}$$

and $F(x) = 0$ for any nonpositive real number x. The square roots in (8.34) denote the principal branch, and Im *takes the imaginary part.*

Remark 8.3.8. The integrand in (8.34) is not too large (except for values of y that are closer to 0 than are typical quadrature nodes), since the real part of $(1 - v_j(y))/(2v_j(y))$ is always nonpositive, so that

$$\left| e^{(\zeta^{(j)})^2 (1 - v_j(y))/(2v_j(y))} \right| \le 1. \tag{8.36}$$

Moreover, the numerator in (8.34) decays reasonably fast (it is sub-Gaussian) when the right-hand side of (8.31) is large.

8.4 Numerical method

Combining Sections 8.2 and 8.3 yields an efficient method for calculating the cdf F of n times the square of the root-sum-square distance between the model and empirical distributions, in the limit that n is large, when the n observed draws are taken i.i.d. from an alternative distribution $p_a = p_0 + a/\sqrt{n}$ (as always, p_0 is the model — a probability distribution over m bins — and a is a vector whose m entries satisfy $\sum_{j=1}^{m} a^{(j)} = 0$). Indeed, Theorem 8.2.1 shows that the desired F is the same as that in (8.16) and (8.34), with the real numbers $\sigma^{(1)}$, $\sigma^{(2)}$, ..., $\sigma^{(\ell)}$ and $\zeta^{(1)}$, $\zeta^{(2)}$, ..., $\zeta^{(\ell)}$ calculated as detailed in Section 8.2 (identifying $\ell = m - 1$). Remark 8.3.2 describes an efficient means of evaluating $F(x)$ in (8.16) that is numerically stable when the right-hand side of (8.31) is not too many orders of magnitude in size. When the right-hand side of (8.31) is many orders of magnitude in size, we can apply quadratures to the representation of $F(x)$ in (8.34) instead (see Remark 8.3.8).

8.5 Plotting the asymptotic statistical power

Let us denote by π the cdf of the P-values for the root-mean-square distance (or, equivalently, for any positive multiple of the square of the root-mean-square distance); π is also the statistical power function of the hypothesis test based on the root-mean-square distance (as a function of the significance level). The method of Section 8.4 is sufficient for plotting π in the limit that the number of draws is large. Indeed, suppose that X denotes n times the square of the root-sum-square distance between the model and empirical distributions, F_0 denotes the cdf for X when taking n draws i.i.d. from the model probability distribution p_0, and F_a denotes the cdf for X when taking n draws i.i.d. from $p_a = p_0 + a/\sqrt{n}$, where a is a vector whose m entries satisfy $\sum_{j=1}^m a^{(j)} = 0$. The P-value P equals $1 - F_0(X)$, in the limit that n is large, and then the cdf π of the P-values for draws from p_a is

$$\begin{aligned}\pi(1 - F_0(x)) = \text{Prob}\{P \le 1 - F_0(x)\} = \text{Prob}\{1 - F_0(X) \le 1 - F_0(x)\}\\ = \text{Prob}\{X \ge x\} = 1 - F_a(x) \quad (8.37)\end{aligned}$$

for any nonnegative real number x; thus, the graph of all points $(\alpha, \pi(\alpha))$ with α ranging from 0 to 1 is the same as the graph of all points $(1-F_0(x), 1-F_a(x))$ with x ranging from 0 to ∞, in the limit that n is large. Section 8.4 describes how to evaluate $F_0(x)$ and $F_a(x)$ for any real number x, in the limit that the number n of draws is large; note that $F_0(x) = F_a(x)$ when the entries of a are all zeros, so the procedure of Section 8.4 can evaluate $F_0(x)$ as well as $F_a(x)$. When the entries of a are all zeros, $\zeta^{(1)} = \zeta^{(2)} = \dots = \zeta^{(\ell)} = 0$ in the method of Section 8.4, and then the right-hand side of (8.31) is exactly 1.

8.6 Numerical examples

This section illustrates the algorithms of the present chapter via several numerical examples. As detailed in the subsections below, we consider three examples for the model p_0 (as always, p_0 is a probability distribution over m bins, that is, a vector whose entries are all positive and satisfy $\sum_{j=1}^m p_0^{(j)} = 1$), taking n i.i.d. draws from the alternative probability distribution

$$p_a = p_0 + a/\sqrt{n}, \tag{8.38}$$

where a is a vector whose m entries satisfy $\sum_{j=1}^m a^{(j)} = 0$ (the subsections below detail several examples for a). Figure 8.1 plots the cdf π of the P-values for n i.i.d. draws taken from the alternative distribution p_a, when n is large; π is also the statistical power function of the hypothesis test based on the root-mean-square distance (as a function of the significance level). For each of the examples, Figure 8.1 plots the cdf π both for n = 1,000,000 draws (computed via Monte-Carlo simulations) and in the limit that n is large (computed via the algorithms of the present chapter); not surprisingly, there is little difference between the plots for n = 1,000,000 and for the limit that n is large. The lines in Figure 8.1 corresponding to n = 1,000,000 draws are colored green; the lines corresponding to the limit of large n are black.

Remark 8.6.1. For each example, we computed the cdf π for n = 1,000,000 draws via 40,000 Monte-Carlo simulations. A straightforward argument based on the binomial distribution, detailed in Remark 1.3.2, shows that the standard errors of the resulting estimates of the P-values P are equal to $\sqrt{P(1-P)/40000} \le 0.0025$, ensuring that the standard errors of the plotted abscissae α for the green points in Figure 8.1 are approximately $\sqrt{\alpha(1-\alpha)/40000} \le 0.0025$ (roughly the size of the radii of the plotted points).

Remark 8.6.2. For each example, we plotted the cdf π in the limit of a large number n of draws via the scheme of Section 8.5. Fig. 8.1 displays points $(\alpha, \pi(\alpha)) = (1-F_0(x), 1-F_a(x))$ for the 10000 values $x = 1/2000$, $2/2000$, ..., $10000/2000$, in the limit that the number n of draws is large, where $F_0(x)$ and $F_a(x)$ are defined in Section 8.5 and computed to at least 6-digit accuracy via the method of Section 8.4.

