Title

stata.com

tabulate twoway — Two-way table of frequencies

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

Description

`tabulate` produces a two-way table of frequency counts, along with various measures of association, including the common Pearson's χ^2, the likelihood-ratio χ^2, Cramér's V, Fisher's exact test, Goodman and Kruskal's gamma, and Kendall's τ_b.

Line size is respected. That is, if you resize the Results window before running `tabulate`, the resulting two-way tabulation will take advantage of the available horizontal space. Stata for Unix(console) users can instead use the `set linesize` command to take advantage of this feature.

`tab2` produces all possible two-way tabulations of the variables specified in *varlist*.

`tabi` displays the $r \times c$ table, using the values specified; rows are separated by '\'. If no options are specified, it is as if `exact` were specified for a 2×2 table and `chi2` were specified otherwise. See **[U] 19 Immediate commands** for a general description of immediate commands. See *Tables with immediate data* below for examples using `tabi`.

See [R] **tabulate oneway** if you want a one-way table of frequencies. See [R] **table** and [R] **tabstat** if you want one-, two-, or n-way table of frequencies and a wide variety of summary statistics. See [R] **tabulate, summarize()** for a description of `tabulate` with the `summarize()` option; it produces a table (breakdowns) of means and standard deviations. `table` is better than `tabulate, summarize()`, but `tabulate, summarize()` is faster. See [R] **Epitab** for a 2×2 table with statistics of interest to epidemiologists.

Quick start

Two-way table of frequencies for `v1` and `v2`

```
tabulate v1 v2
```

Add row percentages

```
tabulate v1 v2, row
```

Frequencies only for observations where `v3` $= 1$

```
tabulate v1 v2 if v3==1
```

Weighted cell counts using frequency weights defined by `wvar`

```
tabulate v1 v2 [fweight=wvar]
```

Pearson's χ^2 test and each cell's contribution

```
tabulate v1 v2, chi2 cchi2
```

All available measures of association

```
tabulate v1 v2, all
```

All possible two-way tables for v1, v2, and v3

```
tab2 v1 v2 v3
```

Input cell frequencies and perform χ^2 test

```
tabi 30 18 38 \ 13 7 22, chi2
```

Menu

tabulate

Statistics > Summaries, tables, and tests > Frequency tables > Two-way table with measures of association

tab2

Statistics > Summaries, tables, and tests > Frequency tables > All possible two-way tables

tabi

Statistics > Summaries, tables, and tests > Frequency tables > Table calculator

Syntax

Two-way table

`tabulate` *varname*$_1$ *varname*$_2$ [*if*] [*in*] [*weight*] [, *options*]

Two-way table for all possible combinations—a convenience tool

`tab2` *varlist* [*if*] [*in*] [*weight*] [, *options*]

Immediate form of two-way tabulations

`tabi` #$_{11}$ #$_{12}$ [...] \ #$_{21}$ #$_{22}$ [...] [\ ...] [, *options*]

options	Description
Main	
`chi2`	report Pearson's χ^2
`exact`[`(#)`]	report Fisher's exact test
`gamma`	report Goodman and Kruskal's gamma
`lrchi2`	report likelihood-ratio χ^2
`taub`	report Kendall's τ_b
`V`	report Cramér's V
`cchi2`	report Pearson's χ^2 in each cell
`column`	report relative frequency within its column of each cell
`row`	report relative frequency within its row of each cell
`clrchi2`	report likelihood-ratio χ^2 in each cell
`cell`	report the relative frequency of each cell
`expected`	report expected frequency in each cell
`nofreq`	do not display frequencies
`rowsort`	list rows in order of observed frequency
`colsort`	list columns in order of observed frequency
`missing`	treat missing values like other values
`wrap`	do not wrap wide tables
[`no`]`key`	report/suppress cell contents key
`nolabel`	display numeric codes rather than value labels
`nolog`	do not display enumeration log for Fisher's exact test
* `firstonly`	show only tables that include the first variable in *varlist*
Advanced	
`matcell(`*matname*`)`	save frequencies in *matname*; programmer's option
`matrow(`*matname*`)`	save unique values of *varname*$_1$ in *matname*; programmer's option
`matcol(`*matname*`)`	save unique values of *varname*$_2$ in *matname*; programmer's option
‡ `replace`	replace current data with given cell frequencies
`all`	equivalent to specifying `chi2 lrchi2 V gamma taub`

*`firstonly` is available only for `tab2`.

