On Rounding Percentages

Persi Diaconis; David Freedman

Journal of the American Statistical Association, Vol. 74, No. 366. (Jun., 1979), pp. 359-364.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197906%2974%3A366%3C359%3AORP%3E2.0.CO%3B2-E

Journal of the American Statistical Association is currently published by American Statistical Association.

On Rounding Percentages

PERSI DIACONIS and DAVID FREEDMAN*

We assess the probability that a table of rounded percentages adds to 100 percent. This extends work of Mosteller, Youtz, and Zahn (1967) who found that the chance of rounding to 100 percent was about $\frac{3}{4}$ with three categories, $\frac{2}{3}$ with four categories, and $(6/\pi c)^{\frac{1}{2}}$ with a large number c of categories. We give a mathematical treatment of this phenomenon when the table is drawn from a multinomial distribution or from a mixture of multinomial distributions. We discuss the very different small-sample behavior and treat Benford's leading digit data as an example.

KEY WORDS: Rounding error; Table of counts; Multinomial; Mixture of multinomials; Limit theorem; Leading digit.

1. INTRODUCTION

Sums of rounded proportions often fail to add to 1. For example, consider Table 1. This table has three *categories*. The number of categories is denoted by c. The proportions are rounded to the nearest .001 or 1 in 1,000. This 1,000 is the *rounding number* n. In Table 1, the rounded proportions add to 1.001 instead of 1. The chance that rounded proportions add to 1 depends on the number c of categories, the rounding number n, and the probability model for generating the table. Our object in this article is to compute this chance.

Failure to add to 1 occurs so frequently that if many sums of proportions add to exactly 1 in a reported set of tables, one begins to suspect the reporter of forcing the proportions to add to 1. An example is discussed in Section 6. Sometimes altering the sample proportions in this way matters: For instance, it can cause large changes in the chi-squared statistic, as shown by an example in Section 6.

In their 1967 article, Mosteller, Youtz, and Zahn (MYZ) proposed several probability models for generating tables. They computed the chance that the rounded proportions add to 1. They concluded that the chance did not depend very much on the rounding number but did depend strongly on the number of categories: They found that with two categories the chance is 100 percent, with three categories the chance is 75 percent, with four categories the chance is 66.66 percent, and with a large number c of categories the chance is about $(6/\pi c)^{\frac{1}{2}}$. Although persuasive and backed by extensive empirical evidence (rounding behavior of 565 tables in the National Halothane Study), the MYZ argument was only heuristic.

In this article we give a rigorous treatment of the main mathematical issues raised by MYZ. To begin with, we consider the multinomial model for generating tables. Due to the law of large numbers the sample proportions round essentially the same way as the theoretical probabilities. Thus, the chance of getting rounded sample proportions that total 1 tends to 0 or 1 as the sample size tends to infinity. The same type of behavior holds for many other models for generating a single table. This analysis is given in Section 2.

Next consider a collection of many tables. Suppose the jth table is drawn from a multinomial distribution with theoretical probability vector $\bar{p}(j) = (p_1(j), \ldots, p_c(j))$ and sample size N_j. It is natural to consider models in which the different $\bar{p}(j)$'s are randomly chosen from a distribution on the simplex

$$\{\bar{p} \mid p_i \geq 0, i = 1, 2, \ldots, c; \sum_{i=1}^{c} p_i = 1\} . \qquad (1)$$

If the sample size in the jth table is large, the proportions in the jth sample will round in the same way as $\bar{p}_j$. Thus, the number of tables that round to 1 will be close to the probability that the random vector $\bar{p}$ rounds to 1.

The principal result of this article concerns the MYZ broken-stick model in which a uniform distribution is put on the simplex (1). It is convenient to introduce the symbol $r_n(x)$ for the result of rounding x to the nearest $1/n$.

If $\dfrac{m+.5}{n} < x \leq \dfrac{m+1}{n}$, then $r_n(x) = \dfrac{m+1}{n}$;

if $\dfrac{m}{n} \leq x < \dfrac{m+.5}{n}$, then $r_n(x) = \dfrac{m}{n}$;

if $x = \dfrac{m+.5}{n}$, then $r_n(x) = \dfrac{m}{n}$ or $\dfrac{m+1}{n}$ according as m is even or odd.

