Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Published as:

Jann, Ben (2008). Multinomial goodness-of-fit: large sample tests with survey design correction and exact tests for small samples. The Stata Journal 8(2): 147-169.

ETH Zurich Sociology Working Paper No. 2

Multinomial goodness-of-fit: large sample tests with survey design correction and exact tests for small samples*

Ben Jann

January 2008

ETH Zurich, Chair of Sociology

SEW E 21, Scheuchzerstrasse 70
8092 Zurich, Switzerland

Tel. +41 44 632 55 56
Fax +41 44 632 10 54
info@soz.gess.ethz.ch
www.socio.ethz.ch

*This research was supported by ETH Research Grant Number TH-8/04-3 on "Empirical studies in scientific misconduct: Prevalence and methods for detecting fraud and errors using the Benford distribution".

Multinomial goodness-of-fit: large sample tests with survey design correction and exact tests for small samples

Ben Jann
ETH Zurich
jann@soz.gess.ethz.ch

January 2008

Abstract. A new Stata command called `mgof` is introduced. The command is used to compute distributional tests for discrete (categorical, multinomial) variables. Apart from classic large sample χ^2-approximation tests based on Pearson's X^2, the likelihood ratio, or any other statistic from the power-divergence family (Cressie and Read 1984), large sample tests for complex survey designs and exact tests for small samples are supported. The complex survey correction is based on the approach by Rao and Scott (1981) and parallels the survey design correction used for independence tests in `svy: tabulate`. The exact tests are computed using Monte Carlo methods or exhaustive enumeration. An exact Kolmogorov-Smirnov test for discrete data is also provided.

Keywords: st0001, multinomial, goodness-of-fit, chi-squared, categorical data, exact tests, Monte Carlo, exhaustive enumeration, combinatorial algorithms, complex survey correction, power-divergence statistic, Kolmogorov-Smirnov, Benford's law

1 Introduction

A fundamental task in statistics is to test whether an observed distribution differs from a theoretical null distribution, yet support for such tests is incomplete in the standard release of Stata. There are tools to test continuous distributions such as normality tests ([R] **sktest**, [R] **swilk**) and a one-sample Kolmogorov-Smirnov test ([R] **ksmirnov**). However, distributional tests for discrete variables are missing.

This lack of tests for discrete variables might not seem to be an issue of serious concern since the classic multinomial goodness-of-fit test is a simple χ^2 test based on Pearson's X^2, which is easy to compute from the values of a frequency table ([R] **tabulate**). Furthermore, several user implementations are available for this test (e.g. Weesie 1997). However, the classic test is only valid in simple random samples and cannot be used with sampling weights or other complex survey features. Furthermore, the classic test is only asymptotic and may be biased in small samples or when the null distribution is very uneven.

I therefore present a new command called `mgof` that performs goodness-of-fit tests

for discrete variables. The command supports complex survey designs and also offers exact tests for small samples. The complex survey tests are based on the approach by Rao and Scott (1981) and parallel the survey design correction used for independence tests in official `tabulate` ([SVY] **svy: tabulate twoway**). The exact tests are computed by sampling from the null distribution (Monte Carlo method) or by enumerating all possible data configurations (exhaustive enumeration technique). Supported test statistics include Pearson's X^2, the likelihood ratio, and any other statistic from the Cressie and Read (1984) family. Further offered tests are the exact multinomial probability test and the exact discrete Kolmogorov-Smirnov test.

2 Syntax and options

The default for `mgof` is to perform classical large sample χ^2 approximation tests, optionally with survey design correction. Alternatively, `mgof` computes exact tests using Monte Carlo methods (`mc` option) or exhaustive enumeration (`ee` option). The syntax is:

`mgof` *varname* [`=`*exp*] [*weight*] [*if*] [*in*] [`,` *options*]

`mgofi` *#1 #2* `...[` `/` *p1 p2* `...]` [`,` *options*]

options	description
Method 1	
`approx[(#)]`	compute large sample χ^2 tests; the default
`svy[(`*svyspec*`)]`	adjust tests for survey design
`vce(`*vcetype*`)`	adjust tests using `proportion` variance estimate
`cluster(`*varname*`)`	adjust tests for intragroup correlation
`noisily`	show output from `proportion`
Method 2	
`mc`	compute Monte Carlo exact tests
`reps(#)`	number of replications for `mc`; default is `reps(10000)`
`level(#)`	set confidence level for `mc`; default is `level(99)`
`citype(`*citype*`)`	set confidence interval type for `mc`; default is `citype(exact)`
Method 3	
`ee`	compute exhaustive enumeration exact tests
Test statistics	
`nox2`	suppress Pearson's X^2 statistic
`nolr`	suppress log likelihood ratio statistic
`cr[(#)]`	include Cressie-Read statistic; # defaults to 2/3
`mlnp`	include log outcome probability statistic (`mc` and `ee` only)
`ksmirnov`	include Kolmogorov-Smirnov statistic (`mc` and `ee` only)

(*Continued on next page*)

Other options

`freq`	display frequency table
`percent`	display frequency table in percent
`matrix(`*name*`)`	provide matrix containing observed and expected counts
`expected(`*name*`)`	provide matrix (column vector) containing expected counts
`nodots`	suppress progress dots (`mc` and `ee` only)

`by` is allowed (unless `svy` is specified); see [D] **by**. `fweight`s, `pweight`s, and `iweight`s are allowed; see [U] **11.1.6 weight**; restrictions: `pweight`s are not allowed with `ee` or `mc`; `iweight`s are not allowed with `ee`, and not with `mc` if the `mlnp` option is specified.

The (theoretical) null distribution (the distribution against which *varname* is tested) is specified by *exp*. *exp* is assumed to evaluate to the hypothesized probabilities of the categories of *varname* or to quantities proportional to these probabilities (e.g. expected counts; the scale does not matter). If *exp* is omitted, the uniform (geometric, equiprobable) distribution is used as the theoretical distribution.

`mgofi` is the immediate form of `mgof` ([U] **19 Immediate commands**) where *#1*, *#2*, etc. specify the observed counts and, optionally, *p1*, *p2*, etc. specify the theoretical probabilities or expected counts.

Method 1 options

`approx`[`(#)`], the default method, computes classical large sample χ^2 approximation tests based on Pearson's X^2 and the log likelihood ratio statistic (see, e.g., Horn 1977, Cressie and Read 1989, Sokal and Rohlf 1995, Ch. 17). The degrees of freedom for χ^2 tests are determined as $k-\#-1$ where k is the number of categories and $\#$, provided by the user, indicates the number of fitted parameters (imposed restrictions) ($\#$'s default is 0). If `pweight`s are specified, the tests are corrected as outlined in Section 4.

