A comparison of the discrete Kolmogorov-Smirnov statistic and the Euclidean distance

Jacob Carruth, Mark Tygert, and Rachel Ward

June 27, 2012

Abstract

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 Euclidean distance between the model and the empirical distributions. In the present paper, both mathematical analysis and its illustration via various data sets indicate that the Kolmogorov-Smirnov statistic tends to be more powerful than the Euclidean distance when there is a natural ordering for the values that the draws can take — that is, when the data is ordinal — whereas the Euclidean 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.

Keywords: significance, hypothesis, chi-square, root-mean-square, mean-square

Contents

1 Introduction

Testing goodness-of-fit is one of the foundations of modern statistics, as elucidated by Rao (2002), for example. The formulation in the discrete setting involves n independent and identically distributed (i.i.d.) draws from a probability distribution over m bins ("categories," "cells," and "classes" are common synonyms for "bins"). In accordance with the standard conventions, we will use p to denote the actual (unknown) underlying distribution of the draws; $p = (p^{(1)}, p^{(2)}, \ldots, p^{(m)})$, with $p^{(1)}$, $p^{(2)}$, $\ldots$, $p^{(m)}$ being nonnegative and

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

We will use p_0 to denote a user-specified distribution, usually called the "model"; again $p_0 = (p_0^{(1)}, p_0^{(2)}, \ldots, p_0^{(m)})$, with $p_0^{(1)}$, $p_0^{(2)}$, $\ldots$, $p_0^{(m)}$ being nonnegative and

$$\sum_{j=1}^{m} p_0^{(j)} = 1. \tag{2}$$

A goodness-of-fit test produces a value — the "P-value" — that gauges the consistency of the observed data with the assumption that $p = p_0$. 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 the assumption that $p = p_0(\hat{\theta})$, where $\hat{\theta}$ is an estimate (taken to be the maximum-likelihood estimate throughout the present paper). We now review the definition of P-values.

P-values are defined via the empirical distribution $\hat{p}$, where $\hat{p} = (\hat{p}^{(1)}, \hat{p}^{(2)}, \ldots, \hat{p}^{(m)})$, 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 $p = p_0(\hat{\theta})$. 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 $p = p_0(\hat{\theta})$.

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

$$d_1(a, b) = \max_{1 \le k \le m} \left| \sum_{j=1}^{k} a^{(j)} - \sum_{j=1}^{k} b^{(j)} \right| \tag{3}$$

and

$$d_2(a,b) = \sqrt{\sum_{j=1}^{m} (a^{(j)} - b^{(j)})^2}, \tag{4}$$

respectively. The P-value for the Kolmogorov-Smirnov statistic is the probability that $d_1(\hat{P}, p_0(\hat{\Theta})) \geq d_1(\hat{p}, p_0(\hat{\theta}))$; the P-value for the Euclidean distance is the probability that $d_2(\hat{P}, p_0(\hat{\Theta})) \geq d_2(\hat{p}, p_0(\hat{\theta}))$. When evaluating the probabilities, we view $\hat{P}$ and $\hat{\Theta}$ as random variables, constructed with i.i.d. draws from the assumed distribution $p = p_0(\hat{\theta})$, 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 (3) is the maximum absolute difference between cumulative distribution functions. The Kolmogorov-Smirnov statistic depends on the ordering of the bins, unlike the Euclidean 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 Euclidean distance is more reliable and more easily understood than the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order). Unlike the Euclidean 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 Euclidean distance and χ^2 statistics are available in Perkins et al. (2011a).

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 paper focuses on the popular Kolmogorov-Smirnov statistic; the Cramér–von-Mises, Watson, and Kuiper variants are very similar.

The remainder of the present paper has the following structure: Section 2 describes how the Euclidean distance is generally preferable to the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order) of the bins. Section 3 describes how the Kolmogorov-Smirnov statistic is generally preferable to the Euclidean distance when there is a natural ordering of the bins. Section 4 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 4, referring back to earlier sections as needed.