‡`replace` is available only for `tabi`.

`by` is allowed with `tabulate` and `tab2`, and `collect` is allowed with `tabulate` and `tabi`; see [U] **11.1.10 Prefix commands**.

`fweights`, `aweights`, and `iweights` are allowed by `tabulate`. `fweights` are allowed by `tab2`. See [U] **11.1.6 weight**.

`all` does not appear in the dialog box.

Options

Main

`chi2` calculates and displays Pearson's χ^2 for the hypothesis that the rows and columns in a two-way table are independent. `chi2` may not be specified if `aweights` or `iweights` are specified.

`exact`[`(#)`] displays the significance calculated by Fisher's exact test and may be applied to $r \times c$ as well as to 2×2 tables. For 2×2 tables, both one- and two-sided probabilities are displayed. For $r \times c$ tables, two-sided probabilities are displayed. The optional positive integer # is a multiplier on the amount of memory that the command is permitted to consume. The default is 1. This option should not be necessary for reasonable $r \times c$ tables. If the command terminates with error 910, try `exact(2)`. The maximum row or column dimension allowed when computing Fisher's exact test is the maximum row or column dimension for `tabulate` (see [R] **Limits**).

`gamma` displays Goodman and Kruskal's gamma along with its asymptotic standard error. `gamma` is appropriate only when both variables are ordinal. `gamma` may not be specified if `aweights` or `iweights` are specified.

`lrchi2` displays the likelihood-ratio χ^2 statistic. `lrchi2` may not be specified if `aweights` or `iweights` are specified.

`taub` displays Kendall's τ_b along with its asymptotic standard error. `taub` is appropriate only when both variables are ordinal. `taub` may not be specified if `aweights` or `iweights` are specified.

`V` (note capitalization) displays Cramér's V. `V` may not be specified if `aweights` or `iweights` are specified.

`cchi2` displays each cell's contribution to Pearson's χ^2 in a two-way table.

`column` displays the relative frequency of each cell within its column in a two-way table.

`row` displays the relative frequency of each cell within its row in a two-way table.

`clrchi2` displays each cell's contribution to the likelihood-ratio χ^2 in a two-way table.

`cell` displays the relative frequency of each cell in a two-way table.

`expected` displays the expected frequency of each cell in a two-way table.

`nofreq` suppresses the printing of the frequencies.

`rowsort` and `colsort` specify that the rows and columns, respectively, be presented in order of observed frequency.

By default, rows and columns are presented in ascending order of the row and column variable. For instance, if you type `tabulate a b` and `a` takes on the values 2, 3, and 5, then the first row of the table will correspond to `a` = 2; the second row will correspond to `a` = 3; and the third row will correspond to `a` = 5.

`rowsort` specifies that the rows instead be presented in descending order of observed frequency of the values. If you type `twoway a b, rowsort`, the most frequently observed value of `a` will be listed in the first row, the second most frequently observed value of `a` in the second row, and

so on. If there are rows with equal frequencies, they will be presented in ascending order of the values of a. If a = 5 occurs with frequency 1,000 and values a = 2 and a = 3 each occur with frequency 500, the rows will be presented in the order a = 5, a = 2, and a = 3.

colsort does the same as rowsort, except with the columns and the column variable.

rowsort and colsort may be specified together.

missing requests that missing values be treated like other values in calculations of counts, percentages, and other statistics.

wrap requests that Stata take no action on wide, two-way tables to make them readable. Unless wrap is specified, wide tables are broken into pieces to enhance readability.

[no]key suppresses or forces the display of a key above two-way tables. The default is to display the key if more than one cell statistic is requested, and otherwise to omit it. key forces the display of the key. nokey suppresses its display.

nolabel causes the numeric codes to be displayed rather than the value labels.

nolog suppresses the display of the log for Fisher's exact test. Using Fisher's exact test requires counting all tables that have a probability exceeding that of the observed table given the observed row and column totals. The log counts down each stage of the network computations, starting from the number of columns and counting down to 1, displaying the number of nodes in the network at each stage. A log is not displayed for 2×2 tables.

firstonly, available only with tab2, restricts the output to only those tables that include the first variable in *varlist*. Use this option to interact one variable with a set of others.

