The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
1-year subscription	$ 79	1-year subscription	$115
2-year subscription	$155	2-year subscription	$225
3-year subscription	$225	3-year subscription	$329
3-year subscription (electronic only)	$210	3-year subscription (electronic only)	$210
1-year student subscription	$ 48	1-year student subscription	$ 79
1-year university library subscription	$ 99	1-year university library subscription	$135
2-year university library subscription	$195	2-year university library subscription	$265
3-year university library subscription	$289	3-year university library subscription	$395
1-year institutional subscription	$225	1-year institutional subscription	$259
2-year institutional subscription	$445	2-year institutional subscription	$510
3-year institutional subscription	$650	3-year institutional subscription	$750

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2012)
12, Number 4, pp. 688–701

A command to calculate age-standardized rates with efficient interval estimation

Dario Consonni
Epidemiology Unit
Fondazione IRCCS Ca' Granda—Ospedale Maggiore Policlinico
Milan, Italy
dario.consonni@unimi.it

Enzo Coviello
Statistics and Epidemiology Unit, ASL BT
Barletta, Italy
enzo.coviello@tin.it

Carlotta Buzzoni
Clinical and Descriptive Epidemiology Unit, ISPO
Firenze, Italy
c.buzzoni@ispo.toscana.it

Carolina Mensi
Department of Preventive Medicine
Fondazione IRCCS Ca' Granda—Ospedale Maggiore Policlinico
and
Lombardy Mesothelioma and Sinonasal Cancer Registry
Milan, Italy
carolina.mensi@unimi.it

Abstract. In this article, we illustrate the command `distrate`, which calculates age-standardized rates with efficient interval estimation by using formulas developed by Tiwari, Clegg, and Zou (2006, *Statistical Methods in Medical Research* 15: 547–569) as a modification of the method proposed by Fay and Feuer (1997, *Statistics in Medicine* 16: 791–801). This method is currently used in the Surveillance, Epidemiology, and End Results Program of the National Cancer Institute in Bethesda, Maryland; the Italian Association of Cancer Registries (Associazione Italiana Registro Tumori, AIRTUM); and the Lombardy Mesothelioma and Sinonasal Cancer Registry in Northern Italy. The command produces a compact output and allows for the possibility of specifying a rate multiplier, for instance, ×100,000 or ×1,000,000. Furthermore, rates and confidence limits can be easily exported to an external dataset for further processing (for example, for making graphs). The command `distrate` is a useful addition to the official Stata command `dstdize`.

Keywords: st0276, distrate, confounding, standardization, incidence rates, mortality rates, confidence intervals

1 Introduction

In observational epidemiology, confounding is a major threat to study validity. Several methods are available to adjust for confounders, including standardization, stratification, and regression modeling (Rothman, Greenland, and Lash 2008). In principle, standardization can be used for any measure: rates, proportions (risks, prevalence rates), or odds. In practice, standardization is mostly used to adjust rates for age differences across populations or exposure groups. The process of age standardization involves the calculation of a weighted average of stratum-specific (that is, age-specific) rates r_j. The weights w_j can be persons in a defined period (that is, person-years) or fractions summing to 1:

$$\mathrm{SR} = \frac{\sum_j w_j r_j}{\sum_j w_j}$$

(The denominator of the formula above is omitted if the weights represent fractions summing to 1.)

In theory, standardization is a unique concept (Miettinen 1972a,b, 1985, 2011; Rothman 1986, 2002; Rothman, Greenland, and Lash 2008); in practice, epidemiologists usually distinguish two types of standardization known as direct standardization and indirect standardization.

1. Directly standardized rates (DSRs) are calculated as weighted averages of the stratum-specific rates r_j for the k groups or populations of interest, taking the weights (person-years Y_j) from a common standard distribution:

   $$\mathrm{DSR}_k = \frac{\sum_{j_k} Y_j r_{j_k}}{\sum_j Y_j}$$

   The (arbitrary) choice of the standard population (region, country, continent, world) depends on the aims of the study. For worldwide comparison of mortality and cancer incidence, Segi's world standard population (18 age groups) is usually employed (Curado et al. 2007). The process could continue by dividing each SR_k by the crude rate in the standard population R to estimate standardized rate ratios (SRRs):

   $$\mathrm{SRR}_k = \frac{\mathrm{SR}_k}{R}$$

   However, SRRs are rarely employed in practice (rate ratios estimated with Poisson regression are most often used), and the process usually stops by calculating standardized rates.