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

The Euclidean distance is generally preferable to the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order) of the bins. As discussed by Perkins et al. (2011b), the interaction of parameter estimation and the Euclidean 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 Euclidean 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 2.1. *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,*

$$\mathbf{E} \max_{1 \le k \le m} \left| \sum_{j=1}^{k} D^{(j)} \right| \Big/ \sqrt{m} \quad \longrightarrow \quad \sqrt{\pi/2} \cdot \ln(2) \tag{5}$$

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

We denote by p the actual underlying distribution of the n observed i.i.d. draws. 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 (1) and (2) 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 Euclidean distance is

$$U = \sqrt{\sum_{j=1}^{m} (\hat{P}^{(j)} - p_0^{(j)})^2}. \tag{6}$$

The Kolmogorov-Smirnov statistic is

$$V = \max_{1 \le k \le m} \left| \sum_{j=1}^{k} (\hat{P}^{(j)} - p_0^{(j)}) \right|. \tag{7}$$

The expected value of the square of the Euclidean distance is

$$\mathbf{E}\, U^2 = \sum_{j=1}^{m} \mathbf{E}\, (\hat{P}^{(j)} - p_0^{(j)})^2 = \sum_{j=1}^{m} \frac{p_0^{(j)}}{n} = \frac{1}{n}. \tag{8}$$

As shown, for example, by Durbin (1972) using Lemma 2.1 above, the expected value of $\sqrt{n}$ times the Kolmogorov-Smirnov statistic is

$$\mathbf{E}\, V\sqrt{n} \to \sqrt{\pi/2} \cdot \ln(2) \approx .8687 \tag{9}$$

in the limit that $n \to \infty$ and $\max_{1 \le j \le m} p_0^{(j)} \to 0$. Comparing (8) and (9), 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 Euclidean 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

$$d^{(j)} = p^{(j)} - p_0^{(j)} \tag{10}$$

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

$$u = \sqrt{\sum_{j=1}^{m} (d^{(j)})^2}. \tag{11}$$

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

$$v = \max_{1 \le k \le m} \left| \sum_{j=1}^{k} d^{(j)} \right|. \tag{12}$$

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,

$$|d^{(j)}| = c_m \tag{13}$$

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 (10), (1), and (2) yields that

$$\sum_{j=1}^{m} d^{(j)} = 0. \tag{14}$$

Together, (14) and (13) 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 (13) and (11) yields that the Euclidean distance is

$$u = \sqrt{m} \cdot c_m. \tag{15}$$

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 (12) could be as small as c_m

or as large as $m \cdot c_m/2$, depending on the ordering of the signs in $d^{(1)}$, $d^{(2)}$, ..., $d^{(m)}$. If all orderings are equally likely (which is equivalent to ordering the bins uniformly at random), then by Lemma 2.1 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 (14)).

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 Euclidean distance and the Kolmogorov-Smirnov statistic have similar statistical power on average, if all orderings of the bins are equally likely. However, the Euclidean 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 Euclidean distance is more reliable than the Kolmogorov-Smirnov statistic when there is no especially natural ordering for the bins.

Remark 2.2. 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 Euclidean distance is the l^2 distance), as remarked at the top of page 396 of Hoeffding (1965). The resulting statistic is no longer cumulative.

3 The case when the bins have a natural order

The Kolmogorov-Smirnov statistic is often preferable to the Euclidean 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 Euclidean 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 from the model distribution p_0, the Euclidean distance is almost surely $1/\sqrt{n}$,

$$\sqrt{\sum_{j=1}^{m} (\hat{P}^{(j)} - p_0^{(j)})^2} = \frac{1}{\sqrt{n}}, \tag{16}$$

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 $j = 1, 2, \ldots, m$, $\hat{P}^{(j)} = 0$ 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 paper). 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.