`svy`[`(`[*vcetype*] [`,` *svy_options*]`)`] specifies that the test results be adjusted for survey design effects according to the `svyset` specifications (see [SVY] **svyset**). *vcetype* and *svy_options* are as described in [SVY] **svy**. The correction procedure is described in Section 4. The `svy` option is not allowed with `mgofi`.

`vce(`*vcetype*`)` specifies that the variance-covariance matrix of the proportions be estimated using the `proportion` command (see [R] **proportion**) and that the tests be adjusted based on this estimate (see Section 4 below). *vcetype* may be `analytic`, `cluster` *clustvar*, `bootstrap`, or `jackknife` (plus possible suboptions as described in [R] ***vce_option***). `analytic` and `cluster` *clustvar* are not allowed with Stata 9. The `vce()` option is not allowed with `mgofi`.

`cluster(`*clustvar*`)` is Stata 9 syntax for `vce(cluster` *clustvar*`)`. The `cluster()` option is not allowed with `mgofi`.

`noisily` displays the output from the `proportion` command, which is used to estimate the variances of the proportions if `svy`, `vce()`, or `cluster()` is specified or if `pweights` are applied.

Method 2 options

`mc` causes the exact p-values to be approximated by sampling from the null distribution (Monte Carlo simulation). The default number of replications for the simulation is 10,000; see the `reps()` option (the same set of samples is used for all test statistics). 99% confidence intervals are displayed for the estimated p-values.

`reps(#)` sets the number of replications for the `mc` method. The default is 10,000.

`level(#)` sets the level for the confidence intervals of the p-values computed by the `mc` method. The default is `level(99)`. Note that, unlike many other Stata commands, `mgof` does *not* depend on `set level` (see [R] **level**).

`citype(`*type*`)` specifies how the binomial confidence intervals for the p-values from the `mc` method are to be calculated. Available types are `exact`, `wald`, `wilson`, `agresti`, and `jeffreys`. See [R] **ci**. `citype(exact)` is the default.

Method 3 option

`ee` causes the exact p-values to be computed by cycling through all possible data compositions given the sample size and the number of categories. Since the number of compositions grows very fast—it is equal to $(n+k-1)!/((k-1)!n!)$ where n is the sample size and k is the number of categories—the `ee` method is only feasible for very small samples and few categories. An important exception is when the null distribution is uniform (and `ksmirnov` is not specified). In this case the tests are based on enumerating partitions, which are much fewer in number than compositions.

Test statistics options

`nox2` suppresses Person's X^2 statistic.

`nolr` suppresses the likelihood ratio statistic.

`cr`$\big[$`(#)`$\big]$ specifies that the Cressie-Read statistic with parameter $\lambda = \#$ be included (Cressie and Read 1984; also see Weesie 1997). The default for $\#$ is `2/3`.

`mlnp` requests that a test based on the (minus log) multinomial probability of the observed outcome be included (see Horn 1977). `mlnp` is not allowed with the `approx` method.

`ksmirnov` requests that the two-sided Kolmogorov-Smirnov statistic be included. The Kolmogorov-Smirnov statistic is sensitive to the order of the categories and should only be used with variables that have a natural order (i.e. ordinal or discrete metric data). Note that the Kolmogorov-Smirnov test implemented in official Stata's `ksmirnov` is conservative in the case of discrete data (see, e.g., Conover 1972). The methods implemented here are exact. `ksmirnov` is not allowed with `approx`.

Other options

`freq` displays a table containing observed and expected frequencies.

`percent` displays a table containing observed and expected percent.

`matrix(`*name*`)` specifies that the observed and expected counts are to be taken from matrix *name* (see [P] **matrix**). The first column of the matrix provides the observed counts and the second column, if present, provides the expected counts or theoretical probabilities. The uniform distribution is used if the matrix does not contain a second column. Do not provide non-integer observed counts with the `ee` or `mc` methods. The `matrix()` option is not allowed with `mgofi`.

`expected(`*name*`)` specifies that the expected counts or theoretical probabilities are to be taken from column vector *name* (see [P] **matrix**). `mgof` aborts if the number of elements in the vector does not match the number outcomes.

`nodots` causes the progress dots for the `ee` and `mc` methods to be suppressed. The default is to display a dot for each 2 percent of completed computations.

Returned results

Scalars

`r(N)`	number of observations	`r(N_pop)`	population size
`r(N_strata)`	number of strata	`r(N_psu)`	number of PSUs
`r(N_clust)`	number of clusters	`r(df_r)`	design degrees of freedom
`r(df)`	degrees of freedom for χ^2	`r(df1)`	numerator d.f. for F
`r(df2)`	denominator d.f. for F	`r(delta)`	mean generalized design effect
`r(a2)`	squared variation coefficient of generalized design effects	`r(reps)`	number of replications
`r(partitions)`	number of partitions	`r(compositions)`	number of compositions
`r(`*stat*`)`	value of test statistic	`r(F_`*stat*`)`	design corrected F
`r(p_`*stat*`)`	(design corrected) p-value	`r(p_`*stat*`_srs)`	uncorrected p-value
`r(p_`*stat*`_lb)`	lower C.I. bound for p-value	`r(p_`*stat*`_ub)`	upper C.I. bound for p-value

where *stat* is `x2` (Pearson's X^2), `lr` (log likelihood ratio), `cr` (Cressie-Read statistic), `mlnp` (minus log outcome probability), or `ksmirnov` (Kolmogorov-Smirnov D)

Macros

`r(depvar)`	name of tabulated variable	`r(h0)`	definition of the theoretical distribution
`r(method)`	test method	`r(stats)`	list of test statistics
`r(lambda)`	Cressie-Read λ	`r(citype)`	Monte Carlo C.I. type
`r(cilevel)`	Monte Carlo confidence level		

