svy: tabulate twoway — Two-way tables for survey data

Description Quick start Menu Syntax
Options Remarks and examples Stored results Methods and formulas
References Also see

Description

`svy: tabulate` produces two-way tabulations with tests of independence for complex survey data. See [SVY] **svy: tabulate oneway** for one-way tabulations for complex survey data.

Quick start

Two-way table of weighted cell proportions for `v1` and `v2` using `svyset` data

```
svy: tabulate v1 v2
```

Same as above, but with a test of independence using Pearson's χ^2 statistic with and without correction for the complex design

```
svy: tabulate v1 v2, pearson
```

Within-row and within-column proportions

```
svy: tabulate v1 v2, row column
```

95% confidence intervals for within-column proportions

```
svy: tabulate v1 v2, column ci
```

Unweighted numbers of observations and weighted counts

```
svy: tabulate v1 v2, obs count
```

Same as above, but display large counts in a more readable format

```
svy: tabulate v1 v2, obs count format(%11.0fc)
```

Weighted counts in the subpopulation defined by `v3` > 0

```
svy, subpop(v3): tabulate v1 v2, count
```

Menu

Statistics > Survey data analysis > Tables > Two-way tables

Syntax

Basic syntax

`svy: tabulate` *varname*$_1$ *varname*$_2$

Full syntax

`svy` [*vcetype*] [, *svy_options*] `:` `tabulate` *varname*$_1$ *varname*$_2$ [*if*] [*in*]
[, *tabulate_options display_items display_options statistic_options*]

Syntax to report results

`svy` [, *display_items display_options statistic_options*]

vcetype	Description
SE	
`linearized`	Taylor-linearized variance estimation
`bootstrap`	bootstrap variance estimation; see [SVY] **svy bootstrap**
`brr`	BRR variance estimation; see [SVY] **svy brr**
`jackknife`	jackknife variance estimation; see [SVY] **svy jackknife**
`sdr`	SDR variance estimation; see [SVY] **svy sdr**

Specifying a *vcetype* overrides the default from `svyset`.

svy_options	Description
if/in	
`subpop(`[*varname*] [*if*]`)`	identify a subpopulation
SE	
bootstrap_options	more options allowed with bootstrap variance estimation; see [SVY] ***bootstrap_options***
brr_options	more options allowed with BRR variance estimation; see [SVY] ***brr_options***
jackknife_options	more options allowed with jackknife variance estimation; see [SVY] ***jackknife_options***
sdr_options	more options allowed with SDR variance estimation; see [SVY] ***sdr_options***

`svy` requires that the survey design variables be identified using `svyset`; see [SVY] **svyset**.

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

Warning: Using `if` or `in` restrictions will often not produce correct variance estimates for subpopulations. To compute estimates for subpopulations, use the `subpop()` option.

tabulate_options	Description
Model	
stdize(*varname*)	variable identifying strata for standardization
stdweight(*varname*)	weight variable for standardization
tab(*varname*)	variable for which to compute cell totals/proportions
missing	treat missing values like other values
Collect	
collect	post results to collection Tabulate
collect([*cname*][, *collect_options*])	post results to a named collection

collect_options	Description
append	append results to an existing collection
replace	replace results of an existing collection
label(*filename*)	specify the collection labels
style(*filename*[, override])	specify the collection style

display_items	Description
Table items	
cell	cell proportions
count	weighted cell counts
column	within-column proportions
row	within-row proportions
se	standard errors
ci	confidence intervals
deff	display the DEFF design effects
deft	display the DEFT design effects
cv	display the coefficient of variation
srssubpop	report design effects assuming SRS within subpopulation
obs	cell observations

When any of se, ci, deff, deft, cv, or srssubpop is specified, only one of cell, count, column, or row can be specified. If none of se, ci, deff, deft, cv, or srssubpop is specified, any of or all cell, count, column, and row can be specified.

display_options	Description
Reporting	
`level(#)`	set confidence level; default is `level(95)`
`proportion`	display proportions; the default
`percent`	display percentages instead of proportions
`vertical`	stack confidence interval endpoints vertically
`nomarginals`	suppress row and column marginals
`nolabel`	suppress displaying value labels
`notable`	suppress displaying the table
`cellwidth(#)`	cell width
`csepwidth(#)`	column-separation width
`stubwidth(#)`	stub width
`format(`*%fmt*`)`	cell format; default is `format(%6.0g)`

`proportion` and `notable` are not shown in the dialog box.

statistic_options	Description
Test statistics	
`pearson`	Pearson's χ^2
`lr`	likelihood ratio
`null`	display null-based statistics
`wald`	adjusted Wald
`llwald`	adjusted log-linear Wald
`noadjust`	report unadjusted Wald statistics

Options

svy_options; see [SVY] **svy**.

Model

`stdize(`*varname*`)` specifies that the point estimates be adjusted by direct standardization across the strata identified by *varname*. This option requires the `stdweight()` option.

`stdweight(`*varname*`)` specifies the weight variable associated with the standard strata identified in the `stdize()` option. The standardization weights must be constant within the standard strata.

`tab(`*varname*`)` specifies that counts be cell totals of this variable and that proportions (or percentages) be relative to (that is, weighted by) this variable. For example, if this variable denotes income, the cell "counts" are instead totals of income for each cell, and the cell proportions are proportions of income for each cell.

`missing` specifies that missing values of *varname*$_1$ and *varname*$_2$ be treated as another row or column category rather than be omitted from the analysis (the default).

Collect

collect and collect([*cname*][, *collect_options*]) specify that results be posted to a collection. This collection produces a table that you can customize and publish to Microsoft Word, Microsoft Excel, PDF, HTML, LaTeX, SMCL, or Markdown. Output does not change when these options are specified. Use collect preview to see the customizeable table.

collect is a shortcut for collect(Tabulate).

cname specifies that a collection named *cname* be associated with the collected results. The default is Tabulate.

append specifies that results be appended to collection *cname*.

replace permits tabulate to overwrite an existing collection. This option is implied for collection Tabulate when append is not specified.

label(*filename*) specifies the *filename* containing the collection labels to use for your table. Labels in *filename* will be loaded into the collection, and any labels not specified in *filename* will be taken from the labels defined in c(collect_label). The default is to use only the collection labels set in c(collect_label); see [TABLES] **set collect_label**.

style(*filename*[, override]) specifies the *filename* containing the collection styles to use for your table. The default collection styles will be discarded, and only the collection styles in *filename* will be applied.

If you prefer the default collection styles but also want to apply any styles in *filename*, specify override. If there are conflicts between the default collection styles and those in *filename*, the ones in *filename* will take precedence.

The default is to use only the collection styles set in c(tabulate_style); see [TABLES] **set tabulate_style**.

Table items

cell requests that cell proportions (or percentages) be displayed. This is the default if none of count, row, or column is specified.

count requests that weighted cell counts be displayed.

column or row requests that column or row proportions (or percentages) be displayed.

se requests that the standard errors of cell proportions (the default), weighted counts, or row or column proportions be displayed. When se (or ci, deff, deft, or cv) is specified, only one of cell, count, row, or column can be selected. The standard error computed is the standard error of the one selected.

ci requests confidence intervals for cell proportions, weighted counts, or row or column proportions. The confidence intervals are constructed using a logit transform so that their endpoints always lie between 0 and 1.

deff and deft request that the design-effect measures DEFF and DEFT be displayed for each cell proportion, count, or row or column proportion. See [SVY] **estat** for details. The mean generalized DEFF is also displayed when deff, deft, or subpop is requested; see *Methods and formulas* for an explanation.

The deff and deft options are not allowed with estimation results that used direct standardization or poststratification.

cv requests that the coefficient of variation be displayed for each cell proportion, count, or row or column proportion. See [SVY] **estat** for details.

srssubpop requests that DEFF and DEFT be computed using an estimate of SRS (simple random sampling) variance for sampling within a subpopulation. By default, DEFF and DEFT are computed using an estimate of the SRS variance for sampling from the entire population. Typically, srssubpop would be given when computing subpopulation estimates by strata or by groups of strata.

obs requests that the number of observations for each cell be displayed.