A discussion of $r_n(x)$ can be found in Wallis and Roberts (1956, p. 175). In the broken-stick model, the sum of the rounded proportions $\sum_{i=1}^{c} r_n(p_i)$ is a random variable with possible values $1, 1 \pm 1/n, 1 \pm 2/n$, and so on. We want to find the chance that $\sum_{i=1}^{c} r_n(p_i) = 1$. To solve

* Persi Diaconis is Associate Professor, Department of Statistics, Stanford University, Stanford, CA 94305, and member, technical staff, Bell Laboratories, 1978–79. David Freedman is Professor of Statistics, Department of Statistics, University of California, Berkeley, CA 94720. This work was supported in part by NSF Grant MPS 74-21416 and by the Energy Research and Development Administration under Contract EY-76-C-03-0515 and (second author) by NSF Grant GP-43085.

© Journal of the American Statistical Association
June 1979, Volume 74, Number 366
Theory and Methods Section

1. Distribution of White Families by Type, United States, 1970

Type	Number (1,000)	Proportion
Husband-Wife	40,802	0.887
Other Male Head	1,036	0.023
Female Head	4,185	0.091
Total	46,023	1.001

Source: *Statistical Abstract*, 1976, Table 53.

this problem it is convenient to introduce $c - 1$ mutually independent random variables V_i, each of which is uniformly distributed over the interval $[-.5, .5]$. Theorem 1 shows that when the rounding number n is large, $\sum_{i=1}^{c} r_n(p_i) - 1$ has approximately the same distribution as $\{r_1(V_1 + \ldots + V_{c-1})\}/n$. In particular, the chance that the rounded proportions sum to 1 converges to the chance that $-.5 < \sum_{i=1}^{c-1} V_i < .5$. The second theorem shows that this last chance is approximately equal to $(6/\pi(c-1))^{\frac{1}{2}}$. These theorems verify a conjecture on p. 857 of MYZ.

Theorem 1: Suppose $p_1, \ldots, p_c$ are uniformly distributed over the simplex (1). As the rounding number n approaches infinity, $n\{\sum_{i=1}^{c} r_n(p_i) - 1\}$ converges in distribution to $r_1(\sum_{i=1}^{c-1} V_i)$ where V_i are independent and uniformly distributed on $[-.5, .5]$.

Theorem 1 will be proved in Section 3. The assumption that $p_1, \ldots, p_c$ are uniformly distributed over the simplex (1) is not critical. Any absolutely continuous distribution will do, as shown in Section 4.

When c is large, the Edgeworth expansion may be used to approximate the distribution of $r_1(V_1 + \ldots + V_{c-1})$.

Theorem 2: Suppose $V_1, \ldots, V_{c-1}$ are independent and uniformly distributed over $[-.5, .5]$. Let $c \to \infty$ and let $j = 0(c^{\frac{1}{2}})$. Then $r_1(V_1 + \ldots + V_{c-1}) = j$ with probability $(6/\pi(c-1))^{\frac{1}{2}} e^{-6j^2/(c-1)} + 0(1/c^{\frac{3}{2}})$. In particular, $-.5 < V_1 + \ldots + V_{c-1} < .5$ with probability $(6/\pi(c-1))^{\frac{1}{2}} + 0(1/c^{\frac{3}{2}})$.

In Section 5 we look at the small-sample behavior of the models discussed before. This behavior can be quite different from what large-sample theory predicts. The final section contains an application of our analysis to Benford's (1938) leading digit data.

2. FIXED CELL PROBABILITY MODELS

First consider the model in which $X_1, X_2, \ldots, X_c$ have a joint multinomial distribution with theoretical probabilities $p_1, p_2, \ldots, p_c$ and sample size N. Then, by the law of large numbers as in Feller (1968, p. 152), the vector of sample proportions $(X_1/N, \ldots, X_c/N)$ must concentrate in smaller and smaller neighborhoods of $(p_1, \ldots, p_c)$ as $N \to \infty$. Thus, $\sum_{i=1}^{c} r_n(X_i/N)$ behaves like the constant $\sum_{i=1}^{c} r_n(p_i)$. There is one exceptional case that is discussed at the end of this section. It follows that for fixed n and c, the rounded proportions add to a constant multiple of $1/n$ with probability tending to 1 as $N \to \infty$.

Similar results hold for many other models. For example, suppose $X_1, \ldots, X_c$ are independent Poisson variables with parameters $\lambda_1, \ldots, \lambda_c$. Let $N = \sum_{i=1}^{c} X_i$, and $\Lambda = \sum_{i=1}^{c} \lambda_i$. If λ_i goes to infinity in such a way that λ_i/Λ tends to a limit, a standard argument shows that $(X_1/N, \ldots, X_c/N)$ tends to the limit of $(\lambda_1/\Lambda, \ldots, \lambda_c/\Lambda)$ in probability.