Matrix

`r(count)`	observed and expected counts

3 Classic large sample tests

The classic large sample goodness-of-fit tests for discrete data are based on the result that statistics such as Pearson's

$$X^2 = \sum_{j=1}^{k} \frac{(f_j - h_j)^2}{h_j}$$

where f_j and h_j are the observed and expected (theoretical) counts for the categories $j = 1, \ldots, k$, the likelihood ratio statistic

$$G^2 = 2 \sum_{j=1}^{k} f_j \ln\left(\frac{f_j}{h_j}\right)$$

or, generally, the power-divergence statistic (Cressie and Read 1984)

$$D^2(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{k} f_j \left[\left(\frac{f_j}{h_j}\right)^{\lambda} - 1\right]$$

are asymptotically $\chi^2(k-1)$ distributed.[1] Significance level α given, an observed distribution is considered significantly different from the null distribution if the chosen test statistic exceeds the $(1-\alpha)$ quantile of the χ^2 distribution with $(k-1)$ degrees of freedom. For guidelines on choosing a test statistic see the "Which Test Statistic?" sections in Read and Cressie (1988). One result, for example, is that in small samples the approximation of the χ^2-distribution is much better for Pearson's X^2 than for the likelihood ratio. Based on various simulations, Read and Cressie (1988) propose the $\lambda = 2/3$ power-divergence statistic as a good compromise in a wide range of situations. $\lambda = 2/3$ is the default for $D^2(\lambda)$ in `mgof`.

As an example, assume testing a sequence of numbers against Benford's law. The law states that under certain conditions the first digit of numbers in the base 10 system follows a probability distribution given as

$$\Pr(d) = \log_{10}(1 + 1/d), \quad d \in \{1, \ldots, 9\}$$

(Newcomb 1881, Benford 1938, Hill 1998). In a small Swiss mail survey, respondents were asked to indicate the first digit of the street number of an acquaintance.[2] The distribution of the indicated digits is astonishingly close to Benford's law, as the results from `mgof` reveal:

1. Note that the Cressie-Read statistic is equal to Pearson's X^2 if $\lambda = 1$ and, as the limiting value, to the likelihood ratio if $\lambda = 0$. Further special cases are the Freeman-Tukey statistic with $\lambda = -1/2$, the Kullback-Leibler information with $\lambda = -1$, and Neyman's modified X^2 statistic with $\lambda = -2$ (see Cressie and Read 1984, Weesie 1997).

2. The survey was conducted in December 2006 and January 2007 by the Sociology Department of the ETH Zurich. Respondents were sampled from the residents of the German speaking part of Switzerland. The street number question was included towards the end of a very short questionnaire on income inequality. It had no relation to the other questions. The response rate of the survey was 41%.

```
. use digits, clear
(2007 Swiss Street Number Data)
. mgof firstdigit = log10(1+1/firstdigit), cr percent
                      Number of obs =     313
                      N of outcomes =       9
                      Chi2 df       =       8
```

Goodness-of-fit	Coef.	P-value
Pearson's X2	6.226606	0.6219
Log likelihood ratio	6.475677	0.5941
Cressie-Read (2/3)	6.303507	0.6133

firstdigit	observed	expected
1	32.59	30.10
2	17.57	17.61
3	14.70	12.49
4	10.86	9.69
5	6.39	7.92
6	6.07	6.69
7	4.47	5.80
8	4.15	5.12
9	3.19	4.58
Total	100.00	100.00

The p-values of the tests based on Pearson's X^2, the Cressie-Read statistic, and the likelihood ratio statistic suggest that the null hypothesis of Benford distributed digits cannot be rejected.

4 Survey design correction for large sample tests

The results of the standard χ^2 tests are only valid for simple random samples. In the case of non-identical sampling probabilities, non-independence, or stratification, the tests may be considerably biased. `mgof` therefore offers a survey design correction which is based on Rao and Scott (1981) and is analogous to the default independence test correction used in `svy: tabulate twoway` (see [SVY] **svy: tabulate twoway** and the references therein). The procedure determines the "design effects" for the variances of the proportion estimates for the single outcomes and then corrects the χ^2 test statistic for the level and variation of these design effects. Rao and Scott (1981) call this a second order correction; a first order correction would ignore the variation in design effects. Finally, the corrected statistic is converted into an F statistic to adjust for the degrees of freedom of the employed variance estimates.

More precisely, let $\widehat{V}/(n-1)$ be a consistent estimate of the variance-covariance matrix of the proportion estimates $\hat{p}_i$, $i = 1, \ldots, k$, where n is the number of observations. Furthermore, let $\hat{v}_{ij}$ denote an element of $\widehat{V}$, m be the number of PSUs or clusters, and

L be the number of strata. The correction then assumes

$$F = \frac{\chi^2}{\hat{\delta}_{\cdot}(\hat{a}^2+1))} \cdot d^{-1} = \frac{\chi^2}{\hat{\delta}_{\cdot}} \cdot (k-1)^{-1}$$

to be $F(d,(m-L)d)$ distributed, where χ^2 stands for X^2, G^2, or $D^2(\lambda)$, and where

$$\hat{\delta}_{\cdot} = (k-1)^{-1}\sum_{i=1}^{k} \hat{v}_{ii}/\hat{p}_i, \qquad \hat{a}^2 = \left[(k-1)^{-1}\sum_{i=1}^{k}\sum_{j=1}^{k} \hat{v}_{ij}^2/(\hat{p}_i\hat{p}_j)\right] \cdot (\hat{\delta}_{\cdot}^2 - 1)^{-1},$$

and $d = (k-1)/(1+\hat{a}^2)$. $\hat{\delta}_{\cdot}$ is the the mean and $\hat{a}^2$ the squared variation coefficient of the "generalized design effects" for the proportions (Rao and Scott 1981). Official `proportion` is used to estimate $\widehat{V}/(n-1)$ taking into account `pweights`, clusters, or other complex survey design settings (see [R] **proportion** for details).

Sribney (1998; also see Thomas et al. 1996) provides simulation evidence indicating that the F-based variant of the second order Rao-Scott correction works well for independence tests in two-way contingency tables. Although it appears reasonable to assume that these results translate to goodness-of-fit tests in one-way tables, it seems important to perform at least a few brief checks. I therefore ran the following simulations.[3]

Simulation 1

Simulation 1 parallels the simulation reported by Sribney (1998). In each replication, a sample was initialized by drawing a number of cluster sizes from a uniform distribution and expanding the clusters to individual observations. Then two sets of variables with categorical values $d \in \{1,\ldots,k\}$ and k varying between 2 and 9 were generated from an underlying continuous variable y, which was $N(0,1)$-distributed and had an intra-class (cluster) correlation of 0.25 (see Sribney 1998 for details on how to generate such a variable). For the first set of categorical variables, y was categorized using normal quantiles with equally spaced probabilities, so that the variables had a geometric (uniform) distribution. For the second set, the cut points were chosen according to Benford's first digit law (see references above) where the base of the number system was set to $k+1$. Hence, the probabilities of the categories of the variables in the second set were given as

$$\Pr(d) = \log(1+1/d)/\log(k+1), \quad d \in \{1,\ldots,k\}$$

For a variable with two categories, for example, the probabilities were $(0.631, 0.369)$; for a variable with nine categories they were $(0.301, 0.176, \ldots, 0.046)$.