Advanced

matcell(*matname*) saves the reported frequencies in *matname*. This option is for use by programmers.

matrow(*matname*) saves the numeric values of the $r \times 1$ row stub in *matname*. This option is for use by programmers. matrow() may not be specified if the row variable is a string.

matcol(*matname*) saves the numeric values of the $1 \times c$ column stub in *matname*. This option is for use by programmers. matcol() may not be specified if the column variable is a string.

replace indicates that the immediate data specified as arguments to the tabi command be left as the current data in place of whatever data were there.

The following option is available with tabulate but is not shown in the dialog box:

all is equivalent to specifying chi2 lrchi2 V gamma taub. Note the omission of exact. When all is specified, no may be placed in front of the other options. all noV requests all association measures except Cramér's V (and Fisher's exact). all exact requests all association measures, including Fisher's exact test. all may not be specified if aweights or iweights are specified.

Limits

Two-way tables may have a maximum of 1,200 rows and 80 columns (Stata/MP and Stata/SE) or 300 rows and 20 columns (Stata/BE). If larger tables are needed, see [R] **table**.

Remarks and examples

stata.com

Remarks are presented under the following headings:

tabulate
Measures of association
N-way tables
Weighted data
Tables with immediate data
tab2
Video examples

For each value of a specified variable (or a set of values for a pair of variables), `tabulate` reports the number of observations with that value. The number of times a value occurs is called its *frequency*.

tabulate

▷ Example 1

`tabulate` will make two-way tables if we specify two variables following the word `tabulate`. In our highway dataset, we have a variable called `rate` that divides the accident rate into three categories: below 4, 4–7, and above 7 per million vehicle miles. Let's make a table of the speed limit category and the accident-rate category:

```
. use https://www.stata-press.com/data/r17/hiway2
(Minnesota highway data, 1973)
. tabulate spdcat rate
```

Speed limit category	Accident rate per million vehicle miles			
	Below 4	4-7	Above 7	Total
40 to 50	3	5	3	11
55 to 50	19	6	1	26
Above 60	2	0	0	2
Total	24	11	4	39

The table indicates that three stretches of highway have an accident rate below 4 and a speed limit of 40 to 50 miles per hour. The table also shows the row and column sums (called the *marginals*). The number of highways with a speed limit of 40 to 50 miles per hour is 11, which is the same result we obtained in our previous one-way tabulations.

Stata can present this basic table in several ways—16, to be precise—and we will show just a few below. It might be easier to read the table if we included the row percentages. For instance, of 11 highways in the lowest speed limit category, three are also in the lowest accident-rate category. Three-elevenths amounts to some 27.3%. We can ask Stata to fill in this information for us by using the `row` option:

```
. tabulate spdcat rate, row
```

Key
frequency *row percentage*

Speed limit category	Accident rate per million vehicle miles Below 4	4-7	Above 7	Total
40 to 50	3 27.27	5 45.45	3 27.27	11 100.00
55 to 50	19 73.08	6 23.08	1 3.85	26 100.00
Above 60	2 100.00	0 0.00	0 0.00	2 100.00
Total	24 61.54	11 28.21	4 10.26	39 100.00

The number listed below each frequency is the percentage of cases that each cell represents out of its row. That is easy to remember because we see 100% listed in the "`Total`" column. The bottom row is also informative. We see that 61.54% of all the highways in our dataset fall into the lowest accident-rate category, that 28.21% are in the middle category, and that 10.26% are in the highest.

`tabulate` can calculate column percentages and cell percentages, as well. It does so when we specify the `column` or `cell` options, respectively. We can even specify them together. Below is a table that includes everything:

```
. tabulate spdcat rate, row column cell
```

Key
frequency *row percentage* *column percentage* *cell percentage*

Speed limit category	Accident rate per million vehicle miles			
	Below 4	4-7	Above 7	Total
40 to 50	3	5	3	11
	27.27	45.45	27.27	100.00
	12.50	45.45	75.00	28.21
	7.69	12.82	7.69	28.21
55 to 50	19	6	1	26
	73.08	23.08	3.85	100.00
	79.17	54.55	25.00	66.67
	48.72	15.38	2.56	66.67
Above 60	2	0	0	2
	100.00	0.00	0.00	100.00
	8.33	0.00	0.00	5.13
	5.13	0.00	0.00	5.13
Total	24	11	4	39
	61.54	28.21	10.26	100.00
	100.00	100.00	100.00	100.00
	61.54	28.21	10.26	100.00