Reporting

level(#) specifies the confidence level, as a percentage, for confidence intervals. The default is level(95) or as set by set level; see [U] **20.8 Specifying the width of confidence intervals**.

proportion, the default, requests that proportions be displayed.

percent requests that percentages be displayed instead of proportions.

vertical requests that the endpoints of confidence intervals be stacked vertically on display.

nomarginals requests that row and column marginals not be displayed.

nolabel requests that variable labels and value labels be ignored.

notable prevents the header and table from being displayed in the output. When specified, only the results of the requested test statistics are displayed. This option may not be specified with any other option in *display_options* except the level() option.

cellwidth(#), csepwidth(#), and stubwidth(#) specify widths of table elements in the output; see [P] **tabdisp**. Acceptable values for the stubwidth() option range from 4 to 32.

format(*%fmt*) specifies a format for the items in the table. The default is format(%6.0g). See [U] **12.5 Formats: Controlling how data are displayed**.

Test statistics

pearson requests that the Pearson χ^2 statistic be computed. By default, this is the test of independence that is displayed. The Pearson χ^2 statistic is corrected for the survey design with the second-order correction of Rao and Scott (1984) and is converted into an F statistic. One term in the correction formula can be calculated using either observed cell proportions or proportions under the null hypothesis (that is, the product of the marginals). By default, observed cell proportions are used. If the null option is selected, then a statistic corrected using proportions under the null hypothesis is displayed as well.

lr requests that the likelihood-ratio test statistic for proportions be computed. This statistic is not defined when there are one or more zero cells in the table. The statistic is corrected for the survey design by using the same correction procedure that is used with the pearson statistic. Again either observed cell proportions or proportions under the null hypothesis can be used in the correction formula. By default, the former is used; specifying the null option gives both the former and the latter. Neither variant of this statistic is recommended for sparse tables. For nonsparse tables, the lr statistics are similar to the corresponding pearson statistics.

null modifies the pearson and lr options only. If null is specified, two corrected statistics are displayed. The statistic labeled “D-B (null)” (“D-B” stands for design-based) uses proportions under the null hypothesis (that is, the product of the marginals) in the Rao and Scott (1984) correction. The statistic labeled merely “Design-based” uses observed cell proportions. If null is not specified, only the correction that uses observed proportions is displayed.

wald requests a Wald test of whether observed weighted proportions equal the product of the marginals (Koch, Freeman, and Freeman 1975). By default, an adjusted F statistic is produced; an unadjusted statistic can be produced by specifying noadjust. The unadjusted F statistic can yield extremely anticonservative p-values (that is, p-values that are too small) when the degrees of freedom of the variance estimates (the number of sampled PSUs minus the number of strata) are small relative to the $(R-1)(C-1)$ degrees of freedom of the table (where R is the number of rows and C is the number of columns). Hence, the statistic produced by wald and noadjust should not be used for inference unless it is essentially identical to the adjusted statistic.

This option must be specified at run time to be used on subsequent calls to svy to report results.

llwald requests a Wald test of the log-linear model of independence (Koch, Freeman, and Freeman 1975). The statistic is not defined when there are one or more zero cells in the table. The adjusted statistic (the default) can produce anticonservative p-values, especially for sparse tables, when the degrees of freedom of the variance estimates are small relative to the degrees of freedom of the table. Specifying noadjust yields a statistic with more severe problems. Neither the adjusted nor the unadjusted statistic is recommended for inference; the statistics are made available only for pedagogical purposes.

noadjust modifies the wald and llwald options only. It requests that an unadjusted F statistic be displayed in addition to the adjusted statistic.

svy: tabulate uses the tabdisp command (see [P] **tabdisp**) to produce the table. Only five items can be displayed in the table at one time. The ci option implies two items. If too many items are selected, a warning will appear immediately. To view more items, redisplay the table while specifying different options.

Remarks and examples

Remarks are presented under the following headings:

Introduction
The Rao and Scott correction
Wald statistics
Properties of the statistics
Publish your tables

Introduction

Despite the long list of options for svy: tabulate, it is a simple command to use. Using the svy: tabulate command is just like using tabulate to produce two-way tables for ordinary data. The main difference is that svy: tabulate computes a test of independence that is appropriate for complex survey data.

The test of independence that is displayed by default is based on the usual Pearson χ^2 statistic for two-way tables. To account for the survey design, the statistic is turned into an F statistic with noninteger degrees of freedom by using a second-order Rao and Scott (1981, 1984) correction. Although the theory behind the Rao and Scott correction is complicated, the p-value for the corrected F statistic can be interpreted in the same way as a p-value for the Pearson χ^2 statistic for "ordinary" data (that is, data that are assumed independent and identically distributed [i.i.d.]).

svy: tabulate, in fact, computes four statistics for the test of independence with two variants of each, for a total of eight statistics. The option combination for each of the eight statistics are the following:

1. pearson (the default)
2. pearson null
3. lr
4. lr null
5. wald
6. wald noadjust
7. llwald
8. llwald noadjust

The wald and llwald options with noadjust yield the statistics developed by Koch, Freeman, and Freeman (1975), which have been implemented in the CROSSTAB procedure of the SUDAAN software (Research Triangle Institute 1997 , release 7.5).

These eight statistics, along with other variants, have been evaluated in simulations (Sribney 1998). On the basis of these simulations, we advise researchers to use the default statistic (the pearson option) in all situations. We recommend that the other statistics be used only for comparative or pedagogical purposes. Sribney (1998) gives a detailed comparison of the statistics; a summary of his conclusions is provided later in this entry.

Other than the test-statistic options (*statistic_options*) and the survey design options (*svy_options*), most of the other options of svy: tabulate simply relate to different choices for what can be displayed in the body of the table. By default, cell proportions are displayed, but viewing either row or column proportions or weighted counts usually makes more sense.

Standard errors and confidence intervals can optionally be displayed for weighted counts or cell, row, or column proportions. The confidence intervals for proportions are constructed using a logit transform so that their endpoints always lie between 0 and 1. Associated design effects (DEFF and DEFT) can be viewed for the variance estimates. The mean generalized DEFF (Rao and Scott 1984) is also displayed when option deff, deft, or srssubpop is specified. The mean generalized DEFF is essentially a design effect for the asymptotic distribution of the test statistic; see the *Methods and formulas* section at the end of this entry.

▷ Example 1

Using data from the Second National Health and Nutrition Examination Survey (NHANES II) (McDowell et al. 1981), we identify the survey design characteristics with `svyset` and then produce a two-way table of cell proportions with `svy: tabulate`.

```
. use https://www.stata-press.com/data/r19/nhanes2b

. svyset psuid [pweight=finalwgt], strata(stratid)

Sampling weights: finalwgt
             VCE: linearized
     Single unit: missing
        Strata 1: stratid
 Sampling unit 1: psuid
           FPC 1: <zero>

. svy: tabulate race diabetes
(running tabulate on estimation sample)

Number of strata = 31                            Number of obs     =      10,349
Number of PSUs   = 62                            Population size = 117,131,111
                                                 Design df         =          31

----------------------------------------
          |       Diabetes status
     Race | Not diab  Diabetic     Total
----------+-----------------------------
          |
    White |     .851     .0281     .8791
    Black |    .0899     .0056     .0955
    Other |    .0248   5.2e-04     .0253
          |
    Total |    .9658     .0342         1
----------------------------------------
Key: Cell proportion

  Pearson:
    Uncorrected   chi2(2)         =   21.3483
    Design-based  F(1.52, 47.26)  =   15.0056     P = 0.0000
```

The default table displays only cell proportions, and this makes it difficult to compare the incidence of diabetes in white, black, and “other” racial groups. It would be better to look at row proportions. This can be done by redisplaying the results (that is, reissuing the command without specifying any variables) with the `row` option.

```
. svy: tabulate, row

Number of strata = 31                            Number of obs     =      10,349
Number of PSUs   = 62                            Population size = 117,131,111
                                                 Design df         =          31

----------------------------------------
          |       Diabetes status
     Race | Not diab  Diabetic     Total
----------+-----------------------------
          |
    White |     .968      .032         1
    Black |     .941      .059         1
    Other |    .9797     .0203         1
          |
    Total |    .9658     .0342         1
----------------------------------------
Key: Row proportion

  Pearson:
    Uncorrected   chi2(2)         =   21.3483
    Design-based  F(1.52, 47.26)  =   15.0056     P = 0.0000
```

This table is much easier to interpret. A larger proportion of blacks have diabetes than do whites or persons in the “other” racial category. The test of independence for a two-way contingency table is equivalent to the test of homogeneity of row (or column) proportions. Hence, we can conclude that there is a highly significant difference between the incidence of diabetes among the three racial groups.