4 Data analysis

This section gives four examples illustrating the performance of the Kolmogorov-Smirnov statistic and the Euclidean distance in various circumstances. The Kolmogorov-Smirnov statistic is more powerful than the Euclidean distance in the first two examples, for which there are natural orderings of the bins. The Euclidean 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 Remark 3.3 of Perkins et al. (2011a). Remark 3.4 of Perkins et al. (2011a) 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.

4.1 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

$$p_0^{(j)} = 2^{-32} \tag{17}$$

for $j = 1, 2, \ldots, 2^{32}$. 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

$$\hat{p}^{(j)} = \begin{cases} 1/n, & j = 1,\ 2,\ \ldots,\ n \\ 0, & j = n+1,\ n+2,\ \ldots,\ 2^{32} \end{cases} \tag{18}$$

for $j = 1, 2, \ldots, 2^{32}$. For these observations, the P-value for the Euclidean distance 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 Euclidean distance has almost no discriminative power for such sparse data, whereas the Kolmogorov-Smirnov statistic easily discerns that the data (18) is inconsistent with the model (17).

Remark 4.1. Like the Euclidean 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 2 of Perkins et al. (2011a).

4.2 A test of Poissonity

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

$$p_0^{(j)} = \frac{100^j}{j! \cdot \exp(100)} \tag{19}$$

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

$$\hat{p}^{(j)} = \begin{cases} 1/10, & j = 100, 101, 102, \ldots, 109 \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

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

- Kolmogorov-Smirnov: .0075
- Euclidean 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 2 of Perkins et al. (2011a). 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 1 plots the model probabilities $p_0^{(0)}, p_0^{(1)}, p_0^{(2)}, \ldots$ defined in (19) along with the observed proportions $\hat{p}^{(0)}, \hat{p}^{(1)}, \hat{p}^{(2)}, \ldots$ defined in (20). Figure 2 plots the model probabilities $p_0^{(0)}, p_0^{(1)}, p_0^{(2)}, \ldots$ along with analogues of the proportions $\hat{p}^{(0)}, \hat{p}^{(1)}, \hat{p}^{(2)}, \ldots$ for a simulation generating 10 i.i.d. draws according to the model.

Figure 3 plots the cumulative model probabilities $p_0^{(0)}, \; p_0^{(0)} + p_0^{(1)}, \; p_0^{(0)} + p_0^{(1)} + p_0^{(2)}, \; \ldots$ along with the cumulative observed proportions $\hat{p}^{(0)}, \; \hat{p}^{(0)} + \hat{p}^{(1)}, \; \hat{p}^{(0)} + \hat{p}^{(1)} + \hat{p}^{(2)}, \; \ldots$. Figure 4 plots the cumulative model probabilities $p_0^{(0)}, \; p_0^{(0)} + p_0^{(1)}, \; p_0^{(0)} + p_0^{(1)} + p_0^{(2)}, \; \ldots$ along with analogues of the cumulative proportions $\hat{p}^{(0)}, \; \hat{p}^{(0)} + \hat{p}^{(1)}, \; \hat{p}^{(0)} + \hat{p}^{(1)} + \hat{p}^{(2)}, \; \ldots$ for the simulation generating 10 i.i.d. draws according to the model.

4.3 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

$$p_0^{(j,k)}(\theta_1, \theta_2, \ldots, \theta_9) = \begin{cases} 2 \cdot \theta_j \cdot \theta_k, & j > k \\ (\theta_k)^2, & j = k \end{cases} \tag{21}$$

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

$$\sum_{j=1}^{9} \theta_j = 1, \tag{22}$$

where the parameters $\theta_1, \theta_2, \ldots, \theta_9$ are the proportions of the nine Rhesus haplotypes in the population (naturally, their maximum-likelihood estimates are the proportions of the haplotypes in the given data). For $j, k = 1, 2, \ldots, 9$ with $j \geq k$, 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 $\theta_1, \theta_2, \ldots, \theta_9$.