The number at the top of each cell is the frequency count. The second number is the row percentage—they sum to 100% going across the table. The third number is the column percentage—they sum to 100% going down the table. The bottom number is the cell percentage—they sum to 100% going down all the columns and across all the rows. For instance, highways with a speed limit above 60 miles per hour and in the lowest accident rate category account for 100% of highways with a speed limit above 60 miles per hour; 8.33% of highways in the lowest accident-rate category; and 5.13% of all our data.

A fourth option, `nofreq`, tells Stata not to print the frequency counts. To construct a table consisting of only row percentages, we type

```
. tabulate spdcat rate, row nofreq
```

Speed limit category	Accident rate per million vehicle miles			
	Below 4	4-7	Above 7	Total
40 to 50	27.27	45.45	27.27	100.00
55 to 50	73.08	23.08	3.85	100.00
Above 60	100.00	0.00	0.00	100.00
Total	61.54	28.21	10.26	100.00

◁

Measures of association

▷ Example 2

`tabulate` will calculate the Pearson χ^2 test for the independence of the rows and columns if we specify the `chi2` option. Suppose that we have 1980 census data on 956 cities in the United States and wish to compare the age distribution across regions of the country. Assume that `agecat` is the median age in each city and that `region` denotes the region of the country in which the city is located.

```
. use https://www.stata-press.com/data/r17/citytemp2
(City temperature data)
. tabulate region agecat, chi2
```

Census region	Age category 19–29	30–34	35+	Total
NE	46	83	37	166
N Cntrl	162	92	30	284
South	139	68	43	250
West	160	73	23	256
Total	507	316	133	956

Pearson chi2(6) = 61.2877 Pr = 0.000

We obtain the standard two-way table and, at the bottom, a summary of the χ^2 test. Stata informs us that the χ^2 associated with this table has 6 degrees of freedom and is 61.29. The observed differences are significant.

The table is, perhaps, easier to understand if we suppress the frequencies and print just the row percentages:

```
. tabulate region agecat, row nofreq chi2
```

Census region	Age category 19–29	30–34	35+	Total
NE	27.71	50.00	22.29	100.00
N Cntrl	57.04	32.39	10.56	100.00
South	55.60	27.20	17.20	100.00
West	62.50	28.52	8.98	100.00
Total	53.03	33.05	13.91	100.00

Pearson chi2(6) = 61.2877 Pr = 0.000

◁

▷ Example 3

We have data on dose level and outcome for a set of patients and wish to evaluate the association between the two variables. We can obtain all the association measures by specifying the `all` and `exact` options:

```
. use https://www.stata-press.com/data/r17/dose
. tabulate dose function, all exact
Enumerating sample-space combinations:
stage 3:  enumerations = 1
stage 2:  enumerations = 9
stage 1:  enumerations = 0
```

Dosage	Function: < 1 hr	1 to 4	4+	Total
1/day	20	10	2	32
2/day	16	12	4	32
3/day	10	16	6	32
Total	46	38	12	96

```
          Pearson chi2(4) =   6.7780   Pr = 0.148
 Likelihood-ratio chi2(4) =   6.9844   Pr = 0.137
               Cramér's V =   0.1879
                    gamma =   0.3689  ASE = 0.129
          Kendall's tau-b =   0.2378  ASE = 0.086
           Fisher's exact =                 0.145
```

We find evidence of association but not enough to be truly convincing.

If we had not also specified the `exact` option, we would not have obtained Fisher's exact test. Stata can calculate this statistic both for 2×2 tables and for $r \times c$. For 2×2 tables, the calculation is almost instant. On more general tables, however, the calculation can take longer.

We carefully constructed our example so that `all` would be meaningful. Kendall's τ_b and Goodman and Kruskal's gamma are relevant only when both dimensions of the table can be ordered, say, from low to high or from worst to best. The other statistics, however, are always applicable.