We may now wish to compute confidence intervals for the row proportions. If we try to redisplay, specifying `ci` along with `row`, we get the following result:

```
. svy: tabulate, row ci
confidence intervals are only available for cells
    To compute row confidence intervals, rerun command with row and ci options.
r(111);
```

There are limits to what `svy: tabulate` can redisplay. Basically, any of the options relating to variance estimation (that is, `se`, `ci`, `deff`, and `deft`) must be specified at run time along with the single item (that is, `count`, `cell`, `row`, or `column`) for which you want standard errors, confidence intervals, DEFF, or DEFT. So to get confidence intervals for row proportions, we must rerun the command. We do so below, requesting not only `ci` but also `se`.

```
. svy: tabulate race diabetes, row se ci format(%7.4f)
(running tabulate on estimation sample)

Number of strata = 31                            Number of obs   =     10,349
Number of PSUs   = 62                            Population size = 117,131,111
                                                 Design df       =          31

-------------------------------------------------------------
          |                  Diabetes status
     Race |        Not diab         Diabetic            Total
----------+--------------------------------------------------
          |
    White |          0.9680           0.0320           1.0000
          |        (0.0020)         (0.0020)
          | [0.9638,0.9718]  [0.0282,0.0362]
          |
    Black |          0.9410           0.0590           1.0000
          |        (0.0061)         (0.0061)
          | [0.9271,0.9523]  [0.0477,0.0729]
          |
    Other |          0.9797           0.0203           1.0000
          |        (0.0076)         (0.0076)
          | [0.9566,0.9906]  [0.0094,0.0434]
          |
    Total |          0.9658           0.0342           1.0000
          |        (0.0018)         (0.0018)
          | [0.9619,0.9693]  [0.0307,0.0381]
----------+--------------------------------------------------

Key: Row proportion
     (Linearized standard error of row proportion)
     [95% confidence interval for row proportion]

  Pearson:
    Uncorrected   chi2(2)         =   21.3483
    Design-based  F(1.52, 47.26)  =   15.0056     P = 0.0000
```

In the above table, we specified a `%7.4f` format rather than using the default `%6.0g` format. The single format applies to every item in the table. We can omit the marginal totals by specifying `nomarginals`. If the above style for displaying the confidence intervals is obtrusive—and it can be in a wider table—we can use the `vertical` option to stack the endpoints of the confidence interval, one over the other, and omit the brackets (the parentheses around the standard errors are also omitted when `vertical` is specified). To express results as percentages, as with the `tabulate` command (see [R] **tabulate twoway**), we can use the percent option. Or we can play around with these display options until we get a table that we are satisfied with, first making changes to the options on redisplay (that is, omitting the cross-tabulated variables when we issue the command).

◁

❑ Technical note

The standard errors computed by svy: tabulate are the same as those produced by svy: mean, svy: proportion, and svy: ratio. Indeed, svy: tabulate uses these commands as subroutines to produce its table.

In the previous example, the estimate of the proportion of African Americans with diabetes (the second proportion in the second row of the preceding table) is simply a ratio estimate; hence, we can also obtain the same estimates by using svy: ratio:

```
. drop black
. generate black = (race==2) if !missing(race)
. generate diablk = diabetes*black
(2 missing values generated)
. svy: ratio diablk/black
(running ratio on estimation sample)

Survey: Ratio estimation

Number of strata = 31          Number of obs   =      10,349
Number of PSUs   = 62          Population size = 117,131,111
                               Design df       =          31

     _ratio_1: diablk/black

─────────────┬────────────────────────────────────────────────
             │             Linearized
             │      Ratio   std. err.     [95% conf. interval]
─────────────┼────────────────────────────────────────────────
    _ratio_1 │   .0590349    .0061443      .0465035    .0715662
─────────────┴────────────────────────────────────────────────
```

Although the standard errors are the same, the confidence intervals are slightly different. The svy: tabulate command produced the confidence interval $[\,0.0477, 0.0729\,]$, and svy: ratio gave $[\,0.0465, 0.0716\,]$. The difference is because svy: tabulate uses a logit transform to produce confidence intervals whose endpoints are always between 0 and 1. This transformation also shifts the confidence intervals slightly toward 0.5, which is beneficial because the untransformed confidence intervals tend to be, on average, biased away from 0.5. See *Methods and formulas* for details.

❑

▷ Example 2: The tab() option

The `tab()` option allows us to compute proportions relative to a certain variable. Suppose that we wish to compare the proportion of total income among different racial groups in males with that of females. We do so below with fictitious data:

```
. use https://www.stata-press.com/data/r19/svy_tabopt, clear

. svy: tabulate gender race, tab(income) row
(running tabulate on estimation sample)

Number of strata = 31                            Number of obs   =      10,351
Number of PSUs   = 62                            Population size = 117,157,513
                                                 Design df       =          31

──────────────────────────────────────
          │          Race
   Gender │ White  Black  Other  Total
──────────┼───────────────────────────
     Male │ .8857  .0875  .0268      1
   Female │  .884   .094   .022      1
          │
    Total │ .8848  .0909  .0243      1
──────────────────────────────────────

  Tabulated variable:  income

Key: Row proportion

  Pearson:
    Uncorrected   chi2(2)         =    3.6241
    Design-based  F(1.91, 59.12)  =    0.8626     P = 0.4227
```

◁

The Rao and Scott correction

`svy: tabulate` can produce eight different statistics for the test of independence. By default, `svy: tabulate` displays the Pearson χ^2 statistic with the Rao and Scott (1981, 1984) second-order correction. On the basis of simulations Sribney (1998), we recommend that you use this statistic in all situations. The statistical literature, however, contains several alternatives, along with other possibilities for implementing the Rao and Scott correction. Hence, for comparative or pedagogical purposes, you may want to view some of the other statistics computed by `svy: tabulate`. This section briefly describes the differences among these statistics; for a more detailed discussion, see Sribney (1998).

Two statistics commonly used for i.i.d. data for the test of independence of $R \times C$ tables (R rows and C columns) are the Pearson χ^2 statistic

$$X_{\mathrm{P}}^2 = m \sum_{r=1}^{R} \sum_{c=1}^{C} \left(\widehat{p}_{rc} - \widehat{p}_{0rc}\right)^2 / \widehat{p}_{0rc}$$

and the likelihood-ratio χ^2 statistic

$$X_{\mathrm{LR}}^2 = 2m \sum_{r=1}^{R} \sum_{c=1}^{C} \widehat{p}_{rc} \ln\left(\widehat{p}_{rc}/\widehat{p}_{0rc}\right)$$

where m is the total number of sampled individuals, $\widehat{p}_{rc}$ is the estimated proportion for the cell in the rth row and cth column of the table, and $\widehat{p}_{0rc}$ is the estimated proportion under the null hypothesis of independence; that is, $\widehat{p}_{0rc} = \widehat{p}_{r\cdot}\widehat{p}_{\cdot c}$, the product of the row and column marginals: $\widehat{p}_{r\cdot} = \sum_{c=1}^{C} \widehat{p}_{rc}$ and $\widehat{p}_{\cdot c} = \sum_{r=1}^{R} \widehat{p}_{rc}$.

For i.i.d. data, both these statistics are distributed asymptotically as $\chi^2_{(R-1)(C-1)}$. The likelihood-ratio statistic is not defined when one or more of the cells in the table are empty. The Pearson statistic, however, can be calculated when one or more cells in the table are empty—the statistic may not have good properties in this case, but the statistic still has a computable value.

For survey data, X^2_{P} and X^2_{LR} can be computed using weighted estimates of $\widehat{p}_{rc}$ and $\widehat{p}_{0rc}$. However, for a complex sampling design, one can no longer claim that they are distributed as $\chi^2_{(R-1)(C-1)}$, but you can estimate the variance of $\widehat{p}_{rc}$ under the sampling design. For instance, in Stata, this variance can be estimated via linearization methods by using `svy: mean` or `svy: ratio`.