Table 8.1 summarizes computational costs of the procedure described in Section 8.4. The headings of Table 8.1 have the following meanings:

- m is the number of bins in the probability distributions p_0 and p_a.
- q_0 is the maximum number of quadrature nodes required in any of the 10000 evaluations of F_0 plotted in Figure 8.1 (Section 8.5 defines F_0), using adaptive Gaussian quadratures as described in Remark 8.3.2.
- q_a is the maximum number of quadrature nodes required in any of the 10000 evaluations of F_a plotted in Figure 8.1 (Section 8.5 defines F_a), using adaptive Gaussian quadratures as described in Remark 8.3.2.
- t is the time in seconds required to perform the quadratures for both $F_0(x)$ and $F_a(x)$ at a single value of x, amortized over the 10000 pairs $(1-F_0(x), 1-F_a(x))$ plotted in Figure 8.1 (Section 8.5 defines F_0 and F_a).

8.6.1 Uniform model

For our first example, we take

$$p_0^{(j)} = 1/10 \tag{8.39}$$

for $j = 1$, 2, ..., 10, and take

$$a^{(j)} = (-1)^j/5 \tag{8.40}$$

for $j = 1$, 2, ..., 10. The root-mean-square distance is equivalent to the canonical χ^2 statistic for this example, since p_0 is a uniform distribution.

8.6.2 Nonuniform model

For our second example, we take

$$p_0^{(j)} = \begin{cases} 1/2, & j = 1 \\ 1/198, & j = 2, 3, \dots, 100 \end{cases} \tag{8.41}$$

for $j = 1$, 2, ..., 100, and take

$$a^{(j)} = \begin{cases} 2/3, & j = 1 \\ -2/297, & j = 2, 3, \dots, 100 \end{cases} \tag{8.42}$$

for $j = 1$, 2, ..., 100.

8.6.3 Poisson model

For our third example, we take

$$p_0^{(j)} = e^{-3}\, 3^{j-1}/(j-1)! \tag{8.43}$$

for $j = 1, 2, 3, \dots$, and take

$$a^{(j)} = \begin{cases} (-1)^j/4, & j = 1, 2, 3, 4 \\ (-1)^j/2, & j = 5, 6 \\ 0, & j = 7, 8, 9, \dots \end{cases} \tag{8.44}$$

for $j = 1, 2, 3, \dots$. For all numerical computations associated with this example, we can truncate to the first 20 bins, since $\sum_{j=21}^{\infty} p_0^{(j)} < 10^{-10}$.

8.6.4 Poisson model with a different alternative

For our fourth example, we again take

$$p_0^{(j)} = e^{-3}\, 3^{j-1}/(j-1)! \tag{8.45}$$

for $j = 1, 2, 3, \dots$, but now take

$$a^{(j)} = \begin{cases} 1, & j = 1 \\ -1/11, & j = 2, 3, \dots, 12 \\ 0, & j = 13, 14, 15, \dots \end{cases} \tag{8.46}$$

for $j = 1, 2, 3, \dots$. For all numerical computations associated with this example, we can truncate to the first 20 bins, since $\sum_{j=21}^{\infty} p_0^{(j)} < 10^{-10}$.

Remark 8.6.3. The right-hand side of (8.31) is 8.233 for Subsection 8.6.1, 2.443 for Subsection 8.6.2, and 24.05 for Subsection 8.6.3. As discussed in Remark 8.3.5, roundoff errors in the numerical evaluation of (8.16) are therefore guaranteed to be negligible for the standard floating-point arithmetic (the mantissa in the standard, double-precision arithmetic has a dynamic range of about $5 \cdot 10^{15} \gg 24.05$). The right-hand side of (8.31) is $1.478 \cdot 10^{16}$ for Subsection 8.6.4, so we used (8.34) rather than (8.16) for the last example (Remark 8.3.8 explains why).

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 $2 \cdot 10^{-16}$). We diagonalized the matrix B defined in (8.7) using the Jacobi algorithm (see, for example, Chapter 8 of Golub and Van Loan (1996)), not taking advantage of Remark 8.2.2; explicitly forming the entries of the matrix B defined in (8.7) can incur a numerical error of at most the machine precision (about $2 \cdot 10^{-16}$) times $\max_{1 \le j \le m} p_0^{(j)} / \min_{1 \le j \le m} p_0^{(j)}$, yielding 6-digit accuracy or better for all our examples. Higher precision is possible via the interlacing properties of eigenvalues, following Gu and Eisenstat (1994). Of course, even 4-digit precision would suffice for most statistical applications; however, modern computers can produce high accuracy very fast, as the examples in this section illustrate.

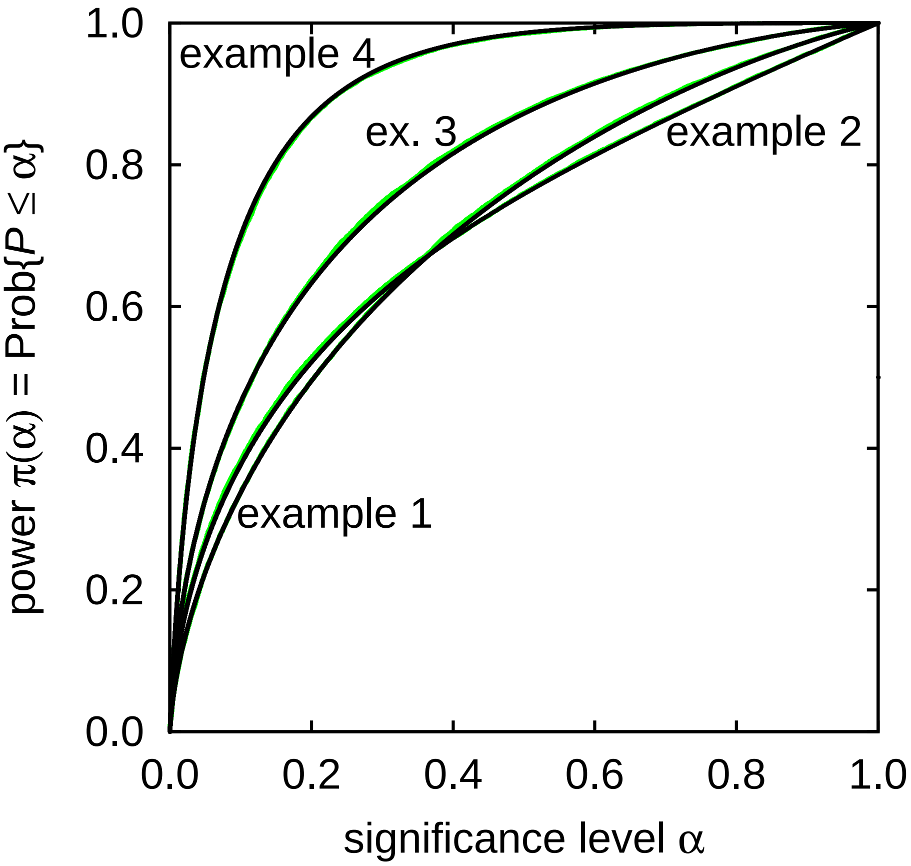


Figure 8.1: Cumulative distribution functions of the P-values P for draws from the alternative distributions defined in Subsections 8.6.1–8.6.4