◁

❑ Technical note

Be careful when attempting to compute the p-value for Fisher's exact test because the number of tables that contribute to the p-value can be extremely large and a solution may not be feasible. The errors that are indicative of this situation are errors 910, exceeded memory limitations, and 1401, integer overflow due to large row-margin frequencies. If execution terminates because of memory limitations, use `exact(2)` to permit the algorithm to consume twice the memory, `exact(3)` for three times the memory, etc. The default memory usage should be sufficient for reasonable tables.

❑

N-way tables

If you need more than two-way tables, your best alternative to is use `table`, not `tabulate`; see [R] **table**.

The technical note below shows you how to use `tabulate` to create a sequence of two-way tables that together form, in effect, a three-way table, but using `table` is easy and produces prettier results:

```
. use https://www.stata-press.com/data/r17/birthcat
(City data)
. table (agecat birthcat) (region), nototals
```

	Census region			
	NE	N Cntrl	South	West
Age category				
19–29				
Birth-rate category				
29–136	11	23	11	11
137–195	31	97	65	46
196–529	4	38	59	91
30–34				
Birth-rate category				
29–136	34	27	10	8
137–195	48	58	45	42
196–529	1	3	12	21
35+				
Birth-rate category				
29–136	34	26	27	18
137–195	3	4	7	4
196–529			4	

❑ Technical note

We can make n-way tables by combining the by *varlist*: prefix with `tabulate`. Continuing with the dataset of 956 cities, say that we want to make a table of age category by birth-rate category by region of the country. The birth-rate category variable is named `birthcat` in our dataset. To make separate tables for each age category, we would type

```
. by agecat, sort: tabulate birthcat region
```

```
-> agecat = 19-29
```

Birth-rate category	Census region NE	N Cntrl	South	West	Total
29–136	11	23	11	11	56
137–195	31	97	65	46	239
196–529	4	38	59	91	192
Total	46	158	135	148	487

```
-> agecat = 30-34
```

Birth-rate category	Census region NE	N Cntrl	South	West	Total
29–136	34	27	10	8	79
137–195	48	58	45	42	193
196–529	1	3	12	21	37
Total	83	88	67	71	309

```
-> agecat = 35+
Birth-rate |                Census region
  category |        NE    N Cntrl       South        West |     Total
-----------+--------------------------------------------+----------
    29-136 |        34         26          27          18 |       105
   137-195 |         3          4           7           4 |        18
   196-529 |         0          0           4           0 |         4
-----------+--------------------------------------------+----------
     Total |        37         30          38          22 |       127
```

❑

Weighted data

▷ Example 4

`tabulate` can process weighted as well as unweighted data. As with all Stata commands, we indicate the weight by specifying the [*weight*] modifier; see [U] **11.1.6 weight**.

Continuing with our dataset of 956 cities, we also have a variable called `pop`, the population of each city. We can make a table of region by age category, weighted by population, by typing

```
. tabulate region agecat [fweight=pop]
   Census |           Age category
   region |     19-29       30-34         35+       Total
----------+---------------------------------------------
       NE |   4721387    10421387     5323610    20466384
  N Cntrl |  16901550     8964756     4015593    29881899
    South |  13894254     7686531     4141863    25722648
     West |  16698276     7755255     2375118    26828649
----------+---------------------------------------------
    Total |  52215467    34827929    15856184   102899580
```

If we specify the `cell`, `column`, or `row` options, they will also be appropriately weighted. Below, we repeat the table, suppressing the counts and substituting row percentages:

```
. tabulate region agecat [fweight=pop], nofreq row
   Census |           Age category
   region |     19-29       30-34         35+       Total
----------+---------------------------------------------
       NE |     23.07       50.92       26.01      100.00
  N Cntrl |     56.56       30.00       13.44      100.00
    South |     54.02       29.88       16.10      100.00
     West |     62.24       28.91        8.85      100.00
----------+---------------------------------------------
    Total |     50.74       33.85       15.41      100.00
```

◁

Tables with immediate data

▷ Example 5

`tabi` ignores the dataset in memory and uses as the table the values that we specify on the command line:

```
. tabi 30 18 \ 38 14
```

row	col 1	2	Total
1	30	18	48
2	38	14	52
Total	68	32	100

```
          Fisher's exact =                 0.289
  1-sided Fisher's exact =                 0.179
```