Rao and Scott (1981, 1984) derived the asymptotic distribution of X^2_{P} and X^2_{LR} in terms of the variance of $\widehat{p}_{rc}$. Unfortunately, the result (see (1) in *Methods and formulas*) is not computationally feasible, but it can be approximated using correction formulas. `svy: tabulate` uses the second-order correction developed by Rao and Scott (1984). By default, or when the `pearson` option is specified, `svy: tabulate` displays the second-order correction of the Pearson statistic. The `lr` option gives the second-order correction of the likelihood-ratio statistic. Because it is the default of `svy: tabulate`, the correction computed with $\widehat{p}_{rc}$ is referred to as the default correction.

The Rao and Scott papers, however, left some details outstanding about the computation of the correction. One term in the correction formula can be computed using either $\widehat{p}_{rc}$ or $\widehat{p}_{0rc}$. Because under the null hypothesis both are asymptotically equivalent, theory offers no guidance about which is best. By default, `svy: tabulate` uses $\widehat{p}_{rc}$ for the corrections of the Pearson and likelihood-ratio statistics. If the `null` option is specified, the correction is computed using $\widehat{p}_{0rc}$. For nonsparse tables, these two correction methods yield almost identical results. However, in simulations of sparse tables, Sribney (1998) found that the null-corrected statistics were extremely anticonservative for 2 × 2 tables (that is, under the null, “significance” was declared too often) and were too conservative for other tables. The default correction, however, had better properties. Hence, we do not recommend using `null`.

For the computational details of the Rao and Scott–corrected statistics, see *Methods and formulas*.

Wald statistics

Prior to the work by Rao and Scott (1981, 1984), Wald tests for the test of independence for two-way tables were developed by Koch, Freeman, and Freeman (1975). Two Wald statistics have been proposed. The first, similar to the Pearson statistic, is based on

$$\widehat{Y}_{rc} = \widehat{N}_{rc} - \widehat{N}_{r\cdot}\widehat{N}_{\cdot c}/\widehat{N}_{\cdot\cdot}$$

where $\widehat{N}_{rc}$ is the estimated weighted count for the r, cth cell. The delta method can be used to approximate the variance of $\widehat{Y}_{rc}$, and a Wald statistic can be calculated as usual. A second Wald statistic can be constructed based on a log-linear model for the table. Like the likelihood-ratio statistic, this statistic is undefined when there is a zero proportion in the table.

These Wald statistics are initially χ^2 statistics, but they have better properties when converted into F statistics with denominator degrees of freedom that account for the degrees of freedom of the variance estimator. They can be converted to F statistics in two ways.

One method is the standard manner: divide by the χ^2 degrees of freedom $d_0 = (R-1)(C-1)$ to get an F statistic with d_0 numerator degrees of freedom and $\nu = n - L$ denominator degrees of freedom. This is the form of the F statistic suggested by Koch, Freeman, and Freeman (1975) and implemented in the CROSSTAB procedure of the SUDAAN software (Research Triangle Institute 1997 , release 7.5), and it is the method used by `svy: tabulate` when the `noadjust` option is specified with `wald` or `llwald`.

Another technique is to adjust the F statistic by using

$$F_{\text{adj}} = (\nu - d_0 + 1)W/(\nu d_0) \qquad \text{with} \qquad F_{\text{adj}} \sim F(d_0, \nu - d_0 + 1)$$

This is the default adjustment for `svy: tabulate`. `test` and the other `svy` estimation commands produce adjusted F statistics by default, using the same adjustment procedure. See Korn and Graubard (1990) for a justification of the procedure.

The adjusted F statistic is identical to the unadjusted F statistic when $d_0 = 1$, that is, for 2 × 2 tables.

As Thomas and Rao (1987) point out (also see Korn and Graubard [1990]), the unadjusted F statistics can become extremely anticonservative as d_0 increases when ν is small or moderate; that is, under the null, the statistics are "significant" far more often than they should be. Because the unadjusted statistics behave so poorly for larger tables when ν is not large, their use can be justified only for small tables or when ν is large. But when the table is small or when ν is large, the unadjusted statistic is essentially identical to the adjusted statistic. Hence, for statistical inference, looking at the unadjusted statistics has no point.

The adjusted "Pearson" Wald F statistic usually behaves reasonably under the null. However, even the adjusted F statistic for the log-linear Wald test tends to be moderately anticonservative when ν is not large (Thomas and Rao 1987 ; Sribney 1998).

▷ Example 3

With the NHANES II data, we tabulate, for the male subpopulation, high blood pressure (highbp) versus a variable (sizplace) that indicates the degree of urbanity/ruralness. We request that all eight statistics for the test of independence be displayed.

```
. use https://www.stata-press.com/data/r19/nhanes2b

. generate male = (sex==1) if !missing(sex)

. svy, subpop(male): tabulate highbp sizplace, col obs pearson lr null wald
> llwald noadj
(running tabulate on estimation sample)

Number of strata = 31                            Number of obs   =      10,351
Number of PSUs   = 62                            Population size = 117,157,513
                                                 Subpop. no. obs =       4,915
                                                 Subpop. size    =  56,159,480
                                                 Design df       =          31

----------+---------------------------------------------------------------
High      |
blood     |                     1=urban, ..., 8=rural
pressure  |      1      2      3      4      5      6      7      8  Total
----------+---------------------------------------------------------------
        0 |  .4949  .5884  .6768  .5308  .5563   .629  .5502  .5618  .5724
          |    241    326    381    228    121    135    186    993   2611
          |
        1 |  .5051  .4116  .3232  .4692  .4437   .371  .4498  .4382  .4276
          |    285    281    241    217    101     95    185    899   2304
          |
    Total |      1      1      1      1      1      1      1      1      1
          |    526    607    622    445    222    230    371   1892   4915
----------+---------------------------------------------------------------

Key: Column proportion
     Number of observations

  Pearson:
    Uncorrected   chi2(7)         =  114.9556
    D-B (null)    F(5.33, 165.13) =    2.1460    P = 0.0584
    Design-based  F(5.48, 169.80) =    2.4281    P = 0.0325

  Likelihood ratio:
    Uncorrected   chi2(7)         =  116.5144
    D-B (null)    F(5.33, 165.13) =    2.1751    P = 0.0552
    Design-based  F(5.48, 169.80) =    2.4610    P = 0.0305

  Wald (Pearson):
    Unadjusted    chi2(7)         =   11.1739
    Unadjusted    F(7, 31)        =    1.5963    P = 0.1735
    Adjusted      F(7, 25)        =    1.2873    P = 0.2967

  Wald (log-linear):
    Unadjusted    chi2(7)         =   14.9598
    Unadjusted    F(7, 31)        =    2.1371    P = 0.0688
    Adjusted      F(7, 25)        =    1.7235    P = 0.1490
```

The *p*-values from the null-corrected Pearson and likelihood-ratio statistics (lines labeled "D-B (null)"; "D-B" stands for "design-based") are bigger than the corresponding default-corrected statistics (lines labeled "Design-based"). Simulations (Sribney 1998) show that the null-corrected statistics are overly conservative for many sparse tables (except 2 × 2 tables); this appears to be the case here, although this table is hardly sparse. The default-corrected Pearson statistic has good properties under the null for both sparse and nonsparse tables; hence, the smaller *p*-value for it should be considered reliable.

The default-corrected likelihood-ratio statistic is usually similar to the default-corrected Pearson statistic except for sparse tables, when it tends to be anticonservative. This example follows this pattern, with its p-value being slightly smaller than that of the default-corrected Pearson statistic.

For tables of these dimensions (2×8), the unadjusted "Pearson" Wald and log-linear Wald F statistics are extremely anticonservative under the null when the variance degrees of freedom is small. Here the variance degrees of freedom is only 31 (62 PSUs minus 31 strata), so we expect that the unadjusted Wald F statistics yield smaller p-values than the adjusted F statistics. Because of their poor behavior under the null for small variance degrees of freedom, they cannot be trusted here. Simulations show that although the adjusted "Pearson" Wald F statistic has good properties under the null, it is often less powerful than the default Rao and Scott–corrected statistics. That is probably the explanation for the larger p-value for the adjusted "Pearson" Wald F statistic than that for the default-corrected Pearson and likelihood-ratio statistics.

The p-value for the adjusted log-linear Wald F statistic is about the same as that for the trustworthy default-corrected Pearson statistic. However, that is probably because of the anticonservatism of the log-linear Wald under the null balancing out its lower power under alternative hypotheses.