Table 8.1: Computational costs

	m	q_0	q_a	t
example 1	10	230	230	0.006
example 2	100	530	550	0.090
example 3	20	250	330	0.013
example 4	20	350	350	0.010

Appendix A

Significance testing without truth

A popular approach to significance testing proposes to decide whether the given hypothesized statistical model is likely to be true (or false). Statistical decision theory provides a basis for this approach by requiring every significance test to make a decision about the truth of the hypothesis/model under consideration. Unfortunately, many interesting and useful models are obviously false (that is, not exactly true) even before considering any data. Fortunately, in practice a significance test need only gauge the consistency (or inconsistency) of the observed data with the assumed hypothesis/model — without enquiring as to whether the assumption is likely to be true (or false), or whether some alternative is likely to be true (or false). In this practical formulation, a significance test rejects a hypothesis/model only if the observed data is highly improbable when calculating the probability while assuming the hypothesis being tested; the significance test only gauges whether the observed data likely invalidates the assumed hypothesis, and cannot decide that the assumption — however unmistakably false — is likely to be false a priori, without any data.

Essentially, all models are wrong, but some are useful. — G. E. P. Box

A.1 Introduction

As pointed out in the above quotation of G. E. P. Box, many interesting models are false (that is, not exactly true), yet are useful nonetheless. Significance testing helps measure the usefulness of a model. Testing the validity of using a model for virtually any purpose requires knowing whether observed discrepancies are due to inaccuracies or inadequacies in the model or (on the contrary) could be due to chance arising from necessarily finite sample sizes. Significance tests gauge whether the discrepancy between the model and the observed data is larger than expected random fluctuations; significance tests gauge the size of the unavoidable random fluctuations.

A traditional approach, along with its modern formulation in statistical decision theory, tries to decide whether a hypothesized model is likely to be true (or false). However, in many practical circumstances, a significance test need only gauge the consistency (or inconsistency) of the observed data with the assumed hypothesis/model — without ever enquiring as to whether the assumption is likely to be true (or false), or whether some alternative is likely to be true (or false). In this practical formulation, a significance test rejects a hypothesis/model only if the observed data is highly improbable when calculating the probability while assuming the hypothesis being tested. Whether or not the assumption could be exactly true in reality is irrelevant.

An illustrative example may help clarify. When testing the goodness of fit for the Poisson regression where the distribution of Y given x is the Poisson distribution of mean $\exp(\theta^{(0)} + \theta^{(1)}x + \theta^{(2)}x^2 + \theta^{(3)}x^3)$, the conventional Neyman-Pearson null hypothesis is

$$\begin{aligned} H_0^{\mathrm{NP}} &: \text{there exist real numbers } \theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)} \text{ such that } y_1, y_2, \dots, y_n \text{ are draws} \\ &\quad \text{from independent Poisson distributions with means } \mu_1, \mu_2, \dots, \mu_n, \text{ respectively,} \end{aligned} \tag{A.1}$$

where

$$\ln(\mu_k) = \theta^{(0)} + \theta^{(1)}x_k + \theta^{(2)}(x_k)^2 + \theta^{(3)}(x_k)^3 \tag{A.2}$$

for $k = 1, 2, \dots, n$, and the observations (x_1, y_1), (x_2, y_2), $\dots$, (x_n, y_n) are ordered pairs of scalars (real numbers paired with nonnegative integers). A related, but perhaps simpler null hypothesis is

$$\begin{aligned} H_0 &: y_1, y_2, \dots, y_n \text{ are draws from independent Poisson distributions} \\ &\qquad\qquad \text{with means } \hat\mu_1, \hat\mu_2, \dots, \hat\mu_n, \text{ respectively,} \end{aligned} \tag{A.3}$$

where

$$\ln(\hat\mu_k) = \hat\theta^{(0)} + \hat\theta^{(1)}x_k + \hat\theta^{(2)}(x_k)^2 + \hat\theta^{(3)}(x_k)^3 \tag{A.4}$$

for $k = 1, 2, \dots, n$, with $\hat\theta$ being a maximum-likelihood estimate. Needless to say, even if the observed data really does arise from Poisson distributions whose means are exponentials of a cubic polynomial, the particular values $\hat\theta^{(0)}$, $\hat\theta^{(1)}$, $\hat\theta^{(2)}$, $\hat\theta^{(3)}$ of the parameters of the fitted polynomial will almost surely not be exactly equal to the true values. Even though the estimated values of the parameters may not be exactly correct, it still makes good sense to enquire as to whether the fitted cubic polynomial is consistent with the data up to random fluctuations inherent in using a finite amount of observed data.

In fact, since subsequent use of the model usually involves the particular fitted polynomial — whose specification includes the observed parameter estimates — analyzing the model including the estimated values of the parameters makes more sense than trying to decide whether the data really did come from Poisson distributions whose means are exponentials of some unspecified cubic polynomial. For instance, any plot of the fit (such as a plot of the means of the Poisson distributions) must use the estimated values of the parameters, and any statistical interpretation of the plot should also depend explicitly on the estimates; a significance test can gauge the consistency of the plotted fit with the observed data, without ever asking whether the plotted fit is the truth (it is almost surely not identical to the underlying reality) and without making some decision about an abstract family of polynomials which may or may not include both the plotted fit and the underlying reality.