As in Sribney (1998), two types of simulations were conducted, one with small variance degrees of freedom (few PSUs) and one with large variance degrees of freedom (many PSUs). For the simulations with few PSUs, 20 clusters were generated with sizes between 30 and 70 observations, resulting in a sample size of approximately 1000 observations. For the simulations with many PSUs, 200 clusters were generated with

3. See Thomas and Rao (1987) and Rai et al. (2001) for additional results.

sizes between 3 and 10 observations, resulting in a sample size of approximately 1200 observations.[4] Ten-thousand replications were computed for both types and the nominal significance level was set to $\alpha = 0.05$. Note that in comparison to the simulation by Sribney (1998), the clusters were re-generated in each replication.

The results of the simulations are depicted in Figures 1 and 2. Tests based on the first and second order Rao-Scott corrections (RS1 and RS2, respectively) according to the following definitions are evaluated:

$$\begin{aligned}
X^2_{\text{RS1}} &= X^2/\hat{\delta}_{\cdot} && \overset{a}{\sim} && \chi^2(k-1) \\
F_{\text{RS1}} &= X^2_{\text{RS1}}/(k-1) && \overset{a}{\sim} && F(k-1,(m-L)(k-1)) \\
X^2_{\text{RS2}} &= X^2_{\text{RS1}}/(1+\hat{a}^2) && \overset{a}{\sim} && \chi^2(d) \\
F_{\text{RS2}} &= X^2_{\text{RS2}}/d && \overset{a}{\sim} && F(d,(m-L)d)
\end{aligned}$$

F_{RS2} corresponds to the default correction method outlined above. Furthermore, an adjusted Wald F test is considered with

$$F^W_{\text{adj}} = W\,\frac{(m-L)-(k-1)+1}{(m-L)(k-1)} \quad \overset{a}{\sim} \quad F(k-1,(m-L)-(k-1)+1)$$

where

$$W = (\hat{p}^* - p^*)^T(\widehat{V}^*/(n-1))^{-1}(\hat{p}^* - p^*)$$

and where $\hat{p} = (\hat{p}_1,\ldots,\hat{p}_k)$ and $p = (p_1,\ldots,p_k)$ are the vectors of observed and expected probabilities and the asterisk indicates that one of the categories is left out (see, e.g., Thomas and Rao 1987).[5]

In the case of small variance degrees of freedom (Figure 1), both the first order Rao-Scott corrected tests (RS1) and the adjusted Wald F test are highly anti-conservative (i.e. rejecting the null hypothesis to often) as the number of categories increases. The second order Rao-Scott corrected X^2 test is also anti-conservative, but the bias does not depend on the number of categories. For the second order Rao-Scott corrected F test, however, the simulated rejection rates match the nominal 5 percent very well. In the case of large variance degrees of freedom (Figure 2), the first order Rao and Scott corrected tests and the adjusted Wald F test are still slightly anti-conservative. The second order Rao-Scott correction again performs well, at least for the uniform variables. With the Benford distributed variables, the second order Rao-Scott correction seems to be slightly conservative.[6]

4. The number of 1300 given in Sribney (1998) seems to be incorrect.

5. Rao-Scott corrected tests based on the likelihood ratio statistic were also conducted. Results were similar to the tests based on Pearson's X^2.

6. Note that in both simulations, the uncorrected tests were highly anti-conservative (51–57% and 11–16% rejection, respectively).

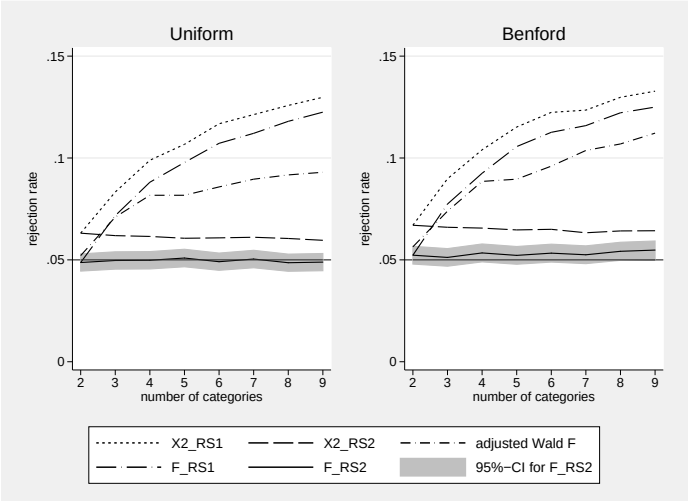

Figure 1: Rejection rates in the small variance degrees of freedom simulations

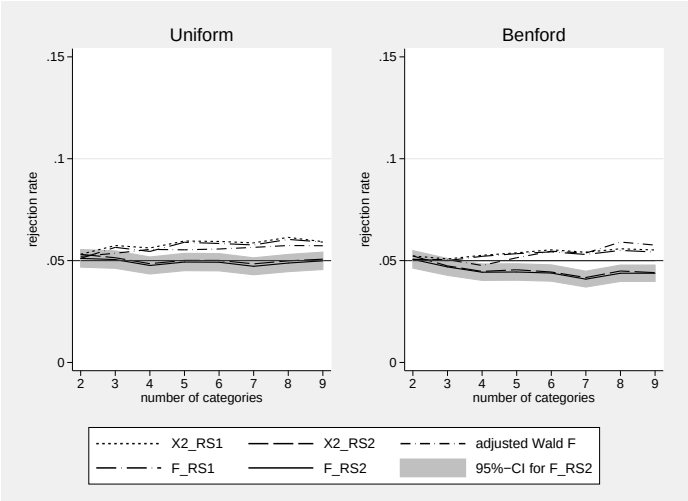

Figure 2: Rejection rates in the large variance degrees of freedom simulations

Simulation 2

Simulation 1 is model-based and somewhat artificial. To get an impression of the performance of the Rao-Scott correction in a more realistic setting, I conducted a second simulation based on sampling from a "real" population, the Swiss 2000 Census.[7]

The sampling plan for the simulation was as follows. First, a total of 100 municipalities was sampled proportional to size (with replacement) from 15 strata (the Swiss cantons, with some aggregations). Second, 10 households were drawn with replacement from each sampled municipality and from each sampled household one individual was selected. The *a priori* sampling probabilities for the individuals varied depending on household size and due to moderate oversampling of small strata. The `gsample` user command was used to draw the samples (Jann 2006).