We have been ignoring one exceptional case that we discuss in the multinomial model for definiteness. Suppose $p_1, \ldots, p_c$ is in the simplex (1), the rounding number is n, and $np_i - [np_i] = .5$ for one or more i's. If X_i/N exceeds p_i, we have to round up; if X_i/N is below p_i, we have to round down. This introduces some randomness into the asymptotic distribution of $\sum_1^c r_n(X_i/N)$.

An example will make the issue clear. Suppose $c = 3$, $p_1 = .35$, $p_2 = .45$, $p_3 = .20$, and $n = 10$. The rounded p_i's are .4, .4, .2, and sum to 1. For large N, X_3/N is nearly .2 and rounds to .2; so the asymptotic behavior of $\sum_1^3 r_{10}(X_i/N)$ depends on whether X_i/N is just above or just below p_i for $i = 1, 2$. The four possibilities are shown in Table 2.

2. Four Cases: $\sum_1^3 r_{10}(X_i/N)$

	$.40 < X_2/N < .45$	$.45 < X_2/N < .50$
$.30 < X_1/N < .35$	.3 + .4 + .2 = .9	.3 + .5 + .2 = 1
$.35 < X_1/N < .40$	.4 + .4 + .2 = 1	.4 + .5 + .2 = 1.1

Asymptotically, the two variables

$$\frac{X_1 - Np_1}{(Np_1(1-p_1))^{\frac{1}{2}}}, \quad \frac{X_2 - Np_2}{(Np_2(1-p_2))^{\frac{1}{2}}}$$

are jointly normal, with mean 0, variance 1, and correlation

$$-\left[\frac{p_1 p_2}{(1-p_1)(1-p_2)}\right]^{\frac{1}{2}} \approx -.66 .$$

The chance that the rounded proportions sum to 1 therefore converges to the chance that two centered normal variables with correlation $-.66$ are of opposite sign.

3. PROOF OF THE MAIN THEOREM IN THE BROKEN-STICK MODEL

In this model, the number of categories c is fixed, $p_1, \ldots, p_c$ are uniformly distributed over the simplex (1), the rounding number n is large, and $r_n(x)$ denotes x rounded to the nearest $1/n$. The random variables $V_1, \ldots, V_{c-1}$ referred to in Theorem 1 may be taken as the rounding errors, defined as follows, for $1 \leq i \leq c-1$:

$$r_n(p_i) = p_i + V_i/n . \tag{3.1}$$

So far, the V_i are neither independent nor uniform. As is easily seen, however,

Lemma 1: If p_i is uniformly distributed over

$[m/n, (m+1)/n]$, then V_i is uniformly distributed over $[-\frac{1}{2}, \frac{1}{2}]$.

These rounding errors V_i are defined only for $1 \leq i \leq c-1$; now the rounding error for p_c must be considered. Here is a preliminary fact:

$$r_n((m+\sigma)/n) = m/n + r_1(\sigma)/n \quad \text{for integer } m \text{ and real } \sigma \ . \tag{3.2}$$

Now (3.1) implies

$$p_c = 1 - \sum_1^{c-1} p_i = 1 - \sum_1^{c-1} r_n(p_i) + \frac{1}{n}\sum_1^{c-1} V_i \ .$$

Using relation (3.2),

$$r_n(p_c) = 1 - \sum_1^{c-1} r_n(p_i) + \frac{1}{n} r_1(\sum_1^{c-1} V_i) \ .$$

Thus,

$$r_n(p_c) = p_c + \frac{1}{n} r_1(\sum_1^{c-1} V_i) - \frac{1}{n}\sum_1^{c-1} V_i \ . \tag{3.3}$$

Summing relation (3.1) for $i = 1$ to $c-1$ and then adding the relation (3.3) to this sum, the term $1/n \sum_1^{c-1} V_i$ cancels. This proves Lemma 2.

Lemma 2: If $r_n(x)$ is x rounded to the nearest $1/n$, and the rounding errors V_i are defined by (3.1), then $\sum_1^c r_n(p_i) = 1 + (1/n)\, r_1(\sum_1^{c-1} V_i)$.