We may specify any of the options of `tabulate` and are not limited to 2×2 tables:

```
. tabi 30 18 38 \ 13 7 22, chi2 exact
Enumerating sample-space combinations:
stage 3:  enumerations = 1
stage 2:  enumerations = 3
stage 1:  enumerations = 0
```

row	col 1	2	3	Total
1	30	18	38	86
2	13	7	22	42
Total	43	25	60	128

```
       Pearson chi2(2) =   0.7967   Pr = 0.671
        Fisher's exact =                 0.707
. tabi 30 13 \ 18 7 \ 38 22, all exact col
```

Key
frequency *column percentage*

```
Enumerating sample-space combinations:
stage 3:  enumerations = 1
stage 2:  enumerations = 3
stage 1:  enumerations = 0
```

row	col 1	2	Total
1	30	13	43
	34.88	30.95	33.59
2	18	7	25
	20.93	16.67	19.53
3	38	22	60
	44.19	52.38	46.88
Total	86	42	128
	100.00	100.00	100.00

```
          Pearson chi2(2) =   0.7967   Pr = 0.671
Likelihood-ratio chi2(2) =   0.7985   Pr = 0.671
                Cramér's V =   0.0789
                     gamma =   0.1204  ASE = 0.160
           Kendall's tau-b =   0.0630  ASE = 0.084
            Fisher's exact =                 0.707
```

For 2 × 2 tables, both one- and two-sided Fisher's exact probabilities are displayed; this is true of both `tabulate` and `tabi`. See *Cumulative incidence data* and *Case–control data* in [R] **Epitab** for more discussion on the relationship between one- and two-sided probabilities.

◁

❑ Technical note

`tabi`, as with all immediate commands, leaves any data in memory undisturbed. With the `replace` option, however, the data in memory are replaced by the data from the table:

```
. tabi 30 18 \ 38 14, replace
           |        col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |        30         18 |        48
         2 |        38         14 |        52
-----------+----------------------+----------
     Total |        68         32 |       100

          Fisher's exact =                 0.289
  1-sided Fisher's exact =                 0.179
. list

     +-------------------+
     | row   col   pop   |
     |-------------------|
  1. |   1     1    30   |
  2. |   1     2    18   |
  3. |   2     1    38   |
  4. |   2     2    14   |
     +-------------------+
```

With this dataset, you could re-create the above table by typing

```
. tabulate row col [fweight=pop], exact
           |        col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |        30         18 |        48
         2 |        38         14 |        52
-----------+----------------------+----------
     Total |        68         32 |       100

          Fisher's exact =                 0.289
  1-sided Fisher's exact =                 0.179
```

❑

tab2

tab2 is a convenience tool. Typing

```
. tab2 myvar thisvar thatvar, chi2
```

is equivalent to typing

```
. tabulate myvar thisvar, chi2
. tabulate myvar thatvar, chi2
. tabulate thisvar thatvar, chi2
```

Video examples

Pearson's chi-squared and Fisher's exact test in Stata

Tables and cross-tabulations in Stata

Cross-tabulations and chi-squared tests calculator

Stored results

tabulate, tab2, and tabi store the following in r():

Scalars

r(N)	number of observations	r(p_exact)	Fisher's exact p
r(r)	number of rows	r(chi2_lr)	likelihood-ratio χ^2
r(c)	number of columns	r(p_lr)	p-value for likelihood-ratio test
r(chi2)	Pearson's χ^2 test	r(CramersV)	Cramér's V
r(p)	p-value for of Pearson's χ^2 test	r(ase_gam)	ASE of gamma
r(gamma)	gamma	r(ase_taub)	ASE of $\tau_{\rm b}$
r(p1_exact)	one-sided Fisher's exact p	r(taub)	$\tau_{\rm b}$

r(p1_exact) is defined only for 2×2 tables. Also, the matrow(), matcol(), and matcell() options allow you to obtain the row values, column values, and frequencies, respectively.

Methods and formulas

Let n_{ij}, $i = 1, \ldots, I$ and $j = 1, \ldots, J$, be the number of observations in the ith row and jth column. If the data are not weighted, n_{ij} is just a count. If the data are weighted, n_{ij} is the sum of the weights of all data corresponding to the (i, j) cell.