A number of categorical variables were tested with the true population distribution as the null hypothesis. The variables included sex (1 = male, 2 = female), nationality (Swiss, foreign), marital status (single, married, divorced, widowed), education (6 levels), and socio-economic status (12 levels). The distributions, representing a broad mix of patterns, are listed in Table 1.

Table 1: Population distributions (in percent)

	1	2	3	4	5	6	7	8	9	10	11	12
Sex	48.5	51.5										
Nationality	80.2	19.8										
Marital status	30.1	56.3	6.9	6.8								
Education	13.0	25.6	36.7	7.9	9.1	7.6						
Socio-economic status	2.1	5.6	5.1	10.1	12.0	4.9	7.4	15.5	7.8	14.2	10.8	4.5

The rejection rates from 5000 replications with a nominal $\alpha = 0.05$ significance level are displayed in Table 2. The test statistics are the same as in Simulation 1, namely, first and second order Rao-Scott corrections of Pearson's X^2, with and without F-conversion, and the adjusted Wald F.[8] In addition, rejection rates for Pearson's X^2 without correction are reported (X^2_{SRS}). As is evident from the rejection rates in the last column in Table 2 (up to 40%), the uncorrected tests yield intolerably anti-conservative results in such a sample. Among the considered correction procedures, the second order Rao-Scott F performed best. The rejections rates for F_{RS2} closely match the nominal 5% for sex, education, and socio-economic status, but are somewhat anti-conservative for nationality and marital status, possibly due to the strongly skewed distributions of

7. The Swiss Census was conducted by the Swiss Federal Statistical Office and covers all residents of Switzerland in December 2000. For the purpose of the simulation, I restricted the population to individuals of age 15 or older. The population then consisted of 6,043,350 individuals, 3,179,246 housholds, and 2,896 municipalities. The strata sizes varied between 200 thousand and one million individuals.

8. Results for the Rao-Scott corrections based on the likelihood ratio statistic were similar.

the two variables. The degree of anti-conservatism, however, is not dramatic, and is smallest among all evaluated procedures. As in Simulation 1, the adjusted Wald F is considerably anti-conservative in some cases.

Table 2: Rejection rates (in percent; 5000 replications)

	X^2_{RS1}	F_{RS1}	X^2_{RS2}	F_{RS2}	F^W_{adj}	X^2_{SRS}
Sex	4.9	4.6	4.9	4.6	4.7	16.6
Nationality	7.3	6.9	7.3	6.9	8.3	23.3
Marital status	9.3	9.0	7.1	6.7	13.8	29.1
Education	7.1	6.8	5.0	4.8	8.5	32.8
Socio-economic status	10.7	10.4	5.8	5.6	12.0	41.2

Overall, the results from Simulations 1 and 2 suggest that the second order Rao-Scott corrected F test outperforms the other considered tests and is a good default choice.

Example

In the example in Section 3 the digit distribution of house numbers of acquaintances was analyzed. The survey from which the data was taken is based on a simple random sample of households. Because only one adult was selected per household, the individual level sampling probabilities depended on the household size: individuals from larger households had smaller sampling probabilities. Therefore one should apply probability weights (inverse to the number of adult household members) to make the data representative of the individual population instead of the population of households. Since the standard tests would not be valid, `mgof` applies the complex survey design correction if `pweights` are specified. In addition to the uncorrected test statistics, corrected F statistics and associated p-values are reported. The results from `mgof` applied to the weighted example data are as follows:

```
. use digits, clear
(2007 Swiss Street Number Data)
. mgof firstdigit = log10(1+1/firstdigit) [pw=w]
                                Number of obs =     313
                                N of outcomes =       9
                                F df1         = 7.93123
                                F df2         = 2474.54
```

Goodness-of-fit	Coef.	F-value	P-value
Pearson's X2	6.144661	0.6435	0.7402
Log likelihood ratio	6.425913	0.6730	0.7145

Again there is virtually no evidence to reject the null hypothesis that the data are distributed according to Benford's law.

A more general syntax to obtain results with design correction is to set the survey properties using `svyset` (see [SVY] **svyset**) and then apply the `svy` option in `mgof`, as in the following example:[9]

```
. svyset [pw=w]
      pweight: w
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
. mgof firstdigit = log10(1+1/firstdigit), svy
Number of strata =       1        Number of obs =     313
Number of PSUs   =     313        Pop size      =     583
                                  Design df     =     312
                                  N of outcomes =       9
                                  F df1         = 7.93123
                                  F df2         = 2474.54
```

Goodness-of-fit	Coef.	F-value	P-value
Pearson's X2	6.144661	0.6435	0.7402
Log likelihood ratio	6.425913	0.6730	0.7145

5 Exact small sample tests

The χ^2 distribution of statistics such as Pearson's X^2 is only asymptotic and the p-values of the standard χ^2 goodness-of-fit tests may be biased when the sample is very small or the null distribution is highly uneven. In such cases it is desirable to compute the exact p-values.

Exhaustive enumeration

The most straightforward approach to compute an exact goodness-of-fit test is to construct all possible data combinations given the number of observations and the number of categories and sum up the probabilities of all configurations that are at least as distant from the null hypothesis as the observed data (e.g. Radlow and Alf 1975). The steps of the procedure may be summarized as follows:

1. Calculate the probability of each data configuration, given the null hypothesis. Under simple random sampling, the probability of a specific configuration $f = (f_1, \ldots, f_k)$ given expected cell probabilities $p = (p_1, \ldots, p_k)$ (the null hypothesis)

9. Due to some technical difficulties, the `svy` prefix command (see [U] **11.1.10 Prefix commands**) cannot be used with `mgof`.

is

$$\Pr(f|p) = \frac{n!}{(f_1!f_2!\cdots f_k!)} \cdot p_1^{f_1} p_2^{f_2} \cdots p_k^{f_k}$$

where $n = \sum f$.

2. Calculate the value of the test statistic, say Pearson's X^2 or the likelihood ratio statistic G^2, for each data configuration, given the null hypothesis.