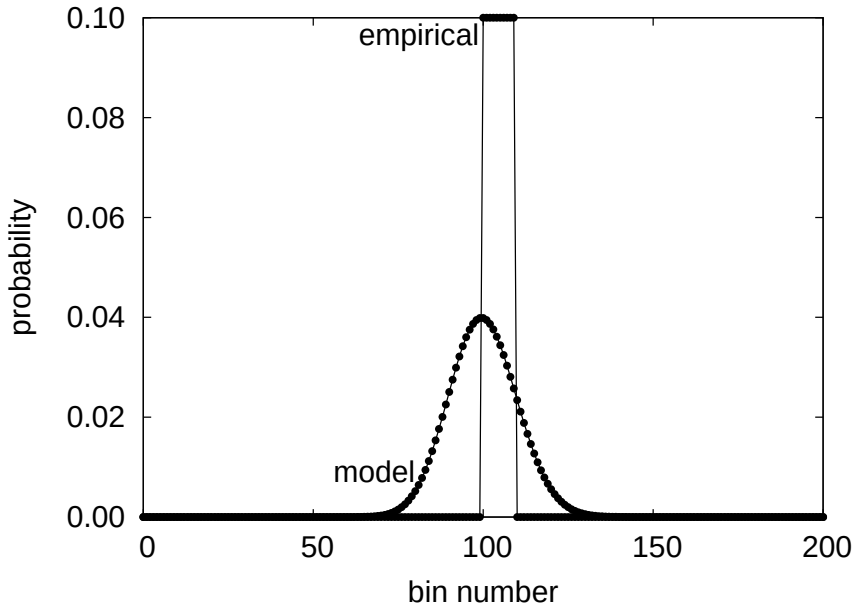

Figure 1: Proportions associated with the bins for the observations

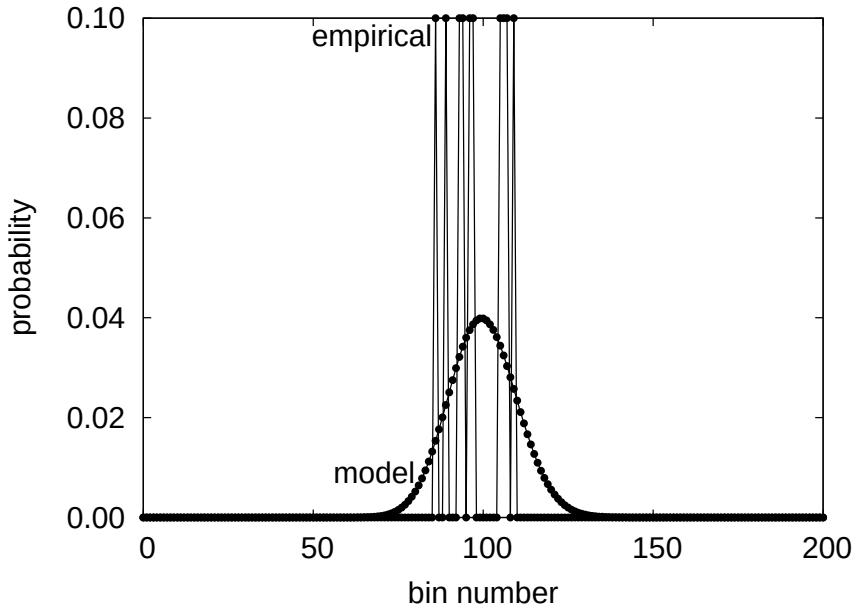

Figure 2: Proportions associated with the bins for a simulation

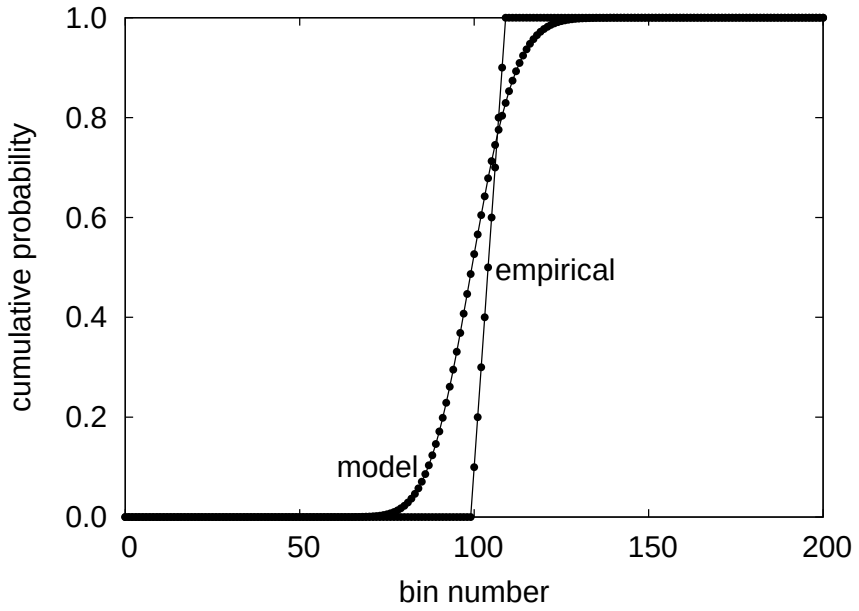

Figure 3: Cumulative proportions associated with the bins for the observations

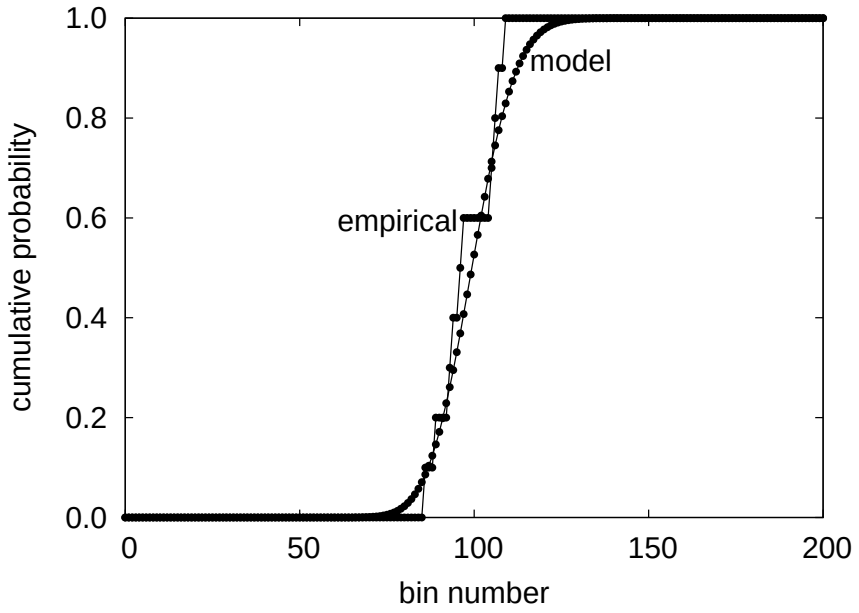

Figure 4: Cumulative proportions associated with the bins for the simulation from Figure 2

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 (21); if a goodness-of-fit statistic rejects the model with high confidence, then we can be confident that mating has not been suitably random.

Table 1 provides data on $n = 8297$ individuals; we duplicated Figure 3 of Guo and Thompson (1992) to obtain Table 1. Figure 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 5, the order of the bins is the lexicographical ordering, namely $(1,1)$, $(2,1)$, $(2,2)$, $(3,1)$, $(3,2)$, $(3,3)$, $\ldots$, $(9,9)$. The nine trials $t = 2, 3, \ldots, 10$ displayed in Figure 5 use pseudorandom orderings of the bins. Please note that the Euclidean distance does not depend on the ordering.