So far, the assumption that the p_i's are uniformly distributed has not been used; it will be needed to compute the distribution of the V_i. Let $m_1, \ldots, m_{c-1}$ be nonnegative integers whose sum is at most $n-(c-1)$. Let $A(m_1, \ldots, m_{c-1})$ be the event that

$$\frac{m_i}{n} \leq p_i < \frac{m_i+1}{n} \quad \text{for} \quad 1 \leq i \leq c-1 \ .$$

Let A_n be the union of these $A(m_1, \ldots, m_{c-1})$ over all choices of $m_1, \ldots, m_{c-1}$. The probability of A_n tends to 1 as $n \to \infty$. In fact, a geometric argument that we omit shows that the chance of A_n is $\{n(n-1)\ldots(n-c+2)\}/n^{c-1}$. Theorem 1, then, can be proved by demonstrating that given A_n, the random variables $V_1, \ldots, V_{c-1}$ are conditionally independent and uniformly distributed over $[-.5, .5]$. In fact, a little more is true:

Lemma 3: Given $A(m_1, \ldots, m_{c-1})$, the random variables $V_1, \ldots, V_{c-1}$ are conditionally independent and uniformly distributed over $[-.5, .5]$.

The proof of Lemma 3 is direct: The distribution of $(p_1, \ldots, p_{c-1})$ is uniform over the region

$$x_i \geq 0 \quad \text{for} \quad 1 \leq i \leq c-1 \ , \quad \sum_1^{c-1} x_i \leq 1 \ . \tag{3.4}$$

Also, the hypercube defining $A(m_1, \ldots, m_{c-1})$ is wholly contained in this region. Therefore, given $A(m_1, \ldots, m_{c-1})$, the first $c-1$ of the p_i's are independent, each being uniformly distributed over its edge of the hypercube. Now Lemma 1 completes the proof. This completes the argument for Theorem 1.

Recall that the argument connecting the distribution of the rounded p_i to tables of counts is as follows: If we observe a multinomial vector $(X_1, X_2, \ldots, X_c)$ drawn with parameters $(p_1, p_2, \ldots, p_c)$ and N then, because of the law of large numbers, the sample proportions will be close enough to the vector of p_i's so that the sample proportions round in the same way that the p_i's round.

When the p_i's are chosen from the uniform distribution of the simplex (1), another approach is available. The distribution of $(X_1, X_2, \ldots, X_c)$ averaged over the simplex (1) follows so-called Bose-Einstein statistics (Feller 1968, p. 40; Hill 1970). Under Bose-Einstein statistics, all partitions of N into c parts are equally likely. Thus, the possible sample vectors are all the lattice points in the simplex $\sum_{i=1}^c x_i = N$, $x_i \geq 0$. All points in this simplex are equally likely. The points that round to a fixed multiple of the rounding number are contained in certain fixed regions of this simplex. As N tends to infinity, the proportion of points in a given region tends to the area of the region. It is essentially this area that was computed in Lemma 3. The approach we have followed seems preferable because it leads to the generalization of the next section.

We conclude this section with a proof of Theorem 2.

Proof: Let $m = c-1$. Let $g_m(x)$ be the probability density of $V_1 + \ldots + V_m$. Note that $E(V_1) = E(V_1^3) = 0$, and $\mathrm{var}(V_1) = \frac{1}{12}$. The Edgeworth expansion, as in Section 16.4 of Feller (1971), shows that

$$g_m(x) = \left(\frac{6}{\pi m}\right)^{\frac{1}{2}} e^{-6x^2/m} + O\left(\frac{1}{m^{\frac{3}{2}}}\right), \text{ uniformly in } x \ .$$

Integrating from $j-.5$ to $j+.5$, and setting $y = j + x$ we find that

$$P\{j-.5 < V_1 + \ldots + V_m < j+.5\} = \left(\frac{6}{\pi m}\right)^{\frac{1}{2}} e^{-6j^2/m} \int_{-.5}^{.5} e^{+12jy/m} e^{-6y^2/m} dy + O\left(\frac{1}{m^{\frac{3}{2}}}\right).$$

Now

$$e^{-6y^2/m} = 1 + O\left(\frac{1}{m}\right).$$

By assumption, $j = O(m^{\frac{1}{2}})$: by calculus,

$$\int_{.5}^{.5} e^{12jy/m} dy = 1 + O\left(\frac{1}{m}\right).$$

This completes the proof.

4. THE GENERALIZATION TO THE ABSOLUTELY CONTINUOUS CASE

Theorem 1 is generalized in the following paragraphs.

Theorem 3: Let μ be a probability measure on the simplex (1). Suppose μ is absolutely continuous with respect to the uniform distribution on (1). Then, as the rounding number $n \to \infty$, the μ distribution of $n\{\sum_1^{c-1} r_n(p_i) - 1\}$ converges to the distribution of

$r_1(\sum_1^{c-1} V_i)$, the V_i being independent and uniformly distributed over $[-.5, .5]$.