A popular measure of divergence from the null hypothesis is the log–likelihood-ratio

$$g^2 = 2\sum_{k=1}^{n} y_k \ln(y_k/\hat\mu_k). \tag{A.5}$$

A P-value (see, for example, Section A.3 below) quantifies whether this divergence is larger than expected from random fluctuations inherent in using only n data points. It is not obvious how to calculate an exact P-value for H_0^{NP} from (A.1) and (A.2), which refers to cubic polynomials with undetermined coefficients. In contrast, H_0 from (A.3) and (A.4) refers explicitly to the particular fitted value $\hat{\theta}$; H_0 concerns the particular fit displayed in a plot, and is natural for the statistical interpretation of such a plot.

Thus, when calculating significance, the assumed model should include the particular values of any parameters estimated from the observed data. Such parameters are known as "nuisance" parameters. As illustrated with H_0 from (A.3) and (A.4), the assumed hypothesis will be "simple" in the Neyman-Pearson sense, but will depend on the observed values of the parameters — that is, the hypothesis will be "data-dependent"; the hypothesis will be "random." Including the particular values of the parameters estimated from the observed data replaces the "composite" hypothesis of the conventional Neyman-Pearson formulation with a "simple" data-dependent hypothesis. As discussed in Section A.4 below, fully conditional tests also incorporate the observed values of the parameters, but make the extra assumption that all possible realizations of the experiment — observed or hypothetical — generate the same observed values of the parameters. The device of a "simple data-dependent hypothesis" such as H_0 incorporates the observed values explicitly without the extra assumption.

For most purposes, a parameterized model is not really operational — that is, suitable for making precise predictions — until its specification is completed via the inclusion of estimates for any nuisance parameters. The results of the significance tests considered below depend on the quality of both the models and the parameter estimators. However, the results are relatively insensitive to the particular observed realizations of the parameter estimators (that is, to the parameter estimates) unless specifically designed to quantify the quality of the parameter estimates. To quantify the quality of the parameter estimates, we recommend testing separately the goodness of fit of the parameter estimates, using confidence intervals, confidence distributions, parametric bootstrapping, or significance tests within parametric models, whose statistical power is focused against alternatives within the parametric family constituting the model (for further discussion of the latter, see Section A.5 below).

The remainder of the present appendix has the following structure: Section A.2 very briefly discusses Bayesian-frequentist hybrids, referring for details to the definitive work of Gelman (2003). Section A.3 defines P-values — also known as "attained significance levels" — which quantify the consistency of the observed data with the assumed models. Section A.4 details several approaches to testing the goodness of fit for distributional profile. Section A.5 discusses testing the goodness of fit for various properties beyond just distributional profile.

Cox (2006) details many advantages of interpreting significance as gauging the consistency of an assumption/hypothesis with observed data, rather than as making decisions about the actual truth of the assumption. However, significance testing is meaningless without any observations, unlike purely Bayesian methods, which can produce results without any data, courtesy of the prior (the prior is the statistician's system of a priori beliefs, accumulated from prior experience, law, morality, religion, etc., without reference to the observed data). Significance tests are deficient in this respect. Those interested in what is to be considered true in reality and in making decisions more generally should use Bayesian, sequential, and multilevel procedures. Significance testing simply gauges the consistency of models with observed data; generally significance testing alone cannot handle the truth.

A.2 Bayesian versus frequentist

Traditionally, significance testing is frequentist. However, there exist Bayesian-frequentist hybrids known as “Bayesian P-values”; Gelman (2003) sets forth a particularly appealing formulation. Bayesian P-values test the consistency of the observed data with the model *used together with a prior* for nuisance parameters. In contrast, the P-values discussed in the present appendix test the consistency of the observed data with the model *used together with a parameter estimator*. In the Bayesian formulation, a small P-value reflects inconsistency with the model *and/or* the prior; in the formulation of the present appendix, a small P-value reflects inconsistency with the model and/or the parameter estimator. Thus, when there are nuisance parameters, the two types of P-values test slightly different hypotheses and provide slightly different information; each type is ideal for its own set-up. Of course, if there are no nuisance parameters, then Bayesian P-values and the P-values discussed below are the same.

A.3 P-values

A P-value for a hypothesis H_0 is a statistic such that, if the P-value is very small, then we can be confident that the observed data is inconsistent with assuming H_0. The P-value associated with a measure of divergence from a hypothesis H_0 is the probability that $D \geq d$, where d is the divergence between the observed and the expected (with the expectation following H_0 for the observations), and D is the divergence between the simulated and the expected (with the expectation following H_0 for the simulations, and with the simulations performed assuming H_0). When taking the probability that $D \geq d$, we view D as a random variable, while viewing d as fixed, not random. For example, when testing the goodness of fit for the model of i.i.d. draws from a probability distribution $p_0(\theta)$, where θ is a nuisance parameter that must be estimated from the data, that is, from observations x_1, x_2, …, x_n, we use the null hypothesis

$$H_0 : x_1, x_2, \dots, x_n \text{ are i.i.d. draws from } p_0(\hat{\theta}), \text{where } \hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n). \tag{A.6}$$

The P-value for H_0 associated with a divergence δ is the probability that $D \geq d$, where

- $d = \delta(\hat{p}, p_0(\hat{\theta}))$,
- $\hat{p}$ is the empirical distribution of x_1, x_2, …, x_n,
- $\hat{\theta}$ is the parameter estimate obtained from the observed draws x_1, x_2, …, x_n,
- $D = \delta(\hat{P}, p_0(\hat{\Theta}))$,
- $\hat{P}$ is the empirical distribution of i.i.d. draws X_1, X_2, …, X_n from $p_0(\hat{\theta})$, and
- $\hat{\Theta}$ is the parameter estimate obtained from the simulated draws X_1, X_2, …, X_n.

If the P-value is very small, then we can be confident that the observed data is inconsistent with assuming H_0. Examples of divergences include χ^2 (for categorical data) and the maximum absolute difference between cumulative distribution functions (for real-valued data).

Remark A.3.1. To compute the P-value assessing the consistency of the experimental data with assuming H_0, we can use Monte-Carlo simulations (very similar to those used by Clauset et al. (2009)). First, we estimate the parameter θ from the n given experimental draws, obtaining $\hat{\theta}$, and calculate the divergence between the empirical distribution and $p_0(\hat{\theta})$. 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 divergence between the empirical distribution of the data generated in Step 1 and $p_0(\tilde{\theta})$, 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 H_0 as the fraction of the divergences calculated in Step 3 that are greater than or equal to the divergence 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; in fact, the standard error of the estimate for an exact P-value P is $\sqrt{P(1-P)/\ell}$, where ℓ is the number of Monte-Carlo simulations conducted to produce the estimate — see Remark 1.3.2.