3. Compute the exact p-value as the sum of the probabilities of all configurations in which the test statistic is at least as large as in the observed data.

A natural variant to using Pearson's X^2 or the likelihood ratio as the test criterion is to directly use $\Pr(f|p)$ to determine whether a configuration adds to the p-value or not (see e.g. Horn 1977). `mgof` offers such a multinomial probability test via the `mlnp` option, where the reported outcome probability statistic is parameterized as $-\ln(\Pr(f|p))$ so that large values indicate departure from the null distribution and the scaling is similar to a χ^2 statistic. Little evidence exists for whether directly using the multinomial probability as the fit statistic is superior to using a statistic such as Pearson's X^2, but both procedures yield "exact" p-values. The only difference is in how discrepancies between the null hypothesis and the data are valued. While statistics such as Pearson's X^2 are defined in terms of differences (or ratios) between expected and observed counts, the direct approach uses the multinomial probability as the measure of discrepancy. In general, different fit statistics give weight to different patterns of deviations from the null and the usefulness of a specific statistic may depend on situation (Read and Cressie 1988, 136–137).

The method outlined above—whichever fit statistic is employed—is called the "exhaustive enumeration" method since all possible data configurations are enumerated. It can be implemented using algorithms to construct k-part compositions of n (Reingold et al. 1977; Nijenguis and Wilf 1978). For example, for $n = 3$ and $k = 2$ the possible compositions are:[10]

```
. mata
------------------------------------------------- mata (type end to exit) -----
: f = mm_compositions(3,2)
: f
       1   2   3   4
    +-------------------+
  1 |  3   2   1   0    |
  2 |  0   1   2   3    |
    +-------------------+
: end
--------------------------------------------------------------------------------
```

Using the sex distribution in a sample with three individuals as an example, possible sample compositions are: 3 females and 0 males, 2 females and 1 male, 1 female and

10. The `mm_compositions()` Mata function is part of the `moremata` package, which provides a number of combinatorial algorithms (among many other functions) (Jann 2005).

2 males, or 0 females and 3 males. Assume that the second outcome, 2 females and 1 male, is the observed outcome. To compute the p-value of the test that the ratio of females to males is, say, 1 to 2, one would determine for each composition Pearson's X^2 and the probability of the composition given the null hypothesis, and then sum up the probabilities of all compositions for which X^2 is at least as large as the observed X^2. For example:[11]

```
. mata
------------------------------------------------- mata (type end to exit) -----
: h = (1, 2)'
: x2 = colsum((f :- h):^2 :/ h)
: x2
        1     2     3     4
  1 |     6   1.5     0   1.5 |
: n = 3
: pr = exp(lnfactorial(n) :- colsum(lnfactorial(f)) :+ colsum(f :* ln(h :/ n)))
: pr
                 1             2             3             4
  1 |  .037037037   .2222222222   .4444444444   .2962962963 |
: pvalue = sum(pr :* (x2 :>= x2[2]))
: pvalue
  .5555555556
: end
--------------------------------------------------------------------------------
```

The p-value is 0.556 and, hence, we cannot reject the null hypothesis (which is not surprising given only three observations). The same result can be obtained by `mgofi` as follows:

```
. mgofi 2 1 / 1 2, ee nodots
                      Number of obs =       3
                      N of outcomes =       2
                      Compositions  =       4
------------------------------------------------------------
                                             Exact
      Goodness-of-fit |       Coef.    P-value
----------------------+-------------------------------------
        Pearson's X2  |         1.5     0.5556
 Log likelihood ratio |    1.386294     0.5556
------------------------------------------------------------
```

The number of possible data configurations increases rapidly with additional observations and categories (the combinatorial explosion), which imposes restrictions for the application of the exhaustive enumeration method. The formula for the number of

11. The logarithmic form of $\Pr(f|p)$ is used here to turn products into sums and thus make computations easier. Also note that the logarithmic form is computationally more robust since `lnfactorial()` can be evaluated for much larger numbers than `factorial()`.

k-fold compositions of n is $(n+k-1)!/((k-1)!n!)$. For example, with $n = 50$ and $k = 5$ the problem size is 316,251, which can be handled (`mgof` takes about about 7 seconds on my computer for a problem of this size). If k is increased to 6, the number of compositions is 3,478,761, taking 80 second to compute. With $k = 7$ the number further increases to 32,468,436 and with $k = 10$ it is 12,565,671,261, taking three to four days.

The examples suggest that the exhaustive enumeration method is only feasible for very small problems. Note, however, that in the case of a uniform null distribution the amount of computations can be reduced a great deal due to redundancies among the compositions. If all elements of h (and hence of p) are equal, then the order of the elements in a composition does not affect the value of $\Pr(f|p)$ or X^2. In the toy example above, for instance, the compositions $\{2,1\}$ and $\{1,2\}$ would yield identical values for $\Pr(f|p)$ and X^2 if p is uniform, as would $\{3,0\}$ and $\{0,3\}$. Hence, the exact p-values can be computed by enumerating only compositions with unique sets of elements. These compositions are equivalent to the (zero-padded) integer partitions of n into k or fewer addends (Andrews 1984; Andrews and Eriksson 2004). For example, in the case of $n = 4$ and $k = 3$ possible partitions are:

```
. mata
------------------------------------------------- mata (type end to exit) -----
: f = mm_partitions(4,3)
: f
       1   2   3   4
    +-----------------+
  1 |  4   3   2   2  |
  2 |  0   1   2   1  |
  3 |  0   0   0   1  |
    +-----------------+

: end
--------------------------------------------------------------------------------
```

As is immediately clear, the number of partitions grows much slower with n and k than the number of compositions. For example, for $n = 50$ and $k = 10$, as above, the number of partitions is only 62,740, compared to 12.6 billion compositions. In the case of a uniform null distribution it is therefore much more efficient to compute the p-value based on partitions, where each partition is weighted by the number of possible permutations (see, e.g., Hirji 1997).[12]

In the example in Section 3 there were 313 observations and 9 categories. This is too large a problem for the exhaustive enumeration method ($2.56 \cdot 10^{15}$ compositions or $1.09 \cdot 10^{10}$ partitions, respectively). For purpose of illustration we apply the exhaustive enumeration tests using a 5% sample of the data:[13]

12. The number of permutations equals $k!/(d_1! \cdots d_I!)$, where d_i denotes the number of repetitions of the ith distinct addend in the partition. The algorithm used by `mgof` to enumerate the partitions is based on Algorithm ZS1 by Zohgbi and Stojmenovic (1998) with some modifications to generate restricted partitions.