This can be proved by the argument of the previous section because given that the p_i's lie in a typical small hypercube, their conditional distribution must be almost uniform. To make this precise, it is convenient to use the idea of Lebesgue points (Dunford and Schwartz, 1957, pp. 210–218).

The distribution θ of $p_1, \ldots, p_{c-1}$ is a probability in the region (3.4) and is absolutely continuous with respect to $(c-1)$-dimensional Lebesgue measure λ. As a result, it admits a derivative f. By definition, x is a Lebesgue point of f provided that:

$$\frac{1}{\lambda(C)} \int_C |f(y) - f(x)| \lambda(dy) \to 0$$

as the $(c-1)$-dimensional hypercube C shrinks to x. As Lebesgue proved, almost all x have this property. Of course, if x is a Lebesgue point for f, then

$$\frac{1}{\lambda(C)} \int_C f(y)\lambda(dy) \to f(x)$$

as the $(c-1)$-dimensional hypercube C shrinks to x.

To prove the argument, the following theorem will be useful. Recall that θ is the distribution of $p_1, \ldots, p_{c-1}$ and $f = d\theta/d\lambda$.

Theorem 4: Suppose $f(x) > 0$ and x is a Lebesgue point of f. Let the hypercube C shrink to x. Let λ_C be the uniform distribution over C, and let θ_C be θ conditioned on C. Let $\| \ \|$ denote variation norm. Then $\|\theta_C - \lambda_C\| \to 0$.

Proof: The following computation is standard.

$$\begin{aligned}
\|\theta_C - \lambda_C\| &= \int_C \left| \frac{f(y)}{\theta(C)} - \frac{1}{\lambda(C)} \right| \lambda(dy) \\
&= \frac{1}{\theta(C)} \int_C \left| f(y) - \frac{\theta(C)}{\lambda(C)} \right| \lambda(dy) \\
&\le \frac{1}{\theta(C)} \int_C |f(y) - f(x)| \lambda(dy) \\
&\quad + \frac{1}{\theta(C)} \int_C \left| f(x) - \frac{\theta(C)}{\lambda(C)} \right| \lambda(dy) \\
&= \frac{\lambda(C)}{\theta(C)} \frac{1}{\lambda(C)} \int_C |f(y) - f(x)| \lambda(dy) \\
&\quad + \frac{\lambda(C)}{\theta(C)} \left| f(x) - \frac{\theta(C)}{\lambda(C)} \right| \to 0 \ .
\end{aligned}$$

We shall say that a statement is true for θ-almost all x if it is true for a set of values of x that has probability 1 under θ.

Corollary 1: Under the conditions of Theorem 4, $\|\theta_C - \lambda_C\| \to 0$ as C shrinks to x, for θ-almost all x.

Corollary 1 is closely related to the martingale proofs of the Radon-Nikodym Theorem. Some references that make the connection clear are Blackwell and Dubins (1962) and Meyer (1966, p. 153).

Returning to the argument for Theorem 3, let $[r]$ denote the greatest integer in r, and define a subset B_n of the region (3.4) by the requirement that $\sum_1^{c-1} [nx_i] \le n - (c-1)$. For $x = (x_1, \ldots, x_{c-1})$, $x \in B_n$, let $C(x)$ be the $(c-1)$-dimensional hypercube

$$\left\{ y : \frac{[nx_i]}{n} \le y_i \le \frac{[nx_i] + 1}{n} \quad \text{for} \quad 1 \le i \le c-1 \right\}.$$

As Corollary 1 implies, $\|\theta_{C(x)} - \lambda_{C(x)}\| \to 0$ as $n \to \infty$ for θ-almost all x. In particular, if $\epsilon > 0$, then for all sufficiently large n,

$$\theta\{x : \|\theta_{C(x)} - \lambda_{C(x)}\| > \epsilon\} < \epsilon \ . \tag{4.1}$$

Now the argument for Theorem 1 applies almost verbatim. Because $\|\theta_{C(x)} - \lambda_{C(x)}\|$ is constant over the hypercube $C(x)$, what (4.1) says is that except for a set of hypercubes $A(m_1, \ldots, m_{c-1})$ of total probability ϵ, the conditional distribution of $p_1, \ldots, p_{c-1}$ is within ϵ of being uniform over $A(m_1, \ldots, m_{c-1})$. This completes the argument for Theorem 3.