A.4 Goodness of fit for distributional profile

Given observations x_1, x_2, $\dots$, x_n, we can test the goodness of fit for the model of i.i.d. draws from a probability distribution $p_0(\theta)$, where θ is a nuisance parameter, via the null hypothesis

$$\begin{aligned} &H_0 : x_1, x_2, \dots, x_n \text{ are i.i.d. draws from } p_0(\hat{\theta}) \\ &\qquad\qquad \text{for the particular observed value of } \hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n). \end{aligned} \tag{A.7}$$

The Neyman-Pearson formulation considers instead the null hypothesis

$$H_0^{\text{NP}} : \text{there exists a value of } \theta \text{ such that } x_1, x_2, \dots, x_n \text{ are i.i.d. draws from } p_0(\theta). \tag{A.8}$$

The fully conditional null hypothesis is

$$\begin{aligned} &H_0^{\text{FC}} : x_1, x_2, \dots, x_n \text{ are i.i.d. draws from } p_0(\hat{\theta}) \\ &\qquad \text{and } \hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n) \text{ takes the same value in all possible realizations.} \end{aligned} \tag{A.9}$$

That is, whereas H_0 supposes that the particular observed realization of the experiment happened to produce a parameter estimate $\hat{\theta}$ that is consistent with having drawn the data from $p_0(\hat{\theta})$, H_0^{FC} assumes that every possible realization of the experiment — observed or hypothetical — produces exactly the same parameter estimate. Few experimental apparatus constrain the parameter estimate to always take the same (a priori unknown) value during repetitions of the experiment, as H_0^{FC} assumes. Assuming H_0^{FC} amounts to conditioning on a statistic that is minimally sufficient for estimating θ; computing the associated P-values is not always trivial. Furthermore, the assumption that H_0^{FC} is true seems to be more extreme, a more substantial departure from H_0^{NP}, than H_0. Finally, testing the significance of assuming H_0 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 $p_0(\hat{\theta})$. We cannot recommend the use of H_0^{FC} in general. Unfortunately, H_0^{NP} also presents problems. . . .

If the probability distributions are discrete, many definitions of the exact P-value for H_0^{NP} involve the actual value of the parameter θ when H_0^{NP} is true (even though the observed data may not determine this value exactly), leaving the P-value undefined when H_0^{NP} is false. An alternative "P-value" is the maximum (or supremum) of P-values calculated separately for every possible value of the parameter; the resulting "maximum" P-value for H_0^{NP} then may not have the desirable asymptotic property discussed in Remark A.4.1 below — a property that many consider necessary to merit the designation "P-value," as discussed by Bayarri and Berger (2000), Gelman (2003), Robins et al. (2000), and many others. The situation may be more favorable when measuring discrepancies with divergences that are "approximately ancillary" with respect to θ, but quantifying the notion of "approximately" seems to be problematic except in the limit of large numbers of draws. (Some divergences are asymptotically ancillary in the limit of large numbers of draws, but this is not especially helpful, as any asymptotically consistent estimator $\hat{\theta}$ converges to the correct value in the limit of large numbers of draws; θ is almost surely known exactly in the limit of large numbers of draws, so there is not too much benefit to being independent of θ only in that limit.) Section 3 of Robins and Wasserman (2000) reviews these and related issues.

Remark A.4.1. Romano (1988), Henze (1996), Bickel et al. (2006), and others have shown that the P-values for H_0 converge in distribution to the uniform distribution over $[0, 1]$ in the limit of large numbers of draws, when H_0^{NP} is true. In particular, Romano (1988) and Henze (1996) prove this convergence for a wide class of divergence measures.

Remark A.4.2. 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 for the null hypothesis H_0 from (A.7) have the following highly desirable property: Suppose that the actual underlying distribution p of the i.i.d. experimental draws is equal to $p_0(\theta)$ for some (unknown) θ. Suppose further that P is the P-value for assuming $p = p_0(\hat{\theta})$ for the observed value of $\hat{\theta}$, 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 $p = p_0(\hat{\theta})$. 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 also holds when there is no parameter in the model. Furthermore, the P-values for H_0 from (A.7) can be viewed as a parametric bootstrap approximation even if the parameter θ is not a permutation, as discussed in the preceding remark.

Remark A.4.3. The surveys of Agresti (1992) and Agresti (2001) discuss exact P-values for contingency-tables/cross-tabulations, including criticism of fully conditional P-values.

Gelman (2003) provides further criticism of fully conditional P-values. Chapter 3 numerically evaluates the different types of P-values for an application in population genetics. Section 4 of Bayarri and Berger (2004) and the references it cites discuss the menagerie of alternative P-values proposed recently.

A.5 Goodness of fit for various properties

For comparative purposes, we first review the null hypothesis of the previous section for testing the goodness of fit for distributional profile, namely

$$H_0 : x_1, x_2, \dots, x_n \text{ are i.i.d. draws from } p_0(\hat{\theta}), \text{where } \hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n), \tag{A.10}$$

with θ being the nuisance parameter. The measure of discrepancy for H_0 is usually taken to be a divergence between the empirical distribution $\hat{p}$ and the model $p_0(\hat{\theta})$ (in the continuous case in one dimension, a common characterization of the empirical distribution is the empirical cumulative distribution function; in the discrete case, a common characterization of the empirical distribution is the empirical probability mass function, that is, the set of empirical proportions). One example for p_0 is the Zipf distribution over m bins with parameter θ, a discrete distribution with the probability mass function

$$p_0^{(j)}(\theta) = \frac{C_\theta}{j^\theta} \tag{A.11}$$

for $j = 1, 2, 3, \dots, m$, where the normalization constant is

$$C_\theta = \frac{1}{\sum_{j=1}^{m} j^{-\theta}} \tag{A.12}$$

and θ is a nonnegative real number.