Generally, a more powerful statistic produces lower P-values. In Figure 5, the P-values for the Kolmogorov-Smirnov statistic are sometimes lower, sometimes higher than the P-values for the Euclidean distance. There is no particularly natural ordering of the bins for Figure 5; Figure 5 displays 10 different orderings corresponding to 10 different trials. Figure 5 demonstrates that the Euclidean distance is more reliable than the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order) for the bins.

Remark 4.2. 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

- Euclidean 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 4.5 of Perkins et al. (2011a). Like the Euclidean distance, the χ^2, G^2, and Freeman-Tukey statistics are all invariant to the ordering of the bins.

4.4 A test of uniformity

Table 2 duplicates Table 1 of Gilchrist (2010), giving the colors of the $n = 62$ pieces of candy in a 2.17 ounce bag. Figure 6 plots the P-values for Table 2 to be consistent up to expected random fluctuations with Table 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 $t = 1, 2, \ldots, 10$ displayed in Figure 6 use pseudorandom orderings of the bins. The Euclidean distance does not depend on the ordering.

Generally, a more powerful statistic produces lower P-values. In Figure 6, the P-values for the Kolmogorov-Smirnov statistic are sometimes lower, sometimes higher than the P-values for the Euclidean distance. There is no particularly natural ordering of the bins for Table 3; Figure 6 displays 10 different pseudorandom orderings corresponding to 10 different trials. Figure 6 illustrates that the Euclidean distance is more reliable than the Kolmogorov-Smirnov statistic when there is no natural ordering (or partial order) for the bins.

Table 1: Frequencies of pairs of Rhesus haplotypes

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

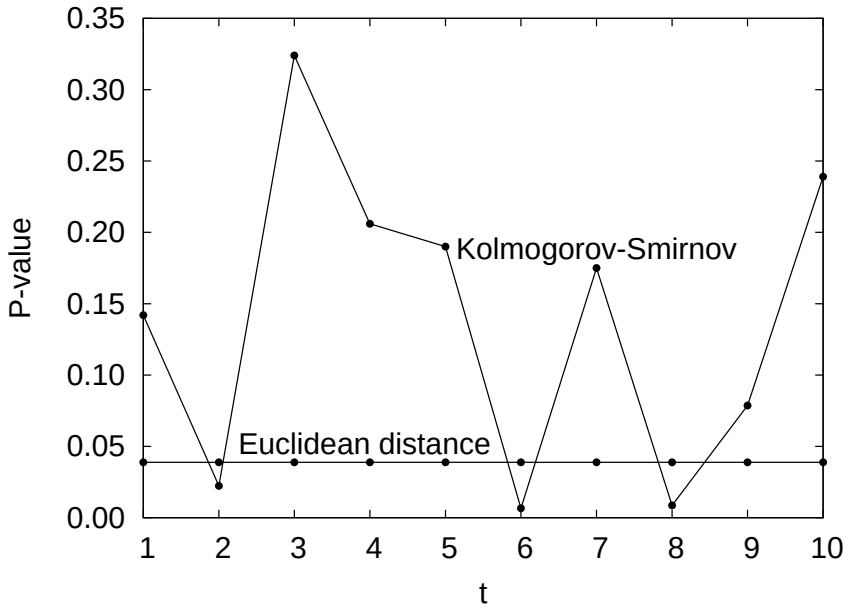

Figure 5: P-values for Table 1 to be consistent with formula (21)

Table 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 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

Remark 4.3. Table 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 2.2.

Remark 4.4. Like the Euclidean 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:

- Euclidean 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 2 of Perkins et al. (2011a). For this example, the Euclidean distance and the χ^2 statistic produce exactly the same P-values: for the model of homogeneous proportions, displayed in Table 3, the Euclidean distance is directly proportional to the square root of the χ^2 statistic, and hence the Euclidean distance is a strictly increasing function of χ^2.