Remark: There is an easy L^1 argument: Let C run over a partition into hypercubes, then

$$\sum_C \theta(C)\|\theta_C - \lambda_C\| = \|f - f_C\|$$

where $f_C(y) = \theta(C)/\lambda(C)$ for $y \in C$, then the right hand side goes to 0 as mesh $(C) \to 0$: approximate f in L^1 by a smooth f^*, and observe $\|f_C - f_C^*\| < \|f - f^*\|$.

For some approximations, the conditional argument may not be useful. For any probability μ on the simplex (1), the exact distribution of $n\{\sum_1^c r_n(p_i) - 1\}$ may be computed exactly, as follows. Let μ^* be the joint distribution of $np_i - [np_i]$ for $1 \le i \le c-1$, a probability measure on the unit cube K_{c-1} in $(c-1)$-dimensional space. For $x = (x_1, \ldots, x_{c-1}) \in K_{c-1}$, let $U_i(x) = r_n(x_i) - x_i$.

Theorem 5: The μ distribution of $n\{\sum_1^c r_n(p_i) - 1\}$ coincides with the μ^* distribution of $r_1(\sum_1^{c-1} U_i)$.

This is immediate from Lemma 1. The point is that μ^* will be almost uniform over K_{c-1} even for many singular measures μ. An example of the use of μ^* is given at the end of the next section.

5. SMALL-SAMPLE RESULTS

MYZ also consider the multinomial model with $p_1, \ldots, p_c$ fixed rather than random. On p. 856 of their article, MYZ seem to assert that with large samples the multinomial model behaves like the broken-stick model. The argument of Section 2 shows that this cannot be correct.

For example, consider the trinomial with equally likely cells: $c = 3$ and $p_1 = p_2 = p_3 = \frac{1}{3}$. For $n = 10$, or any other decimal rounding, the rounded p's add to one unit less than 1:

$$\tfrac{1}{3} \approx .3 \quad \text{and} \quad .3 + .3 + .3 = 1 - \tfrac{1}{10} \ .$$

If X_1, X_2, X_3 are the counts in this model, as $N \to \infty$, the chance that $\sum_1^3 r_{10}(X_i/N) = 1$ must approach 0.

3. Joint Distribution of ($X_1/100$, $X_2/100$, $X_3/100$) When (X_1,X_2,X_3) Have a Multinomial Distribution With Parameters 100 and ($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$).*

$X_1/100$ \ $X_2/100$	.2					.25					.30					.35					.40					.45					.5
.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
.25	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	2	2	2	2	1	1	1	1	1	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	2	2	2	2	2	1	1	1	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	1	1	1	2	2	3	3	3	3	3	3	2	2	1	1	1	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	1	1	2	2	3	3	4	4	4	4	4	3	3	2	2	1	1	0	0	0	0	0	0
	0	0	0	0	0	0	0	1	1	2	2	3	4	5	5	5	5	5	5	4	3	2	2	1	1	0	0	0	0	0	0
.30	0	0	0	0	0	0	1	1	2	2	3	4	5	6	6	6	6	6	5	4	3	2	2	1	1	0	0	0	0	0	0
	0	0	0	0	0	1	1	1	2	3	4	5	6	7	7	7	7	6	5	4	3	2	1	1	1	0	0	0	0	0	0
	0	0	0	0	0	1	1	2	3	4	5	6	7	8	8	8	7	6	5	4	3	2	1	1	0	0	0	0	0	0	0
	0	0	0	0	1	1	2	2	3	5	6	7	8	8	8	8	7	6	5	3	2	2	1	1	0	0	0	0	0	0	0
	0	0	0	0	1	1	2	3	4	5	6	7	8	8	8	7	6	5	4	3	2	1	1	0	0	0	0	0	0	0	0
.35	0	0	0	0	1	1	2	3	4	5	6	7	8	8	7	6	5	4	3	2	1	1	0	0	0	0	0	0	0	0	0
	0	0	0	1	1	2	2	3	4	5	6	7	7	7	6	5	4	3	2	2	1	1	0	0	0	0	0	0	0	0	0
	0	0	0	1	1	2	2	3	4	5	6	6	6	6	5	4	3	2	2	1	1	0	0	0	0	0	0	0	0	0	0
	0	0	0	1	1	2	2	3	4	5	5	5	5	5	4	3	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	1	1	2	2	3	3	4	4	4	4	3	3	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
.40	0	0	0	1	1	1	2	2	3	3	3	3	3	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	1	1	1	2	2	2	2	2	2	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	1	1	1	2	2	2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

* All entries should be multiplied by 1/1,000.