When testing the goodness of fit for parameter estimates, we use the null hypothesis

$$H_0' : x_1, x_2, \dots, x_n \text{ are i.i.d. draws from } p_0(\varphi_0, \hat{\theta}), \text{where } \hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n), \tag{A.13}$$

with θ being the nuisance parameter and φ being the parameter of interest (and with φ_0 being the value of φ assumed under the model). Please note that H_0 and H_0' are actually equivalent, via the identification $p_0(\theta) = p_0(\varphi_0, \theta)$. However, the measure of discrepancy for H_0' is usually taken to be a divergence between $\hat{\varphi}$ and φ_0 rather than the divergence between $\hat{p}$ and $p_0(\hat{\theta})$ that is more natural for H_0. Also, if φ is scalar-valued, then confidence intervals, confidence distributions, and parametric bootstrap distributions are more informative than a significance test. A significance test is appropriate if φ is vector-valued. One example for p_0 is the sorted Zipf distribution over m bins with θ being the power in the power law and with the maximum-likelihood estimate $\hat{\varphi}$ being a permutation that sorts the bins into rank-order, that is, p_0 is the discrete distribution with the probability mass function

$$p_0^{(j)}(\varphi, \theta) = \frac{C_\theta}{(\varphi(j))^\theta} \tag{A.14}$$

for $j = 1, 2, 3, \ldots, m$, where the normalization constant C_θ is defined in (A.12) with θ being a nonnegative real number, and φ is a permutation of the numbers 1, 2, …, m. The choice for φ_0 that is of widest interest in applications is the identity permutation (i.e., the "rearrangement" of the bins that does not permute any bins: $\varphi_0(j) = j$ for $j = 1, 2, \ldots, m$).

When testing the goodness of fit for the standard Poisson regression with the distribution of Y given x being the Poisson distribution of mean $\exp\left(\theta^{(0)} + \sum_{j=1}^m \theta^{(j)} x^{(j)}\right)$, we use the null hypothesis

$$H_0'' : y_1, y_2, \ldots, y_n \text{ are draws from independent Poisson distributions with means}$$
$$\exp\left(\hat\theta^{(0)} + \sum_{j=1}^m \hat\theta^{(j)} x_1^{(j)}\right),\ \exp\left(\hat\theta^{(0)} + \sum_{j=1}^m \hat\theta^{(j)} x_2^{(j)}\right),\ \ldots,\ \exp\left(\hat\theta^{(0)} + \sum_{j=1}^m \hat\theta^{(j)} x_n^{(j)}\right),$$
$$\text{respectively,} \tag{A.15}$$

where θ is the nuisance parameter and $\hat\theta$ is its maximum-likelihood estimate. The measure of discrepancy for H_0'' is usually taken to be the log–likelihood-ratio (also known as the deviance)

$$g^2 = 2 \sum_{k=1}^n y_k \ln(y_k / \hat\mu_k), \tag{A.16}$$

where $\hat\mu_k$ is the mean of the Poisson distribution associated with y_k in H_0'', namely,

$$\hat\mu_k = \exp\left(\hat\theta^{(0)} + \sum_{j=1}^m \hat\theta^{(j)} x_k^{(j)}\right). \tag{A.17}$$

One example is the cubic polynomial

$$\ln(\mu_k) = \theta^{(0)} + \theta^{(1)} x_k + \theta^{(2)} (x_k)^2 + \theta^{(3)} (x_k)^3 \tag{A.18}$$

for $k = 1, 2, \ldots, n$, which comes from the choice $m = 3$ and

$$x_k^{(1)} = x_k; \quad x_k^{(2)} = (x_k)^2; \quad x_k^{(3)} = (x_k)^3 \tag{A.19}$$

for $k = 1, 2, \ldots, n$, given observations as ordered pairs of scalars (x_1, y_1), (x_2, y_2), …, (x_n, y_n). Of course, there are similar formulations for other generalized linear models, such as those discussed by McCullagh and Nelder (1989). Chapter 5 above discusses logistic regression, too.

Acknowledgements

Raymond Carroll and Gary Simon are co-authors on articles closely related to Chapters 3 and 8, respectively. Their help has been tremendous. In addition, they introduced us to the standard conventions of statistical writing.

We would also like to thank Alex Barnett, Andrew Barron, Gérard Ben Arous, James Berger, Tony Cai, Sourav Chatterjee, Ronald Raphael 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, Amit Singer, Alan Sokal, Michael Stein, Stephen Stigler, John Strain, Joel Tropp, Larry Wasserman, Douglas A. Wolfe, and Bin Yu. We would especially like to thank Jiayang Gao for her many observations, which include pointing out the identity in (1.4), showing that the Freeman-Tukey/Hellinger-distance statistic is just a weighted version of the root-mean-square.

Will Perkins was supported in part by an NSF Postdoctoral Research Fellowship and NSF Grant OISE-0730136. 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 Research Fellowship from the Alfred P. Sloan Foundation, a Donald D. Harrington Faculty Fellowship, ONR Grant N00014-12-1-0743, an NSF CAREER Award, and an AFOSR Young Investigator Program Award.

Bibliography

Agresti, A. (1992), A survey of exact inference for contingency tables, *Statist. Sci.*, **7**, 131–153.

Agresti, A. (2001), Exact inference for categorical data: recent advances and continuing controversies, *Stat. Med.*, **20**, 2709–2722.

Allison, P. (2012), *Logistic Regression Using SAS: Theory and Application*, Cary, North Carolina: SAS Institute, 2nd ed.

Ampadu, C. (2008), On the powers of some new chi-square type statistics, *Far East J. Theoretical Statist.*, **26**, 59–72.

Ampadu, C., Wang, D., and Steele, M. (2009), Simulated power of some discrete goodness-of-fit test statistics for testing the null hypothesis of a zig-zag distribution, *Far East J. Theoretical Statist.*, **28**, 157–171.

Andersen, E. B. (1990), *The Statistical Analysis of Categorical Data*, Berlin: Springer-Verlag.

Ayres, K. and Balding, D. (1998), Measuring departures from Hardy-Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient, *Heredity*, **80**, 769–777.

Bayarri, M. J. and Berger, J. O. (2000), P-values for composite null models, *J. Amer. Statist. Assoc.*, **95**, 1127–1142.

Bayarri, M. J. and Berger, J. O. (2004), The interplay of Bayesian and frequentist analysis, *Statist. Sci.*, **19**, 58–80.

Beran, R. J. and Millar, P. W. (1991), Tests of fit for logistic models, Tech. Rep. 160, U.C. Berkeley Dept. of Statistics.