Define the row and column marginals as

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij} \qquad\qquad n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$$

and let $n = \sum_i \sum_j n_{ij}$ be the overall sum. Also, define the concordance and discordance as

$$A_{ij} = \sum_{k>i}\sum_{l>j} n_{kl} + \sum_{k<i}\sum_{l<j} n_{kl} \qquad\qquad D_{ij} = \sum_{k>i}\sum_{l<j} n_{kl} + \sum_{k<i}\sum_{l>j} n_{kl}$$

along with twice the number of concordances $P = \sum_i \sum_j n_{ij} A_{ij}$ and twice the number of discordances $Q = \sum_i \sum_j n_{ij} D_{ij}$.

The Pearson χ^2 statistic with $(I-1)(J-1)$ degrees of freedom (so called because it is based on Pearson (1900); see Conover [1999, 240] and Fienberg [1980, 9]) is defined as

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

where $m_{ij} = n_{i\cdot}n_{\cdot j}/n$.

The likelihood-ratio χ^2 statistic with $(I-1)(J-1)$ degrees of freedom (Fienberg 1980, 40) is defined as

$$G^2 = 2\sum_i \sum_j n_{ij}\ln(n_{ij}/m_{ij})$$

Cramér's V (Cramér 1946) is a measure of association designed so that the attainable upper bound is 1. For 2×2 tables, $-1 \leq V \leq 1$, and otherwise, $0 \leq V \leq 1$.

$$V = \begin{cases} (n_{11}n_{22} - n_{12}n_{21})/(n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2})^{1/2} & \text{for } 2 \times 2 \\ \left\{(X^2/n)/\min(I-1, J-1)\right\}^{1/2} & \text{otherwise} \end{cases}$$

Gamma (Goodman and Kruskal 1954, 1959, 1963, 1972; also see Agresti [2010,186–188]) ignores tied pairs and is based only on the number of concordant and discordant pairs of observations, $-1 \leq \gamma \leq 1$,

$$\gamma = (P - Q)/(P + Q)$$

with asymptotic variance

$$16\sum_i \sum_j n_{ij}(QA_{ij} - PD_{ij})^2/(P+Q)^4$$

Kendall's τ_b (Kendall 1945; also see Agresti 2010, 188–189), $-1 \leq \tau_b \leq 1$, is similar to gamma, except that it uses a correction for ties,

$$\tau_b = (P - Q)/(w_r w_c)^{1/2}$$

with asymptotic variance

$$\frac{\sum_i \sum_j n_{ij}(2w_r w_c d_{ij} + \tau_b v_{ij})^2 - n^3\tau_b^2(w_r + w_c)^2}{(w_r w_c)^4}$$

where

$$w_r = n^2 - \sum_i n_{i.}^2$$
$$w_c = n^2 - \sum_j n_{.j}^2$$
$$d_{ij} = A_{ij} - D_{ij}$$
$$v_{ij} = n_{i.}w_c + n_{.j}w_r$$

Fisher's exact test (Fisher 1935; Finney 1948; see Zelterman and Louis [1992, 293–301] for the 2×2 case) yields the probability of observing a table that gives at least as much evidence of association as the one actually observed under the assumption of no association. Holding row and column marginals fixed, the hypergeometric probability P of every possible table A is computed, and the

$$P = \sum_{T \in A} \Pr(T)$$

where A is the set of all tables with the same marginals as the observed table, T^*, such that $\Pr(T) \leq \Pr(T^*)$. For 2×2 tables, the one-sided probability is calculated by further restricting A to tables in the same tail as T^*. The first algorithm extending this calculation to $r \times c$ tables was Pagano and Halvorsen (1981); the one implemented here is the FEXACT algorithm by Mehta and Patel (1986). This is a search-tree clipping method originally published by Mehta and Patel (1983) with further refinements by Joe (1988) and Clarkson, Fan, and Joe (1993). Fisher's exact test is a permutation test. For more information on permutation tests, see Good (2005 and 2006) and Pesarin (2001).

References

Agresti, A. 2010. *Analysis of Ordinal Categorical Data*. 2nd ed. Hoboken, NJ: Wiley.

Campbell, M. J., D. Machin, and S. J. Walters. 2007. *Medical Statistics: A Textbook for the Health Sciences*. 4th ed. Chichester, UK: Wiley.