MYZ's Table 3 shows, however, that for samples of size $N = 1$ to 20, this chance is close to 75 percent. We now explain this.

We began by recomputing the table. To our dismay, it checked perfectly. Some further numerical exploration, however, suggested a tentative answer: From the point of view of rounding calculations, the law of large numbers works very slowly. Even when $N = 100$ in the preceding example, the distribution of ($X_1/100$, $X_2/100$, $X_3/100$) is much closer to uniform than it is to a point mass at ($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$)—and $P\{\sum_1^3 r_{10}(X_i/100) = 1\} = .74$!

The joint distribution of ($X_1/100$, $X_2/100$, $X_3/100$) is shown in Table 3. The values of $X_1/100$ appear at the left of the table; the values of $X_2/100$ are across the top. Of course, $X_3/100 = 1 - X_1/100 - X_2/100$. The corresponding probabilities are reported in the body of the table, rounded to integer multiples of 1/1,000. For instance, the chance that $X_1/100 = .33$ and $X_2/100 = .33$—so that $X_3/100 = .34$ and the rounded proportions add to .9—is about 8/1,000. The chance that $X_1/100 = .33$ and $X_2/100 = .36$—so that $X_3/100 = .31$ and the rounded proportions add to 1—is about 7/1,000. A zero in the table means that the corresponding chance is below .0005. For instance, the chance that $X_1/100 = .25$ and $X_2/100 = .30$ is shown as 0; in fact, it is .0004.

Table 3 demonstrates that ($X_1/100$, $X_2/100$, $X_3/100$) spreads out around ($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$) in a way that really matters when rounding to tenths. The discussion leading up to Theorem 5 suggests examining the measure μ^*, which in

4. Joint Distribution of μ^* When (X_1,X_2,X_3) Have a Multinomial Distribution With Parameters 100 and (1/3, 1/3, 1/3). All Entries Should Be Multiplied by 1/1,000

	$X_2/10 - [X_2/10]$									
	10	10	10	10	10	10	10	10	10	10
	10	10	10	10	10	10	10	10	10	10
	10	10	10	10	11	10	10	10	10	10
	10	10	11	11	11	10	10	10	10	10
$X_1/10 - [X_1/10]$	10	10	10	10	10	10	10	10	10	10
	10	10	10	10	10	10	10	9	10	10
	10	10	10	10	10	10	9	9	10	10
	10	10	10	10	10	10	10	10	10	10
	10	10	10	10	10	10	10	10	10	10

this case amounts to the joint distribution of

$$X_1/10 - [X_1/10] \ , \quad X_2/10 - [X_2/10] \ .$$

This distribution is shown in Table 4, the values of the first variable being given along the left edge, the value of the second variable across the top, and the corresponding probability in the body of the table, rounded to an integer multiple of 1/1,000. It is essentially uniform.

In this example, we have been rounding to tenths. When rounding to halves, for instance, the spread in Table 3 would be relatively small; and $\sum_1^3 r_2(X_i/100) = 1$ with probability only 14 percent, compared with the 75 percent predicted by the broken-stick model on p. 856 of MYZ.

6. AN EXAMPLE

While investigating the behavior of leading digits in typical data, Benford (1938) (also see Diaconis 1977, Raimi 1976, and Ylvisaker 1977) collected a sample of size 20,229 from a total of 20 sources. These data are presented in Table 5. For example, Benford looked at the areas of 335 rivers and found that 31.0 percent of the areas began with 1, 16.4 percent began with 2, and so on.

Each row in Table 5 adds to 100 percent. How likely is this? On the broken-stick model, the chance of a given row rounding to 100 percent is approximately $(6/8\pi)^{\frac{1}{2}} \doteq \frac{1}{2}$. Numerical calculations show that this approximation is quite accurate. Assuming the rows are independent, the chance of all rows simultaneously rounding to 100 percent is astronomically small. We conclude that Benford's table does not follow the broken-stick model—or any of the probability models introduced in Sections 2, 3, or 4. This raises the suspicion that Benford manipulated the data to make the rows round properly. This suspicion is not hard to verify. Consider the first row of Table 5. The percentage of numbers with leading digit 7 is reported as 5.5, with a total of 335 cases. The only proportions compatible with 5.5 are 18/335, which rounds to 5.4, or 19/335, which rounds to 5.7: There is no proportion possible that rounds to 5.5.