Best, D. J. and Rayner, J. C. W. (1997), Goodness-of-fit for the ordered categories discrete uniform distribution, *Comm. Statist. Theory Meth.*, **26**, 899–909.

Bickel, P. J., Ritov, Y., and Stoker, T. M. (2006), Tailor-made tests for goodness of fit to semiparametric hypotheses, *Ann. Statist.*, **34**, 721–741.

Bowden, C. L., Brugger, A. M., Swann, A. C., Calabrese, J. R., Janicak, P. G., Petty, F., Dilsaver, S. C., Davis, J. M., Rush, A. J., Small, J. G., Garza-Treviño, E. S., Risch, S. C., Goodnick, P. J., and Morris, D. D. (1994), Efficacy of divalproex vs. lithium and placebo in the treatment of mania, *J. Amer. Med. Assoc.*, **271**, 918–924.

Brownlee, K. (1965), *Statistical series and methodology in science and engineering*, New York: Wiley.

Chen, J. and Thomson, G. (1999), The variance for the disequilibrium coefficient in the individual Hardy-Weinberg test, *Biometrics*, **55**, 1269–1272.

Choulakian, V., Lockhart, R. A., and Stephens, M. A. (1994), Cramér–von-Mises statistics for discrete distributions, *Canadian J. Statist.*, **22**, 125–137.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009), Power-law distributions in empirical data, *SIAM Review*, **51**, 661–703.

Consonni, G., Moreno, E., and Venturini, S. (2011), Testing Hardy-Weinberg equilibrium: an objective Bayesian analysis, *Statistics in Medicine*, **30**, 62–74.

Cox, D. R. (2006), *Principles of Statistical Inference*, Cambridge, UK: Cambridge University Press.

D'Agostino, R. B. and Stephens, M. A. (1986), *Goodness-of-Fit Techniques*, New York: Marcel Dekker.

Davies, R. B. (1980), Algorithm AS 155: The distribution of a linear combination of χ^2 random variables, *J. Roy. Statist. Soc. Ser. C*, **29**, 323–333.

del Barrio, E., Cuesta-Albertos, J. A., and Matrán, C. (2000), Contributions of empirical and quantile processes to the asympototic theory of goodness-of-fit tests, *Test*, **9**, 1–53.

Diaconis, P. and Sturmfels, B. (1998), Algebraic algorithms for sampling from conditional distributions, *Annals of Statistics*, **26**, 363–397.

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

Durbin, J. (1972), *Distribution Theory for Tests Based on the Sample Distribution Function*, CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia: Society for Industrial and Applied Mathematics.

Eldridge, R. C. (2010), *Six Thousand Common English Words: Their Comparative Frequency and What Can Be Done with Them*, Charleston, SC: Nabu Press, reprint of the 1911 edition; available online at http://www.archive.org/details/sixthousandcomm00eldrgoog.

Engels, W. (2009), Exact tests for Hardy-Weinberg proportions, *Genetics*, **183**, 1431–1441.

Erosheva, E., Walton, E. C., and Takeuchi, D. T. (2007), Self-rated health among foreign- and US-born Asian Americans: A test of comparability, *Med. Care*, **45**, 80–87.

Finney, D. J. (1947), The estimation from individual records of the relationship between dose and quantal response, *Biometrika*, **34**, 320–334.

Fisher, R. A. (1925), *Statistical methods for research workers*, Edinburgh: Oliver and Boyd.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943), The relation between the number of species and the number of individuals in a random sample of an animal population, *J. Animal Ecology*, **12**, 42–58.

From, S. G. (1996), A new goodness-of-fit test for the equality of multinomial cell probabilities versus trend alternatives, *Comm. Statist. Theory Meth.*, **25**, 3167–3183.

Gelman, A. (2003), A Bayesian formulation of exploratory data analysis and goodness-of-fit testing, *Internat. Stat. Rev.*, **71**, 369–382.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995), *Bayesian Data Analysis*, London: Chapman and Hall.

Gibbons, J. and Pratt, J. (1975), P-values: interpretation and methodology, *American Statistician*, **29**, 20–25.

Gilchrist, E. (2010), A sweet approach to teaching the one-variable chi-square test, *Communication Teacher*, **24**, 14–18.

Golub, G. H. and Van Loan, C. F. (1996), *Matrix Computations*, Baltimore, Maryland: Johns Hopkins University Press, 3rd ed.

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

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

Guo, S. W. and Thompson, E. A. (1992), Performing the exact test of Hardy-Weinberg proportion for multiple alleles, *Biometrics*, **48**, 361–372.

Haldane, J. (1954), An exact test for randomness of mating, *Journal of Genetics*, **52**, 631–635.

Hardy, G. (1908), Mendelian proportions in a mixed population, *Science*, **28**, 49–50.

Haschenburger, J. K. and Spinelli, J. J. (2005), Assessing the goodness-of-fit of statistical distributions when data are grouped, *Math. Geology*, **37**, 261–276.

Henze, N. (1996), Empirical-distribution-function goodness-of-fit tests for discrete models, *Canad. J. Statist.*, **24**, 81–93.

Hoeffding, W. (1965), Asymptotically optimal tests for multinomial distributions, *Ann. Math. Statist.*, **36**, 369–401.

Hollander, M. and Wolfe, D. A. (1999), *Nonparametric Statistical Methods*, Wiley, 2nd ed.

Horn, S. D. (1977), Goodness-of-fit tests for discrete data: a review and an application to a health impairment scale, *Biometrics*, **33**, 237–247.

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997), A comparison of goodness-of-fit tests for the logistic regression model, *Statistics in Medicine*, **16**, 965–980.

Hosmer, D. W. and Lemeshow, S. (1980), Goodness of fit tests for the multiple logistic regression model, *Comm. Statist. Theory Meth.*, **9**, 1043–1069.

Hosmer, D. W. and Lemeshow, S. (2000), *Applied Logistic Regression*, Wiley, 2nd ed.

Imhof, J. P. (1961), Computing the distribution of quadratic forms in normal variables, *Biometrika*, **48**, 419–426.

Kendall, M. G., Stuart, A., Ord, K., and Arnold, S. (2009), *Kendall's Advanced Theory of Statistics*, vol. 1 and 2A, Wiley, 6th ed.

Khoury, M., Little, J., and Burke, W. (2004), *Human genome epidemiology: a scientific foundation for using genetic information to improve health and prevent disease*, Oxford: Oxford University Press.