Acknowledgements

We would like to thank Alex Barnett, Gérard Ben Arous, James Berger, Tony Cai, Sourav Chatterjee, Ronald Raphael Coifman, Ingrid Daubechies, Jianqing Fan, Jiayang Gao, Andrew Gelman, Leslie Greengard, Peter W. Jones, Deborah Mayo, Peter McCullagh, Michael O'Neil, Ron Peled, William Perkins, William H. Press, Vladimir Rokhlin, Joseph Romano, Gary Simon, Amit Singer, Michael Stein, Stephen Stigler, Joel Tropp, Larry Wasserman, and Douglas A. Wolfe. This work was supported in part by Alfred P. Sloan Research Fellowships, a Donald D. Harrington Faculty Fellowship, and a DARPA Young Faculty Award.

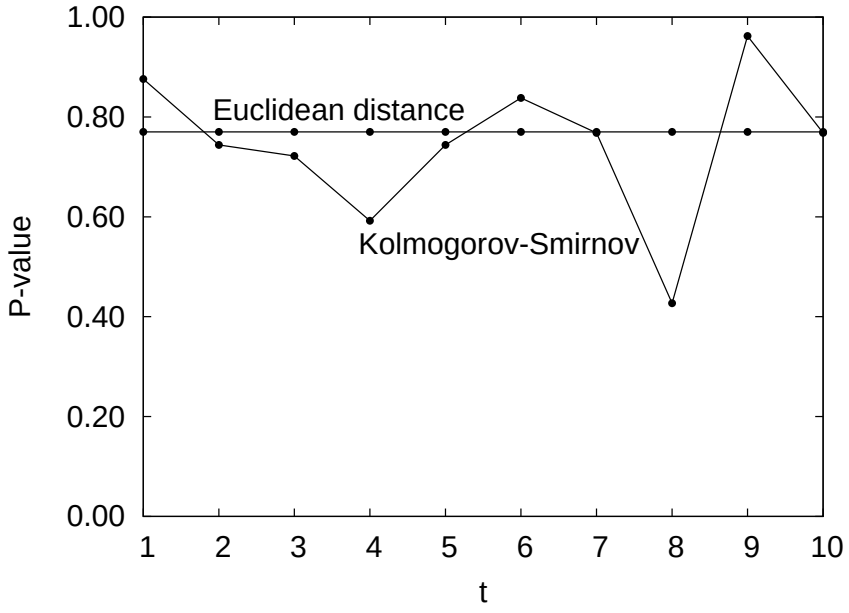

Figure 6: P-values for Table 2 to be consistent with the model displayed in Table 3

References

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.

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.

Choulakian, V., Lockhart, R. A., and Stephens, M. A. (1994), Cramér–von-Mises statistics for discrete distributions, *Canadian J. Statist.*, **22**, 125–137.

D'Agostino, R. B. and Stephens, M. A. (1986), *Goodness-of-Fit Techniques*, New York: Marcel Dekker.

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.

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.

Gilchrist, E. (2010), A sweet approach to teaching the one-variable chi-square test, *Communication Teacher*, **24**, 14–18.

Guo, S. W. and Thompson, E. A. (1992), Performing the exact test of Hardy-Weinberg proportion for multiple alleles, *Biometrics*, **48**, 361–372.

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.

Hoeffding, W. (1965), Asymptotically optimal tests for multinomial distributions, *Ann. Math. Statist.*, **36**, 369–401.

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.

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., Tsang, W. W., and Wang, J. (2003), Evaluating Kolmogorov's distribution, *J. Statist. Soft.*, **8**, 1–4.

Perkins, W., Tygert, M., and Ward, R. (2011a), χ^2 and classical exact tests often wildly misreport significance; the remedy lies in computers, Tech. Rep. 1108.4126, arXiv, http://cims.nyu.edu/∼tygert/abbreviated.pdf.

Perkins, W., Tygert, M., and Ward, R. (2011b), Computing the confidence levels for a root-mean-square test of goodness-of-fit, II, Tech. Rep. 1009.2260, arXiv.

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.

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.

Rényi, A. (1953), On the theory of order statistics, *Acta Math. Acad. Sci. Hungar.*, **4**, 191–231.

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.