The "uncorrected" χ^2 Pearson and likelihood-ratio statistics displayed in the table are misspecified statistics; that is, they are based on an i.i.d. assumption, which is not valid for complex survey data. Hence, they are not correct, even asymptotically. The "unadjusted" Wald χ^2 statistics, on the other hand, are completely different. They are valid asymptotically as the variance degrees of freedom becomes large.

◁

Properties of the statistics

This section briefly summarizes the properties of the eight statistics computed by `svy: tabulate`. For details, see Sribney (1998), Rao and Thomas (1989), Thomas and Rao (1987), and Korn and Graubard (1990).

`pearson` is the Rao and Scott (1984) second-order corrected Pearson statistic, computed using $\widehat{p}_{rc}$ in the correction (default correction). It is displayed by default. Simulations show it to have good properties under the null for both sparse and nonsparse tables. Its power is similar to that of the `lr` statistic in most situations. It often appears to be more powerful than the adjusted "Pearson" Wald F statistic (`wald` option), especially for larger tables. We recommend using this statistic in all situations.

`pearson null` is the Rao and Scott second-order corrected Pearson statistic, computed using $\widehat{p}_{0rc}$ in the correction. It is numerically similar to the `pearson` statistic for nonsparse tables. For sparse tables, it can be erratic. Under the null, it can be anticonservative for sparse 2 × 2 tables but conservative for larger sparse tables.

`lr` is the Rao and Scott second-order corrected likelihood-ratio statistic, computed using $\widehat{p}_{rc}$ in the correction (default correction). The correction is identical to that for `pearson`. It is numerically similar to the `pearson` statistic for nonsparse tables. It can be anticonservative (p-values too small) in sparse tables. If there is a zero cell, it cannot be computed.

`lr null` is the Rao and Scott second-order corrected likelihood-ratio statistic, computed using $\widehat{p}_{0rc}$ in the correction. The correction is identical to that for `pearson null`. It is numerically similar to the `lr` statistic for nonsparse tables. For sparse tables, it can be overly conservative. If there is a zero cell, it cannot be computed.

`wald` statistic is the adjusted "Pearson" Wald F statistic. It has good properties under the null for nonsparse tables. It can be erratic for sparse 2 × 2 tables and some sparse large tables. The `pearson` statistic often appears to be more powerful.

`wald noadjust` is the unadjusted "Pearson" Wald F statistic. It can be extremely anticonservative under the null when the table degrees of freedom (number of rows minus one times the number of columns minus one) approaches the variance degrees of freedom (number of sampled PSUs minus the number of strata). It is the same as the adjusted `wald` statistic for 2 × 2 tables. It is similar to the adjusted `wald` statistic for small tables, large variance degrees of freedom, or both.

`llwald` statistic is the adjusted log-linear Wald F statistic. It can be anticonservative for both sparse and nonsparse tables. If there is a zero cell, it cannot be computed.

`llwald noadjust` statistic is the unadjusted log-linear Wald F statistic. Like `wald noadjust`, it can be extremely anticonservative under the null when the table degrees of freedom approaches the variance degrees of freedom. It also suffers from the same general anticonservatism of the `llwald` statistic. If there is a zero cell, it cannot be computed.

Publish your tables

With the `collect` option, `svy: tabulate` posts the tabulated values to a collection named `Tabulate` and sets it as the current collection. With collections, you can customize the look of your table and then publish it to HTML, Word, LaTeX, PDF, Excel, or another format appropriate for your report.

If you are not familiar with collections, see [TABLES] **Intro**. The predefined styles for `svy: tabulate` are documented in [TABLES] **Predefined styles**.

▷ Example 4

Recall the final table from example 1. Let's rerun that command but add the `collect` option to produce a collection with the reported values.

```
. svy: tabulate race diabetes, row se ci format(%7.4f) collect
(running tabulate on estimation sample)

Number of strata = 31                            Number of obs   =      10,349
Number of PSUs   = 62                            Population size = 117,131,111
                                                 Design df       =          31

------------------------------------------------------------
          |              Diabetes status
     Race |        Not diab          Diabetic          Total
----------+-------------------------------------------------
    White |          0.9680            0.0320         1.0000
          |        (0.0020)          (0.0020)
          | [0.9638,0.9718]   [0.0282,0.0362]
          |
    Black |          0.9410            0.0590         1.0000
          |        (0.0061)          (0.0061)
          | [0.9271,0.9523]   [0.0477,0.0729]
          |
    Other |          0.9797            0.0203         1.0000
          |        (0.0076)          (0.0076)
          | [0.9566,0.9906]   [0.0094,0.0434]
          |
    Total |          0.9658            0.0342         1.0000
          |        (0.0018)          (0.0018)
          | [0.9619,0.9693]   [0.0307,0.0381]
------------------------------------------------------------
Key: Row proportion
     (Linearized standard error of row proportion)
     [95% confidence interval for row proportion]

  Pearson:
    Uncorrected   chi2(2)         =   21.3483
    Design-based  F(1.52, 47.26)  =   15.0056     P = 0.0000
```

The output does not change; however, we can use the `collect dir` command to see that `svy: tabulate` created a collection named `Tabulate`.

```
. collect dir

Collections in memory
Current: Tabulate
--------------------
    Name   No. items
--------------------
Tabulate   42
 default   0
--------------------
```

In this collection, within-row proportions are tagged with `result[row_proportion]`, standard errors with `result[se]`, and confidence bounds with `result[_r_lb]` and `result[_r_ub]`. Here is a table of all *display_items* and their corresponding `result[`*level*`]` tag elements.

display_item	*level*
cell proportion	proportion
cell percent	percent
weighted cell count	count
column proportion	column_proportion
column percent	column_percent
row proportion	row_proportion
row percent	row_percent
se	se
ci	_r_lb
	_r_ub
deff	deff
deft	deft
cv	cv
obs	obs

The tabulated variables (that is, `race` and `diabetes`) are added to the collection as dimensions and are used to tag the collected results. In addition to the name, label, level values, and value labels of the tabulated variables, these dimensions also have the `__margCode__` level with the `Total` label for tagging the marginal values. Here we use `collect label list` to show the levels and labels of the dimension for each tabulated variable.

```
. collect label list race

  Collection: Tabulate
   Dimension: race
       Label: Race
Level labels:
           1  White
           2  Black
           3  Other
__margCode__  Total

. collect label list diabetes

  Collection: Tabulate
   Dimension: diabetes
       Label: Diabetes status
Level labels:
           0  Not diabetic
           1  Diabetic
__margCode__  Total
```

When tests of independence are reported, `svy: tabulate` uses a style that defines composite results for them. Each test of independence has a corresponding composite result that composes it in a single cell of a second table. Here is a table of all the tests and the corresponding name of the composite result.

statistic_options	Description	Composite result
`pearson`	Uncorrected Pearson's χ^2	`UncorrectedPearson_ma`
`pearson`	Design-based Pearson's F	`DesignbasedPearson_ma`
`pearson null`	Design-based (null) Pearson's F	`DBnullPearson_ma`
`lr`	Uncorrected likelihood ratio χ^2	`UncorrectedLR_ma`
`lr`	Design-based likelihood ratio F	`DesignbasedLR_ma`
`lr null`	Design-based (null) likelihood ratio F	`DBnullLR_ma`
`wald`	Unadjusted Wald χ^2	`UnadjustedWaldX2_ma`
`wald`	Adjusted Wald F	`AdjustedWaldF_ma`
`wald noadjust`	Unadjusted Wald F	`UnadjustedWaldF_ma`
`llwald`	Unadjusted Wald (log-linear) χ^2	`UnadjustedLLWX2_ma`
`llwald`	Adjusted Wald (log-linear) F	`AdjustedLLWF_ma`
`llwald noadjust`	Unadjusted Wald (log-linear) F	`UnadjustedLLWF_ma`

`svy: tabulate` constructs a default layout, so you can view your customizable table with the `collect preview` command. Here we use the `collect layout` command to report the default layout specification and corresponding table.

```
. collect layout

Collection: Tabulate
      Rows: race#result result
   Columns: diabetes _overall
    Tables: cmdset#item_type
   Table 1: 17 x 3
   Table 2: 2 x 1

--------------------------------------------------------------
                                Diabetes status
                        Not diabetic          Diabetic   Total
--------------------------------------------------------------
Race
  White
    Row proportion            0.9680            0.0320  1.0000
    Standard error          (0.0020)          (0.0020)
    95% CI           [0.9638 0.9718]   [0.0282 0.0362]
  Black
    Row proportion            0.9410            0.0590  1.0000
    Standard error          (0.0061)          (0.0061)
    95% CI           [0.9271 0.9523]   [0.0477 0.0729]
  Other
    Row proportion            0.9797            0.0203  1.0000
    Standard error          (0.0076)          (0.0076)
    95% CI           [0.9566 0.9906]   [0.0094 0.0434]
  Total
    Row proportion            0.9658            0.0342  1.0000
    Standard error          (0.0018)          (0.0018)
    95% CI           [0.9619 0.9693]   [0.0307 0.0381]
--------------------------------------------------------------

                       Uncorrected Pearson χ2(2) = 21.3483
Design-based Pearson F(1.52, 47.26) = 15.0056 P = <0.0001
```

We can make further changes to the table with the `collect` suite of commands. But we are happy with this layout and ready to publish the table to an HTML file with `collect export`. We simply specify the filename to which we want to export it.

```
. collect export table2.html
(collection Tabulate exported to file table2.html)
```

With `collect export`, you can publish the table to several formats, such as HTML, PDF, and LaTeX files, by specifying the appropriate file extension.