13. The `gsample` command is provided by Jann (2006).

```
. use digits, clear
(2007 Swiss Street Number Data)
. gsample 5, percent generate(freq)
. mgof firstdigit = log10(1+1/firstdigit) [fw=freq], ee
Percent completed (735471 compositions)
0 ----- 20 ------- 40 ------- 60 ------- 80 ------ 100
..................................................

                      Number of obs =      16
                      N of outcomes =       9
                      Compositions  =  735471

------------------------------------------------------
                                             Exact
      Goodness-of-fit        Coef.       P-value
------------------------------------------------------
         Pearson's X2      7.902401        0.4207
 Log likelihood ratio      9.157266        0.4564
------------------------------------------------------

. mgof firstdigit [fw=freq], ee
Percent completed (201 partitions)
0 ----- 20 ------- 40 ------- 60 ------- 80 ------ 100
..................................................

                      Number of obs =      16
                      N of outcomes =       9
                      Partitions    =     201

------------------------------------------------------
                                             Exact
      Goodness-of-fit        Coef.       P-value
------------------------------------------------------
         Pearson's X2         15.5        0.0557
 Log likelihood ratio     18.13734        0.0348
------------------------------------------------------
```

The results indicate that the data in the sample are closer to Benford's law than they are to the uniform distribution. Note that, especially for the likelihood ratio, the exact p-values can be considerably different from the approximate p-values in such small samples. The approximate p-values are 0.443 and 0.329 for the tests against Benford's law and 0.050 and 0.020 for the equal probability case.

❑ Technical note

In the example above, the sample was marked using frequency weights instead of constructing a dataset containing the sampled observations. The rationale behind this was to ensure that all original categories remained visible to `mgof` even if some of them were missing in the sample. In general, a goodness-of-fit test will be biased if categories with theoretical probabilities greater zero are omitted from the test due to lack of corresponding observations. Such "missing" categories can be introduced to `mgof` by adding extra observations to the dataset and assigning zero weights to them. Alternatively, zero observed counts can be specified using the immediate `mgofi` or the `matrix()` option.

Monte Carlo method

Many problems are too large for exhaustive enumeration but one might still want to compute the exact p-values. An approach which is easy to implement and can handle larger problems in reasonable time is to approximate the exact p-values by Monte Carlo simulation. The approach is made by sampling from the null distribution[14] and computing the p-value as the fraction of replications in which the test statistic is at least as large as in the observed data. A drawback of the procedure is that the computed p-values are subject to random variation (which, however, can be made arbitrarily small by increasing the number of replications). Ninety-nine percent confidence intervals (computed using `cii`; see [R] **ci**) are therefore reported by `mgof`.

Using the Benford example from Section 3, the Monte Carlo method can simulate the exact p-values in a few seconds:

```
. use digits, clear
(2007 Swiss Street Number Data)
. mgof firstdigit = log10(1+1/firstdigit), mc
Percent completed (10000 replications)
0 ——— 20 ——— 40 ——— 60 ——— 80 ——— 100
..................................................
                                          Number of obs =     313
                                          N of outcomes =       9
                                          Replications  =   10000
```

Goodness-of-fit	Coef.	Exact P-value	[99% Conf.	Interval]
Pearson's X2	6.226606	0.6147	0.6021	0.6272
Log likelihood ratio	6.475677	0.5921	0.5793	0.6048

As can be expected for a sample of this size, the approximate p-values (0.6219 and 0.5941; see Section 3) are very close to the simulated exact p-values and lie within the computed 99% confidence limits.

Discrete Kolmogorov-Smirnov test

In addition to the multinomial tests based on, say, Pearson's X^2 or the likelihood ratio, `mgof` also offers an exact (two-sided) Kolmogorov-Smirnov test for ordered discrete data. The Kolmogorov-Smirnov test has higher power than the multinomial tests, but is only appropriate if the categories have a natural order.

The (two-sided) Kolmogorov-Smirnov test statistic is defined as the supremum (least upper bound) of the absolute difference between the theoretical and the empirical dis-

14. Sampling from a null distribution is equivalent to sampling n units with replacement from a population of k elements, where the sampling weights for the elements correspond to the theoretical probabilities of the categories.

tribution function. In the discrete case, the statistic can be expressed as

$$D = \max_j |H(j) - F(j)|, \quad j = 1, \ldots, k$$

with

$$H(j) = \frac{1}{n}\sum_{i=1}^{j} h_i \quad \text{and} \quad F(j) = \frac{1}{n}\sum_{i=1}^{j} f_i$$

where h_j and f_j denote the expected and observed counts for category j (also see, e.g., Wood and Altavela 1978).

The distribution of the Kolmogorov-Smirnov statistic for continuous data is well known (see [R] **ksmirnov**), but does not hold for discrete data (e.g. Conover 1972). `mgof` therefore performs the discrete Kolmogorov-Smirnov test without making assumptions about the distribution of D using Monte Carlo simulation or exhaustive enumeration.

In the Benford example in Section 3, an argument could be put forward that the Kolmogorov-Smirnov test is more appropriate than the multinomial tests because the digits do have a natural order. Indeed, the p-value from the Kolmogorov-Smirnov test is considerably lower than the p-values based on Pearson's X^2 or the likelihood ratio:

```
. mgof firstdigit = log10(1+1/firstdigit), mc ksmirnov
Percent completed (10000 replications)
0 ——— 20 ——— 40 ——— 60 ——— 80 ——— 100
..................................................
                                          Number of obs =     313
                                          N of outcomes =       9
                                          Replications  =   10000
```

Goodness-of-fit	Coef.	Exact P-value	[99% Conf.	Interval]
Pearson's X2	6.226606	0.6311	0.6186	0.6435
Log likelihood ratio	6.475677	0.6093	0.5966	0.6219
Kolmogorov-Smirnov D	.0582185	0.0967	0.0892	0.1046

6 Concluding remarks

A new and flexible command for multinomial goodness-of-fit tests was introduced. The main features of the command are that it can be used with complex survey designs and that it offers methods to determine exact p-values in small samples. Two limitations should be mentioned.