2. In calculating an indirectly standardized rate (ISR), the weights (person-years) are taken from the k groups or populations of interest y_{j_k}, while the age-specific rates R_j come from a unique external population (most commonly, the whole country or a region):

$$\text{ISR}_k = \frac{\sum_{j_k} y_{j_k} R_j}{\sum_{j_k} y_{j_k}}$$

The process is then further carried on by taking the ratio between the crude rates r_k in each group or population and the ISR_k (equivalent to the ratio of the observed and expected number of deaths or diseased cases in each population) to calculate the standardized mortality or morbidity ratio (SMR):

$$\text{SMR}_k = \frac{r_k}{\text{ISR}_k} = \frac{\text{Observed}_k}{\text{Expected}_k}$$

This method is mostly used in the analysis of occupational cohorts (Checkoway, Pearce, and Kriebel 2004) and in small-area geographical studies. A largely used synonym for SMR is standardized incidence ratio.

The key distinction between the two approaches lies in the fact that the first method employs a common set of weights for all the index groups or populations; therefore, rates are mutually standardized and can be safely compared (Miettinen 1972b). In the second form, weights are usually different, so rates are not mutually standardized, and there may be issues of noncomparability between SMRs calculated for different groups or populations (Miettinen 1972b; Breslow and Day 1987; Clayton and Hills 1993; Rothman 1986, 2002; Checkoway, Pearce, and Kriebel 2004; Rothman, Greenland, and Lash 2008). Notwithstanding this potential problem, the SMR is still widely used in occupational and small-area epidemiology because of better statistical properties in case of sparse numbers (Breslow and Day 1987) and because it is the fundamental causal component of a crude risk or rate ratio (Miettinen 1972a).

Miettinen (1972a,b) underlined the uniqueness of the standardization process. For this reason, he used the terms "direct" and "indirect" between quotes and noted that "[SMR]... should be regarded as the ratio of DSR for the exposed and nonexposed ... with the exposed group as the standard ... " (Miettinen 1972a). In his 1985 book, he wrote more explicitly: "There are those who believe that there are two types of mutually standardized rate pairs or rate sets, 'directly' and 'indirectly' standardized. This is a misapprehension. As noted, this issue is singular, modification of weights, and the role of the 'standard' is to supply those weights." In his recent book, he reminded readers: "The misunderstanding in this has been exposed long ago but it persists ... " (Miettinen 2011).

Rothman (1986, 2002) also remarked that the terms "direct" and "indirect" are misnomers; for this reason, Rothman, Greenland, and Lash (2008) and Checkoway, Pearce, and Kriebel (2004) expressly avoid them. However, the two terms have been and are widely used by epidemiologists and can be found in popular epidemiology books (Breslow and Day 1987; Clayton and Hills 1993).

In this article, we focus on (direct) standardization, in which a common set of weights is used for standardizing rates in several groups or populations of interest. Different large-sample approximate formulas are available for calculating confidence intervals of DSRs when numbers are large. The fundamental publication *Cancer incidence in five continents* (CI5) has become the recognized reference source on the incidence of cancers in populations around the world. The last edition (volume IX) reports verified, good-quality data for 300 populations in 225 cancer registries across 60 countries (Curado et al. 2007). The formula used in CI5 for the variance of the SR_k, credited to Keyfitz (1966), is a weighted average of the age-specific rate variances assuming that each age-specific rate r_{jk} is binomially distributed with $\mathrm{Var}(r_{jk}) = r_{jk}(1-r_{jk})/y_{jk}$:

$$\mathrm{Var}(\mathrm{SR}_k) = \frac{\sum_{j_k} \left\{ \frac{d_{jk}(y_{jk}-d_{jk})w_j^2}{y_{jk}^3} \right\}}{\left(\sum_j w_j \right)^2} \tag{1}$$

where j indicates the age groups, d_{jk} and y_{jk} are the number of age-specific cases and person-years in each kth cancer registry, and w_j are the weights (in this case, the person-years in the 18 age groups of Segi's world standard population). The 95% confidence limits of each SR_k are then calculated based on the normal assumption

$$\mathrm{CL}_{95\%} = \mathrm{SR}_k \pm 1.96\sqrt{\mathrm{Var}(\mathrm{SR}_k)}$$

The official Stata command `dstdize` uses the algebraically equivalent Cochran's (1977) formula.

A slightly different formula—in which a weighted average of the age-specific rate variances is calculated based on the assumption that each stratum-specific rate has a Poisson distribution with variance $\mathrm{Var}(r_{jk}) = d_{jk}/y_{jk}^2$—is illustrated by Rothman, Greenland, and Lash (2008):

$$\mathrm{Var}(\mathrm{SR}_k) = \frac{\sum_{j_k} \left(\frac{d_{jk}}{y_{jk}^2} w_j^2 \right)}{\left(\sum_j w_j \right)^2} \tag{2}$$

This expression is algebraically equivalent to those (using proportional weights summing to 1) reported in Breslow and Day (1987) and Clayton and Hills (1993). Equations (1) and (2) give very similar results when $d_{jk} << y_{jk}$.