◁

Stored results

In addition to the results documented in [SVY] **svy**, `svy: tabulate` stores the following in `e()`:

Scalars

`e(r)`	number of rows
`e(c)`	number of columns
`e(cvgdeff)`	coefficient of variation of generalized DEFF eigenvalues
`e(mgdeff)`	mean generalized DEFF
`e(total)`	weighted sum of `tab()` variable
`e(F_Pear)`	default-corrected Pearson F
`e(F_Penl)`	null-corrected Pearson F
`e(df1_Pear)`	numerator d.f. for `e(F_Pear)`
`e(df2_Pear)`	denominator d.f. for `e(F_Pear)`
`e(df1_Penl)`	numerator d.f. for `e(F_Penl)`
`e(df2_Penl)`	denominator d.f. for `e(F_Penl)`
`e(p_Pear)`	p-value for `e(F_Pear)`
`e(p_Penl)`	p-value for `e(F_Penl)`
`e(cun_Pear)`	uncorrected Pearson χ^2
`e(cun_Penl)`	null variant uncorrected Pearson χ^2
`e(F_LR)`	default-corrected likelihood-ratio F
`e(F_LRnl)`	null-corrected likelihood-ratio F
`e(df1_LR)`	numerator d.f. for `e(F_LR)`
`e(df2_LR)`	denominator d.f. for `e(F_LR)`
`e(df1_LRnl)`	numerator d.f. for `e(F_LRnl)`
`e(df2_LRnl)`	denominator d.f. for `e(F_LRnl)`
`e(p_LR)`	p-value for `e(F_LR)`
`e(p_LRnl)`	p-value for `e(F_LRnl)`
`e(cun_LR)`	uncorrected likelihood-ratio χ^2
`e(cun_LRnl)`	null variant uncorrected likelihood-ratio χ^2
`e(F_Wald)`	adjusted “Pearson” Wald F
`e(F_LLW)`	adjusted log-linear Wald F
`e(p_Wald)`	p-value for `e(F_Wald)`
`e(p_LLW)`	p-value for `e(F_LLW)`
`e(Fun_Wald)`	unadjusted “Pearson” Wald F
`e(Fun_LLW)`	unadjusted log-linear Wald F
`e(pun_Wald)`	p-value for `e(Fun_Wald)`
`e(pun_LLW)`	p-value for `e(Fun_LLW)`
`e(cun_Wald)`	unadjusted “Pearson” Wald χ^2
`e(cun_LLW)`	unadjusted log-linear Wald χ^2

Macros

`e(cmd)`	`tabulate`
`e(tab)`	`tab()` variable
`e(rowlab)`	label or empty
`e(collab)`	label or empty
`e(rowvlab)`	row variable label

`e(colvlab)`	column variable label
`e(rowvar)`	*varname*$_1$, the row variable
`e(colvar)`	*varname*$_2$, the column variable
`e(setype)`	`cell`, `count`, `column`, or `row`
Matrices	
`e(Prop)`	matrix of cell proportions
`e(Obs)`	matrix of observation counts
`e(Deff)`	DEFF vector for `e(setype)` items
`e(Deft)`	DEFT vector for `e(setype)` items
`e(Row)`	values for row variable
`e(Col)`	values for column variable
`e(V_row)`	variance for row totals
`e(V_col)`	variance for column totals
`e(V_srs_row)`	$\mathrm{V}_{\mathrm{srs}}$ for row totals
`e(V_srs_col)`	$\mathrm{V}_{\mathrm{srs}}$ for column totals
`e(Deff_row)`	DEFF for row totals
`e(Deff_col)`	DEFF for column totals
`e(Deft_row)`	DEFT for row totals
`e(Deft_col)`	DEFT for column totals

Methods and formulas

Methods and formulas are presented under the following headings:

The table items
Confidence intervals
The test statistics

See *Coefficient of variation* under *Methods and formulas* of [SVY] **estat** for information on the coefficient of variation (the `cv` option).

The table items

For a table of R rows by C columns with cells indexed by r, c, let

$$y_{(rc)j} = \begin{cases} 1 & \text{if the } j\text{th observation of the data is in the } r,c\text{th cell} \\ 0 & \text{otherwise} \end{cases}$$

where $j = 1, \dots, m$ indexes individuals in the sample. Weighted cell counts (`count` option) are

$$\widehat{N}_{rc} = \sum_{j=1}^{m} w_j \, y_{(rc)j}$$

where w_j is a sampling weight. If a variable, x_j, is specified with the `tab()` option, $\widehat{N}_{rc}$ becomes

$$\widehat{N}_{rc} = \sum_{j=1}^{m} w_j \, x_j \, y_{(rc)j}$$

Let

$$\widehat{N}_{r\cdot} = \sum_{c=1}^{C} \widehat{N}_{rc}\,, \qquad \widehat{N}_{\cdot c} = \sum_{r=1}^{R} \widehat{N}_{rc}\,, \qquad \text{and} \qquad \widehat{N}_{\cdot\cdot} = \sum_{r=1}^{R}\sum_{c=1}^{C} \widehat{N}_{rc}$$

Estimated cell proportions are $\widehat{p}_{rc} = \widehat{N}_{rc}/\widehat{N}_{\cdot\cdot}$; estimated row proportions (row option) are $\widehat{p}_{\mathrm{row}\,rc} = \widehat{N}_{rc}/\widehat{N}_{r\cdot}$; estimated column proportions (column option) are $\widehat{p}_{\mathrm{col}\,rc} = \widehat{N}_{rc}/\widehat{N}_{\cdot c}$; estimated row marginals are $\widehat{p}_{r\cdot} = \widehat{N}_{r\cdot}/\widehat{N}_{\cdot\cdot}$; and estimated column marginals are $\widehat{p}_{\cdot c} = \widehat{N}_{\cdot c}/\widehat{N}_{\cdot\cdot}$.

$\widehat{N}_{rc}$ is a total, the proportion estimators are ratios, and their variances can be estimated using linearization methods as outlined in [SVY] **Variance estimation**. svy: tabulate computes the variance estimates by using svy: mean, svy: ratio, and svy: total.

Confidence intervals

Confidence intervals for proportions are calculated using a logit transform so that the endpoints lie between 0 and 1. Let $\widehat{p}$ be an estimated proportion and $\widehat{s}$ be an estimate of its standard error. Let

$$f(\widehat{p}) = \ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right)$$

be the logit transform of the proportion. In this metric, an estimate of the standard error is

$$\widehat{\mathrm{SE}}\{f(\widehat{p})\} = f'(\widehat{p})\widehat{s} = \frac{\widehat{s}}{\widehat{p}(1-\widehat{p})}$$

Thus a $100(1-\alpha)\%$ confidence interval in this metric is

$$\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm \frac{t_{1-\alpha/2,\nu}\,\widehat{s}}{\widehat{p}(1-\widehat{p})}$$

where $t_{1-\alpha/2,\nu}$ is the $(1-\alpha/2)$th quantile of Student's t distribution with ν degrees of freedom. The endpoints of this confidence interval are transformed back to the proportion metric by using the inverse of the logit transform

$$f^{-1}(y) = \frac{e^y}{1+e^y}$$

Hence, the displayed confidence intervals for proportions are

$$f^{-1}\left\{\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm \frac{t_{1-\alpha/2,\nu}\,\widehat{s}}{\widehat{p}(1-\widehat{p})}\right\}$$

Confidence intervals for weighted counts are untransformed and are identical to the intervals produced by svy: total.