The second-order Rao-Scott correction, which is used by the command to account for survey design is an improvement over performing uncorrected tests or using the Wald statistic, as was illustrated in the simulations in Section 4. However, there is also some evidence that the correction is not always optimal. For example Magnussen and Köhl (2006) is a study in which the second-order Rao-Scott correction did not perform

as well as certain other procedures in the context of single-stage cluster sampling. A comprehensive simulation in which different procedures are systematically evaluated under various survey designs and data structures would be valuable.

Further, computational speed is a major concern when computing exact p-values. The exhaustive enumeration method is slow or even unfeasible unless the sample is very small. Some speed gains could be made if the underlying combinatorial algorithms, which are currently implemented in Mata, would be ported to C, although this would not much increase the range of feasible applications. A more promising approach would be to implement fast algorithms for exact p-values extending the work of Mehta and Patel (1983), Baglivo et al. (1992), or Hirji (1997). However, the returns on this appeared limited to me given the availability of the Monte Carlo approximation method and given the result that for Pearson's X^2 or Cressie-Read's $D^2(2/3)$ (but not for G^2) the χ^2 approximation is usually quite good even with relatively small samples (e.g. Read 1984; for a somewhat more conservative view Formann 1995). An exception may be if many very small p-values have to be estimated with great accuracy (Keich and Nagarajan 2006). Although currently not planned, a future extension of the command to cover such fast algorithms might therefore be desirable.

7 Acknowledgements

I would like to thank Bill Gould and Debra Hevenstone for their helpful comments.

This research was supported by ETH Research Grant Number TH-8/04-3 on "Empirical studies in scientific misconduct: Prevalence and methods for detecting fraud and errors using the Benford distribution".

8 References

Andrews, G. E. 1984. *The Theory of Partitions*. Cambridge: Cambridge University Press.

Andrews, G. E., and K. Eriksson. 2004. *Integer Partitions*. Cambridge: Cambridge University Press.

Baglivo, J., D. Olivier, and M. Pagano. 1992. Methods for Exact Goodness-of-Fit Tests. *Journal of the American Statistical Association* 87(418): 464–469.

Benford, F. 1938. The law of anomalous numbers. *Proceedings of the American Philosophical Society* 78(4): 551572.

Conover, W. J. 1972. A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions. *Journal of the American Statistical Association* 67(339): 591–596.

Cressie, N., and T. R. C. Read. 1984. Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society. Series B (Methodological)* 46(3): 440–464.

———. 1989. Pearson's X^2 and the Loglikelihood Ratio Statistic G^2: A Comparative Review. *International Statistical Review* 57(1): 19–43.

Formann, A. K. 1995. Small-sample comparison of the exact and asymptotic upper tail probabilities of chi-squared goodness-of-fit statistics: Pearson's X^2, likelihood ratio, and power-divergence statistic ($\lambda = 2/3$). *Journal of Statistical Computation and Simulation* 51(2-4): 369–384.

Hill, T. P. 1998. The first digit phenomenon. *American Scientist* 86(4): 358.

Hirji, K. F. 1997. A Comparison of Algorithms for Exact Goodness-of-Fit Tests for Multinomial Data. *Communications in StatisticsSimulation and Computation* 26(3): 1197–1227.

Horn, S. D. 1977. Goodness-of-Fit Tests for Discrete Data: A Review and an Application to a Health Impairment Scale. *Biometrics* 33(1): 237–247.

Jann, B. 2005. moremata: module to provide various Mata functions. Available from http://ideas.repec.org/c/boc/bocode/s455001.html.

———. 2006. gsample: Stata module to draw a random sample. Available from http://ideas.repec.org/c/boc/bocode/s456716.html.

Keich, U., and N. Nagarajan. 2006. A fast and numerically robust method for exact multinomial goodness-of-fit test. *Journal of Computational & Graphical Statistics* 15(4): 779–802.

Magnussen, S., and M. Köhl. 2006. A better alternative to Wald's test-statistic for simple goodness-of-fit tests under one-stage cluster sampling. *Forest Ecology and Management* 221(1-3): 123–132.

Mehta, C. R., and N. R. Patel. 1983. A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables. *Journal of the American Statistical Association* 78(382): 427–434.

Newcomb, S. 1881. Note on the Frequency of Use of the Different Digits in Natural Numbers. *American Journal of Mathematics* 4(1): 39–40.

Nijenguis, A., and H. S. Wilf. 1978. *Combinatorial Algorithms. For Computers and Calculators*. Second edition ed. New York: Academic Press.

Radlow, R., and J. Alf, Edward F. 1975. An Alternate Multinomial Assessment of the Accuracy of the χ^2 Test of Goodness of Fit. *Journal of the American Statistical Association* 70(352): 811–813.

Rai, A., A. K. Srivastava, and H. C. Gupta. 2001. Small sample comparison of modified chi-square test statistics for survey data. *Biometrical Journal* 43(4): 483–495.

Rao, J. N. K., and A. J. Scott. 1981. The Analysis of Categorical Data From Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association* 76(374): 221–230.

Read, T. R. C. 1984. Small-Sample Comparisons for the Power Divergence Goodness-of-Fit Statistics. *Journal of the American Statistical Association* 79(388): 929–935.

Read, T. R. C., and N. A. C. Cressie. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer.

Reingold, E. M., J. Nievergelt, and N. Deo. 1977. *Combinatorial Algorithms: Theory and Practice*. Englewood Cliffs, NJ: Prentice-Hall.

Sokal, R. R., and F. J. Rohlf. 1995. *Biometry. The Principles and Practice of Statistics in Biological Research*. Third edition ed. New York: W. H. Freeman and Company.

Sribney, W. M. 1998. Two-way contingency tables for survey or clustered data. *Stata Technical Bulletin Reprints* 8: 297–322.

Thomas, D. R., and J. N. K. Rao. 1987. Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics Under Cluster Sampling. *Journal of the American Statistical Association* 82(398): 630–636.

Thomas, D. R., A. C. Singh, and G. R. Roberts. 1996. Tests of independence on two-way tables under cluster sampling: An evaluation. *International Statistical Review* 64(3): 295–311.

Weesie, J. 1997. sg68: Goodness-of-fit statistics for multinomial distributions. *Stata Technical Bulletin Reprints* 6: 183–186.

Wood, C. L., and M. M. Altavela. 1978. Large-Sample Results for Kolmogorov-Smirnov Statistics for Discrete Distributions. *Biometrika* 65(1): 235–239.

Zohgbi, A., and I. Stojmenovic. 1998. Fast Algorithms for Generating Partitions. *International Journal of Computer Mathematics* 70: 319–332.