Kleinbaum, D. G. and Klein, M. (2010), *Logistic Regression*, Statistics for Biology and Health, Springer, 3rd ed.

Lauretto, M., Nakano, F., Faria, S., Pereira, C., and Stern, J. (2009), A straightforward multiallelic significance test for the Hardy-Weinberg equilibrium law, *Genetics and Molecular Biology*, **32**, 619–625.

Levene, H. (1949), On a matching problem arising in genetics, *Annals of Mathematical Statistics*, **20**, 91–94.

Li, Y. and Graubard, B. (2009), Testing Hardy-Weinberg equilibrium and homogeneity of Hardy-Weinberg disequilibrium using complex survey data, *Biometrics*, **65**, 1096–1104.

Lockhart, R. A., Spinelli, J. J., and Stephens, M. A. (2007), Cramér–von-Mises statistics for discrete distributions with unknown parameters, *Canadian J. Statist.*, **35**, 125–133.

Marsaglia, G. (2003), Random number generators, *J. Modern Appl. Stat. Meth.*, **2**, 2–13.

Marsaglia, G., Tsang, W. W., and Wang, J. (2003), Evaluating Kolmogorov's distribution, *J. Statist. Soft.*, **8**, 1–4.

Mayo, D. G. (1996), *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman and Hall, 2nd ed.

Mehta, C. and Patel, N. (1983), A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables, *J. Amer. Statist. Assoc.*, **78**, 427–434.

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

Motwani, R. and Raghavan, P. (1995), *Randomized Algorithms*, Cambridge, UK: Cambridge University Press.

Munk, A. and Czado, C. (1998), Nonparametric validation of similar distributions and assessment of goodness of fit, *J. Roy. Statist. Soc. Ser. B*, **60**, 223–241.

Pan, Z. and Lin, D. (2005), Goodness-of-fit methods for generalized linear mixed models, *Biometrics*, **61**, 1000–1009.

Pearson, K. (1900), 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**, 157–175.

Pettitt, A. N. and Stephens, M. A. (1977), The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data, *Technometrics*, **19**, 205–210.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (2007), *Numerical Recipes*, Cambridge, UK: Cambridge University Press, 3rd ed.

Press, W. H. (2005), "What is better than chi-square?" and related koans, Available at http://www.nr.com/whp/notes/betterthanchi.pdf.

Radlow, R. and Alf, E. (1975), An alternative multinomial assessment of the accuracy of the χ^2 test of goodness of fit, *J. Amer. Statist. Assoc.*, **70**, 811–813.

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

Raymond, M. and Rousset, F. (1995), An exact test for population differentiation, *Evolution*, **49**, 1280–1283.

Read, T. R. C. and Cressie, N. A. C. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*, New York: Springer.

Rényi, A. (1953), On the theory of order statistics, *Acta Math. Acad. Sci. Hungar.*, **4**, 191–231.

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

Robins, J. M., van der Vaart, A., and Ventura, V. (2000), Asymptotic distribution of P-values in composite null models, *J. Amer. Statist. Assoc.*, **95**, 1143–1156.

Robins, J. M. and Wasserman, L. (2000), Conditioning, likelihood, and coherence: a review of some foundational concepts, *J. Amer. Statist. Assoc.*, **95**, 1340–1346.

Rohlfs, R. V. and Weir, B. S. (2008), Distributions of Hardy-Weinberg equilibrium test statistics, *Genetics*, **180**, 1609–1616.

Romano, J. P. (1988), A bootstrap revival of some nonparametric distance tests, *J. Amer. Statist. Assoc.*, **83**, 698–708.

Royston, P. (1992), The use of cusums and other techniques in modelling continuous covariates in logistic regression, *Statistics in Medicine*, **11**, 1115–1129.

Ruben, H. (1974), Non-central chi-square and gamma revisited, *Commun. Statist.*, **3**, 607–633.

Rutherford, E., Geiger, H., and Bateman, H. (1910), The probability variations in the distribution of α-particles, *Philos. Mag. (Ser. 6)*, **20**, 698–707.

Sepkoski, Jr., J. J. and Rex, M. A. (1974), Distribution of freshwater mussels: coastal rivers as biogeographic islands, *Systematic Zoology*, **23**, 165–188.

Sham, P. (2001), *Statistics in Human Genetics*, London: Arnold.

Shoemaker, J., Painter, I., and Weir, B. (1998), A Bayesian characterization of Hardy-Weinberg disequilibrium, *Genetics*, **149**, 2079–2088.

Steele, M. and Chaseling, J. (2006), Powers of discrete goodness-of-fit statistics for a uniform null against a selection of alternative distributions, *Comm. Statist. Simul. Comput.*, **35**, 1067–1075.

Stephens, M. A. (1970), Use of the Kolmogorov-Smirnov, Cramér–Von-Mises and related statistics without extensive tables, *J. Roy. Statist. Soc. Ser. B*, **32**, 115–122.

Student (1907), On the error of counting with a haemacytometer, *Biometrika*, **5**, 351–360.

Stute, W. and Zhu, L.-X. (2002), Model checks for generalized linear models, *Scandinavian J. Statist.*, **29**, 535–545.

Su, J. Q. and Wei, L.-J. (1991), A lack-of-fit test for the mean function in a generalized linear model, *J. Amer. Statist. Assoc.*, **86**, 420–426.

Varadhan, S. R. S., Levandowsky, M., and Rubin, N. (1974), *Mathematical Statistics*, Lecture Notes Series, New York: Courant Institute of Mathematical Sciences, NYU.

Wasserman, L. (2003), *All of Statistics*, Springer.

Weinberg, W. (1908), Über den nachweis der vererbung beim menschen, *Jh. Ver. Vaterl. Naturk. Wurttemb.*, **64**, 369–382.

Weising, K. (2005), *DNA Fingerprinting in Plants: Principles, Methods, and Applications*, Boca Raton, Florida: Taylor and Francis.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., and Mor, V. (2004), Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder, *Pharmacoepidem. Drug Safety*, **14**, 227–238.

Wigginton, J., Cutler, D., and Abecasis, G. (2005), A note on exact tests of Hardy-Weinberg equilibrium, *Am. J. Hum. Genet.*, **76**, 887–893.

Zipf, G. K. (1935), *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, Boston, Massachusetts: Houghton Mifflin.