The test statistics

The uncorrected Pearson χ^2 statistic is

$$X_{\mathrm{P}}^2 = m \sum_{r=1}^{R} \sum_{c=1}^{C} \left(\widehat{p}_{rc} - \widehat{p}_{0rc}\right)^2 / \widehat{p}_{0rc}$$

and the uncorrected likelihood-ratio χ^2 statistic is

$$X_{\mathrm{LR}}^2 = 2m \sum_{r=1}^{R} \sum_{c=1}^{C} \widehat{p}_{rc} \ln \left(\widehat{p}_{rc}/\widehat{p}_{0rc}\right)$$

where m is the total number of sampled individuals, $\widehat{p}_{rc}$ is the estimated proportion for the cell in the rth row and cth column of the table as defined earlier, and $\widehat{p}_{0rc}$ is the estimated proportion under the null hypothesis of independence; that is, $\widehat{p}_{0rc} = \widehat{p}_{r\cdot}\widehat{p}_{\cdot c}$, the product of the row and column marginals.

Rao and Scott (1981, 1984) show that, asymptotically, X_{P}^2 and X_{LR}^2 are distributed as

$$X^2 \sim \sum_{k=1}^{(R-1)(C-1)} \delta_k W_k \qquad (1)$$

where the W_k are independent χ_1^2 variables and the δ_k are the eigenvalues of

$$\Delta = (\widetilde{\mathbf{X}}_2' \mathbf{V}_{\mathrm{srs}} \widetilde{\mathbf{X}}_2)^{-1} (\widetilde{\mathbf{X}}_2' \mathbf{V} \widetilde{\mathbf{X}}_2) \qquad (2)$$

where $\mathbf{V}$ is the variance of the $\widehat{p}_{rc}$ under the survey design and $\mathbf{V}_{\mathrm{srs}}$ is the variance of the $\widehat{p}_{rc}$ that you would have if the design were simple random sampling; namely, $\mathbf{V}_{\mathrm{srs}}$ has diagonal elements $p_{rc}(1-p_{rc})/m$ and off-diagonal elements $-p_{rc}p_{st}/m$.

$\widetilde{\mathbf{X}}_2$ is calculated as follows. Rao and Scott do their development in a log-linear modeling context, so consider $[\,\mathbf{1}\,|\,\mathbf{X}_1\,|\,\mathbf{X}_2\,]$ as predictors for the cell counts of the $R \times C$ table in a log-linear model. The $\mathbf{X}_1$ matrix of dimension $RC \times (R+C-2)$ contains the $R-1$ "main effects" for the rows and the $C-1$ "main effects" for the columns. The $\mathbf{X}_2$ matrix of dimension $RC \times (R-1)(C-1)$ contains the row and column "interactions". Hence, fitting $[\,\mathbf{1}\,|\,\mathbf{X}_1\,|\,\mathbf{X}_2\,]$ gives the fully saturated model (that is, fits the observed values perfectly) and $[\,\mathbf{1}\,|\,\mathbf{X}_1\,]$ gives the independence model. The $\widetilde{\mathbf{X}}_2$ matrix is the projection of $\mathbf{X}_2$ onto the orthogonal complement of the space spanned by the columns of $\mathbf{X}_1$, where the orthogonality is defined with respect to $\mathbf{V}_{\mathrm{srs}}$; that is, $\widetilde{\mathbf{X}}_2' \mathbf{V}_{\mathrm{srs}} \mathbf{X}_1 = \mathbf{0}$.

See Rao and Scott (1984) for the proof justifying (1) and (2). However, even without a full understanding, you can get a feeling for Δ. It is like a ratio (although remember that it is a matrix) of two variances. The variance in the numerator involves the variance under the true survey design, and the variance in the denominator involves the variance assuming that the design was simple random sampling. The design effect DEFF for an estimated proportion (see [SVY] **estat**) is defined as

$$\mathrm{DEFF} = \frac{\widehat{V}(\widehat{p}_{rc})}{\widetilde{V}_{\mathrm{srsor}}(\widetilde{p}_{rc})}$$

Hence, Δ can be regarded as a design-effects matrix, and Rao and Scott call its eigenvalues, the δ_ks, the "generalized design effects".

Computing an estimate for Δ by using estimates for $\mathbf{V}$ and $\mathbf{V}_{\rm srs}$ is easy. Rao and Scott (1984) derive a simpler formula for $\widehat{\Delta}$:

$$\widehat{\Delta} = \left(\mathbf{C}'\mathbf{D}_{\widehat{\mathbf{p}}}^{-1}\widehat{V}_{\rm srs}\mathbf{D}_{\widehat{\mathbf{p}}}^{-1}\mathbf{C}\right)^{-1}\left(\mathbf{C}'\mathbf{D}_{\widehat{\mathbf{p}}}^{-1}\widehat{V}\mathbf{D}_{\widehat{\mathbf{p}}}^{-1}\mathbf{C}\right)$$

Here $\mathbf{C}$ is a contrast matrix that is any $RC \times (R-1)(C-1)$ full-rank matrix orthogonal to $[\mathbf{1}\,|\,\mathbf{X}_1]$; that is, $\mathbf{C}'\mathbf{1} = \mathbf{0}$ and $\mathbf{C}'\mathbf{X}_1 = \mathbf{0}$. $\mathbf{D}_{\widehat{\mathbf{p}}}$ is a diagonal matrix with the estimated proportions $\widehat{p}_{rc}$ on the diagonal. When one of the $\widehat{p}_{rc}$ is zero, the corresponding variance estimate is also zero; hence, the corresponding element for $\mathbf{D}_{\widehat{\mathbf{p}}}^{-1}$ is immaterial for computing $\widehat{\Delta}$.

Unfortunately, (1) is not practical for computing a p-value. However, you can compute simple first-order and second-order corrections based on it. A first-order correction is based on downweighting the i.i.d. statistics by the average eigenvalue of $\widehat{\Delta}$; namely, you compute

$$X^2_{\rm P}(\widehat{\delta}_{\cdot}) = X^2_{\rm P}/\widehat{\delta}_{\cdot} \qquad \text{and} \qquad X^2_{\rm LR}(\widehat{\delta}_{\cdot}) = X^2_{\rm LR}/\widehat{\delta}_{\cdot}$$

where $\widehat{\delta}_{\cdot}$ is the mean-generalized DEFF

$$\widehat{\delta}_{\cdot} = \frac{1}{(R-1)(C-1)} \sum_{k=1}^{(R-1)(C-1)} \delta_k$$

These corrected statistics are asymptotically distributed as $\chi^2_{(R-1)(C-1)}$. Thus, to first-order, you can view the i.i.d. statistics $X^2_{\rm P}$ and $X^2_{\rm LR}$ as being "too big" by a factor of $\widehat{\delta}_{\cdot}$ for true survey design.

A better second-order correction can be obtained by using the Satterthwaite approximation to the distribution of a weighted sum of χ^2_1 variables. Here the Pearson statistic becomes

$$X^2_{\rm P}(\widehat{\delta}_{\cdot}, \widehat{a}) = \frac{X^2_{\rm P}}{\widehat{\delta}_{\cdot}(\widehat{a}^2+1)} \tag{3}$$

where $\widehat{a}$ is the coefficient of variation of the eigenvalues:

$$\widehat{a}^2 = \frac{\sum \widehat{\delta}_k^2}{(R-1)(C-1)\widehat{\delta}_{\cdot}^2} - 1$$

Because $\sum \widehat{\delta}_k = \text{tr}\,\widehat{\Delta}$ and $\sum \widehat{\delta}_k^2 = \text{tr}\,\widehat{\Delta}^2$, (3) can be written in an easily computable form as

$$X^2_{\rm P}(\widehat{\delta}_{\cdot}, \widehat{a}) = \frac{\text{tr}\,\widehat{\Delta}}{\text{tr}\,\widehat{\Delta}^2}\, X^2_{\rm P}$$

These corrected statistics are asymptotically distributed as χ^2_d, with

$$d = \frac{(R-1)(C-1)}{\widehat{a}^2+1} = \frac{(\text{tr}\,\widehat{\Delta})^2}{\text{tr}\,\widehat{\Delta}^2}$$

that is, a χ^2 with, in general, noninteger degrees of freedom. The likelihood-ratio statistic $X^2_{\rm LR}$ can also be given this second-order correction in an identical manner.