Clarkson, D. B., Y.-A. Fan, and H. Joe. 1993. A remark on Algorithm 643: FEXACT: An algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *ACM Transactions on Mathematical Software* 19: 484–488. https://doi.org/10.1145/168173.168412.

Conover, W. J. 1999. *Practical Nonparametric Statistics*. 3rd ed. New York: Wiley.

Cox, N. J. 2009. Speaking Stata: I. J. Good and quasi-Bayes smoothing of categorical frequencies. *Stata Journal* 9: 306–314.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Donath, S. 2018. baselinetable: A command for creating one- and two-way tables of summary statistics. *Stata Journal* 18: 327–344.

Fienberg, S. E. 1980. *The Analysis of Cross-Classified Categorical Data*. 2nd ed. Cambridge, MA: MIT Press.

Finney, D. J. 1948. The Fisher–Yates test of significance in 2×2 contingency tables. *Biometrika* 35: 145–156. https://doi.org/10.2307/2332635.

Fisher, R. A. 1935. The logic of inductive inference. *Journal of the Royal Statistical Society* 98: 39–82. https://doi.org/10.2307/2342435.

Good, P. I. 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses: A Practical Guide to Resampling Methods for Testing Hypotheses*. 3rd ed. New York: Springer.

——. 2006. *Resampling Methods: A Practical Guide to Data Analysis*. 3rd ed. Boston: Birkhäuser.

Goodman, L. A., and W. H. Kruskal. 1954. Measures of association for cross classifications. *Journal of the American Statistical Association* 49: 732–764. https://doi.org/10.1080/01621459.1954.10501231.

——. 1959. Measures of association for cross classifications II: Further discussion and references. *Journal of the American Statistical Association* 54: 123–163. https://doi.org/10.1080/01621459.1959.10501503.

——. 1963. Measures of association for cross classifications III: Approximate sampling theory. *Journal of the American Statistical Association* 58: 310–364. https://doi.org/10.2307/2283271.

——. 1972. Measures of association for cross classifications IV: Simplification of asymptotic variances. *Journal of the American Statistical Association* 67: 415–421. https://doi.org/10.2307/2284396.

Harrison, D. A. 2006. Stata tip 34: Tabulation by listing. *Stata Journal* 6: 425–427.

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.

Joe, H. 1988. Extreme probabilities for contingency tables under row and column independence with application to Fisher's exact test. *Communications in Statistics—Theory and Methods* 17: 3677–3685. https://doi.org/10.1080/03610928808829827.

Kendall, M. G. 1945. The treatment of ties in rank problems. *Biometrika* 33: 239–251. https://doi.org/10.2307/2332303.

Longest, K. C. 2014. *Using Stata for Quantitative Analysis*. 2nd ed. Thousand Oaks, CA: SAGE.

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: 427–434. https://doi.org/10.1080/01621459.1983.10477989.

——. 1986. Algorithm 643 FEXACT: A FORTRAN subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Transactions on Mathematical Software* 12: 154–161. https://doi.org/10.1145/6497.214326.

Newson, R. B. 2002. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal* 2: 45–64.

Pagano, M., and K. T. Halvorsen. 1981. An algorithm for finding the exact significance levels of $r \times c$ contingency tables. *Journal of the American Statistical Association* 76: 931–934. https://doi.org/10.2307/2287590.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, Series 5 50: 157–175. https://doi.org/10.1080/14786440009463897.

Pesarin, F. 2001. *Multivariate Permutation Tests: With Applications in Biostatistics*. Chichester, UK: Wiley.

Zelterman, D., and T. A. Louis. 1992. Contingency tables in medical studies. In *Medical Uses of Statistics*, 2nd ed, ed. J. C. Bailar III and C. F. Mosteller, 293–310. Boston: Dekker.

Also see

[R] **Epitab** — Tables for epidemiologists

[R] **table** — Table of frequencies, summaries, and command results

[R] **table twoway** — Two-way tabulation

[R] **tabstat** — Compact table of summary statistics

[R] **tabulate oneway** — One-way table of frequencies

[R] **tabulate, summarize()** — One- and two-way tables of summary statistics

[D] **collapse** — Make dataset of summary statistics

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

[SVY] **svy: tabulate twoway** — Two-way tables for survey data

[XT] **xttab** — Tabulate xt data

[U] **12.6.3 Value labels**

[U] **26 Working with categorical data and factor variables**