The bottom row of averages also rounds to 100 percent. Direct calculation shows that the entries in columns 3 and 9 have been incorrectly rounded. Benford was trying to

5. Benford Data (Percentages)

Group	Title	1	2	3	4	5	6	7	8	9	Count
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3,259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1,389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H. P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Molecular Weight	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1,800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Weight	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	n^{-1}, $\sqrt{n}$, . . .	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5,000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	Digest	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1,458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1,165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	n^1,n^2, . . . ,n!	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	20,229

show that $\hat{p}_i$, the proportion of numbers that begin with the leading digit i, follows the theoretical leading digit law: $p_i = 100 \log_{10}(1 + 1/i)$. It turns out that in both columns 3 and 9, Benford incorrectly rounded toward the theoretical proportions p_i. For column 3, 12.26 was rounded to 12.4. For column 9, 4.775 was rounded to 4.7. The theoretical percentages are $p_3 \doteq 12.5$ and $p_9 \doteq 4.6$.

Changes in rounded proportions to make tables round to 100 percent can affect the results of statistical tests such as chi-square. The chi-squared statistic for goodness of fit of c sample proportions $\hat{p}_i$ based on a sample size of N to theoretical probability p_i is $\chi^2 = N \sum_{i=1}^{c} (\hat{p}_i - p_i)^2 / p_i$. If the $\hat{p}_i$ did not sum to 1, then adjusting the $\hat{p}_i$ that correspond to small p_i can change the value of χ^2 appreciably for large N. Of course, it becomes easier to change the value of χ^2 as the rounding number decreases.

For example, consider Benford's data in Table 5. The proportion of all 20,229 numbers that begin with a 1 can be found by taking a weighted average of the proportions in the first column. Doing this for each digit yields Table 6.

Ylvisaker (1977) gives χ^2 from Table 6 as 85. To show the effect of rounding, Table 7 gives the results of rounding the numbers in Table 6 to the nearest 1 percent. The χ^2 statistic for goodness of fit of data to theory is approximately 192. Both rows of Table 7 add to 101 percent.

If 1 percent is subtracted from the data row in the eighth position and 1 percent is subtracted from the theory row in the seventh position so that both rows sum to 100 percent, the χ^2 statistic becomes approximately 118. Thus, rounding to help the data fit the theory can make a difference. This example also shows that it is important to calculate with many-digit accuracy when computing χ^2 for large sample sizes.

6. Proportion of Benford Data Beginning With Digit Leading i and Theoretical Proportions $100 \log_{10}(1 + 1/i)$

	Digit								
	1	2	3	4	5	6	7	8	9
Data	28.9	19.5	12.7	9.1	7.5	6.4	5.4	5.5	5.0
Theory	30.1	17.6	12.3	9.7	7.9	6.7	5.8	5.1	4.6

7. Numbers in Table 6 Rounded to Nearest 1 Percent

	Digit								
	1	2	3	4	5	6	7	8	9
Data	29	20	13	9	8	6	5	6	5
Theory	30	18	12	10	8	7	6	5	5

[*Received January 1978. Revised December 1978.*]

REFERENCES

Benford, Frank (1938), "The Law of Anomalous Numbers," *Proceedings of the American Philosophical Society*, 78, 551–572.

Blackwell, David, and Dubins, Lester (1962), "Merging of Opinions With Increasing Information," *Annals of Mathematical Statistics*, 33, 882–886.

Diaconis, Persi (1977), "The Distribution of Leading Digits and Uniform Distribution Mod 1," *Annals of Probability*, 5, 72–81.

Dunford, Nelson, and Schwartz, Jacob (1957), *Linear Operators*, New York: Wiley Interscience.

Feller, William (1968), *An Introduction to Probability Theory and Its Applications*, Vol. I (3rd ed.), New York: John Wiley & Sons.

——— (1971), *An Introduction to Probability Theory and Its Applications*, Vol. II (2nd ed.), New York: John Wiley & Sons.

Hill, Bruce (1970), "Zipf's Law and Prior Distributions for the Composition of a Population," *Journal of the American Statistical Association*, 65, 1220–1232.

Meyer, Paul Andre (1966), *Probability and Potentials*, Mass.: Blaisdell.

Mosteller, Frederick, Youtz, Cleo, and Zahn, Douglas (1967), "The Distribution of Sums of Rounded Percentages," *Demography*, 4, 850–858.

Raimi, Ralph (1976), "The First Digit Problem," *American Mathematics Monthly*, 83, 521–538.

Wallis, Allen, and Roberts, Harley (1956), *Statistics: A New Approach*, Chicago: Free Press.

Ylvisaker, Donald (1977), "Test Resistance," *Journal of the American Statistical Association*, 72, 551–556.