Two issues remain. First, there are two possible ways to compute the variance estimate $\widehat{V}_{\rm srs}$, which is used to compute $\widehat{\Delta}$. $\mathbf{V}_{\rm srs}$ has diagonal elements $p_{rc}(1-p_{rc})/m$ and off-diagonal elements $-p_{rc}p_{st}/m$, but here p_{rc} is the true, not estimated, proportion. Hence, the question is what to use to estimate p_{rc}: the observed proportions, $\widehat{p}_{rc}$, or the proportions estimated under the null hypothesis of independence, $\widehat{p}_{0rc} = \widehat{p}_{r\cdot}\widehat{p}_{\cdot c}$? Rao and Scott (1984, 53) leave this as an open question.

Because of the question of using $\widehat{p}_{rc}$ or $\widehat{p}_{0rc}$ to compute $\widehat{V}_{\rm srs}$, `svy: tabulate` can compute both corrections. By default, when the `null` option is not specified, only the correction based on $\widehat{p}_{rc}$ is displayed. If `null` is specified, two corrected statistics and corresponding p-values are displayed, one computed using $\widehat{p}_{rc}$ and the other using $\widehat{p}_{0rc}$.

The second outstanding issue concerns the degrees of freedom resulting from the variance estimate, $\widehat{V}$, of the cell proportions under the survey design. The customary degrees of freedom for t statistics resulting from this variance estimate is $\nu = n - L$, where n is the number of PSUs in the sample and L is the number of strata.

Rao and Thomas (1989) suggest turning the corrected χ^2 statistic into an F statistic by dividing it by its degrees of freedom, $d_0 = (R-1)(C-1)$. The F statistic is then taken to have numerator degrees of freedom equal to d_0 and denominator degrees of freedom equal to νd_0. Hence, the corrected Pearson F statistic is

$$F_{\rm P} = \frac{X^2_{\rm P}}{{\rm tr}\,\widehat{\Delta}} \quad \text{with} \quad F_{\rm P} \sim F(d, \nu d) \quad \text{where} \quad d = \frac{({\rm tr}\,\widehat{\Delta})^2}{{\rm tr}\,\widehat{\Delta}^2} \quad \text{and} \quad \nu = n - L \tag{4}$$

This is the corrected statistic that `svy: tabulate` displays by default or when the `pearson` option is specified. When the `lr` option is specified, an identical correction is produced for the likelihood-ratio statistic $X^2_{\rm LR}$. When `null` is specified, (4) is also used. For the statistic labeled "D-B (null)", $\widehat{\Delta}$ is computed using $\widehat{p}_{0rc}$. For the statistic labeled "Design-based", $\widehat{\Delta}$ is computed using $\widehat{p}_{rc}$.

The Wald statistics computed by `svy: tabulate` with the `wald` and `llwald` options were developed by Koch, Freeman, and Freeman (1975). The statistic given by the `wald` option is similar to the Pearson statistic because it is based on

$$\widehat{Y}_{rc} = \widehat{N}_{rc} - \widehat{N}_{r\cdot}\widehat{N}_{\cdot c}/\widehat{N}_{\cdot\cdot}$$

where $r = 1, \dots, R-1$ and $c = 1, \dots, C-1$. The delta method can be used to estimate the variance of $\widehat{\mathbf{Y}}$ (which is $\widehat{Y}_{rc}$ stacked into a vector), and a Wald statistic can be constructed in the usual manner:

$$W = \widehat{\mathbf{Y}}'\{\mathbf{J}_{\rm N}\widehat{V}(\widehat{\mathbf{N}})\mathbf{J}_{\rm N}'\}^{-1}\widehat{\mathbf{Y}} \qquad \text{where} \qquad \mathbf{J}_{\rm N} = \partial\widehat{\mathbf{Y}}/\partial\widehat{\mathbf{N}}'$$

The statistic given by the `llwald` option is based on the log-linear model with predictors $[\,\mathbf{1}\,|\,\mathbf{X}_1\,|\,\mathbf{X}_2\,]$ that was mentioned earlier. This Wald statistic is

$$W_{\rm LL} = (\mathbf{X}_2'\ln\widehat{\mathbf{p}})'\{\mathbf{X}_2'\mathbf{J}_{\rm p}\widehat{V}(\widehat{\mathbf{p}})\mathbf{J}_{\rm p}'\mathbf{X}_2\}^{-1}(\mathbf{X}_2'\ln\widehat{\mathbf{p}})$$

where $\mathbf{J}_{\rm p}$ is the matrix of first derivatives of $\ln\widehat{\mathbf{p}}$ with respect to $\widehat{\mathbf{p}}$, which is, of course, just a matrix with $\widehat{p}_{rc}^{-1}$ on the diagonal and zero elsewhere. This log-linear Wald statistic is undefined when there is a zero cell in the table.

Unadjusted F statistics (noadjust option) are produced using

$$F_{\text{unadj}} = W/d_0 \qquad \text{with} \qquad F_{\text{unadj}} \sim F(d_0, \nu)$$

Adjusted F statistics are produced using

$$F_{\text{adj}} = (\nu - d_0 + 1)W/(\nu d_0) \qquad \text{with} \qquad F_{\text{adj}} \sim F(d_0, \nu - d_0 + 1)$$

The other svy estimators also use this adjustment procedure for F statistics. See Korn and Graubard (1990) for a justification of the procedure.

References

Fuller, W. A., W. J. Kennedy, Jr., D. Schnell, G. Sullivan, and H. J. Park. 1986. *PC CARP.* Software package. Ames, IA: Statistical Laboratory, Iowa State University.

Jann, B. 2008. Multinomial goodness-of-fit: Large-sample tests with survey design correction and exact tests for small samples. *Stata Journal* 8: 147–169.

Koch, G. G., D. H. Freeman, Jr., and J. L. Freeman. 1975. Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review* 43: 59–78. https://doi.org/10.2307/1402660.

Korn, E. L., and B. I. Graubard. 1990. Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics. *American Statistician* 44: 270–276. https://doi.org/10.2307/2684345.

McDowell, A., A. Engel, J. T. Massey, and K. Maurer. 1981. “Plan and operation of the Second National Health and Nutrition Examination Survey, 1976–1980”. In *Vital and Health Statistics*, ser. 1, no. 15. Hyattsville, MD: National Center for Health Statistics.

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: 221–230. https://doi.org/10.2307/2287815.

———. 1984. On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics* 12: 46–60. https://doi.org/10.1214/aos/1176346391.

Rao, J. N. K., and D. R. Thomas. 1989. “Chi-squared tests for contingency tables”. In *Analysis of Complex Surveys*, edited by C. J. Skinner, D. Holt, and T. M. F. Smith, 89–114. New York: Wiley.

Research Triangle Institute. 1997. *SUDAAN User’s Manual, Release 7.5.* Research Triangle Park, NC: Research Triangle Institute.

Sribney, W. M. 1998. svy7: Two-way contingency tables for survey or clustered data. *Stata Technical Bulletin* 45: 33–49. Reprinted in *Stata Technical Bulletin Reprints*, vol. 8, pp. 297–322. College Station, TX: Stata Press.

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: 630–636. https://doi.org/10.2307/2289475.

Also see

[SVY] **svy postestimation** — Postestimation tools for svy

[SVY] **svy** — The survey prefix command

[SVY] **svy: tabulate oneway** — One-way tables for survey data

[SVY] **svydescribe** — Describe survey data

[SVY] **Calibration** — Calibration for survey data

[SVY] **Direct standardization** — Direct standardization of means, proportions, and ratios

[SVY] **Poststratification** — Poststratification for survey data

[SVY] **Subpopulation estimation** — Subpopulation estimation for survey data

[SVY] **Variance estimation** — Variance estimation for survey data

[R] **tabulate twoway** — Two-way table of frequencies

[R] **test** — Test linear hypotheses after estimation

[U] 20 Estimation and postestimation commands

For suggested citations, see the FAQ on citing Stata documentation.