Dobson et al. (1991) proposed formulas based on the χ^2 distribution; these formulas do not require large cell counts as do the formulas above. Later, Fay and Feuer (1997) developed more conservative confidence limits assuming that a mixture of Poisson distributions is approximately distributed as a gamma distribution. More recently, Tiwari, Clegg, and Zou (2006) proposed modified gamma intervals and showed that they are more efficient: they have empirical coverage probabilities less than or equal to those of Fay and Feuer (1997), and they also retain the nominal level. The lower $L(\mathrm{SR}_k)$ and upper $U(\mathrm{SR}_k)$ confidence limits are defined as

$$L(\mathrm{SR}_k, \alpha) = \frac{\nu_k}{2\mathrm{SR}_k} \left(\chi^2_{2\mathrm{SR}^2_k/\nu_k}\right)^{-1} \left(\frac{\alpha}{2}\right); U(\mathrm{SR}_k, \alpha) = \frac{\widetilde{\nu}_k}{2\widetilde{\mathrm{SR}}_k} \left(\chi^2_{2\widetilde{\mathrm{SR}}^2_k/\widetilde{\nu}_k}\right)^{-1} \left(1 - \frac{\alpha}{2}\right)$$

where $\nu_k = \sum_{j=1}^{j}\{(d_{jk}w_j^2)/(y_{jk}^2)\}$, $\widetilde{\mathrm{SR}}_k = \sum_{j=1}^{j}[\{(d_{jk} + 1/J)w_j\}/(y_{jk})]$, $\widetilde{\nu}_k = \sum_{j=1}^{j}[\{(d_{jk} + 1/J)w_j^2\}/(y_{jk}^2)]$, and $(\chi_l^2)^{-1}(\alpha)$ correspond to the 100αth percentile of the chi-squared distribution with l degrees of freedom (formulas slightly modified from Tiwari, Clegg, and Zou [2006]). When $\mathrm{SR}_k = 0$, note that $L(\mathrm{SR}_k)$ is not defined and is set to 0. The methods of Fay and Feuer (1997) and Tiwari, Clegg, and Zou (2006) are currently used in the Surveillance, Epidemiology, and End Results (SEER) Program[1] of the National Cancer Institute in Bethesda, Maryland. The formulas of Tiwari, Clegg, and Zou (2006) are used by the Italian Association of Cancer Registries (Associazione Italiana Registri Tumori, AIRTUM)[2] and the Lombardy Mesothelioma and Sinonasal Cancer Registry.

As noted above, the official Stata command `dstdize` implements a widely used formula. Several options are available in choosing the standard population, internal or external to the study dataset. `dstdize` produces a long output that favors detailed examinations of the standardization process but can be somewhat cumbersome to read when analyzing several populations. Furthermore, exporting the estimated rates is not easy (it involves some matrix manipulation).

In this article, we describe the command `distrate` (written by Enzo Coviello), which implements the formulas of Tiwari, Clegg, and Zou (2006) for the confidence interval of standardized rates, which might be preferable in the case of rare diseases. Useful characteristics of the program are the compact output, the possibility of specifying the desired multiplier for rates (for instance, 100,000 or 1,000,000), and the easy output of rates and confidence limits to an external file for further processing. We illustrate its performance on data from the Lombardy Mesothelioma and Sinonasal Cancer Registry in Northern Italy.

2 The distrate command

In theory, `distrate` can be run to analyze individual data. In practice, data for age-standardization are most often organized in an aggregate form, with each record containing the age category, other relevant covariates, the number of events (diseases or deaths), and an appropriate denominator (population-time).

1. http://seer.cancer.gov/
2. http://www.registri-tumori.it/cms/

2.1 Syntax

`distrate` *casevar popvar* `using` *filename* [*if*] [*in*] `,` `standstrata(`*stratavars*`)`
[`by(`*varlist*`)` `popstand(`*varname*`)` `list(`*varlist*`)` `sepby(`*varlist*`)` `format(%`*fmt*`)`
`formatn(`*#*`)` `mult(`*#*`)` `level(`*#*`)` `dobson` `saving(`*filename*[`, replace`]`)`
`prefix(`*string*`)` `postfix(`*string*`)`]

casevar specifies a variable containing the rate numerator (number of cases of death or disease).

popvar specifies a variable containing the denominator (number of person-years over the study period).

`using` *filename* specifies a Stata dataset containing the standard population providing the common set of weights for standardization.

2.2 Options

`standstrata(`*stratavars*`)` specifies the variables defining strata across which to average stratum-specific rates. These variables must be present in the study population and in the standard population file. This is most often a unique variable containing age categories. `standstrata()` is required.

`by(`*varlist*`)` produces DSRs (that is, sharing a common standard) for each group identified by equal values of the `by()` variables taking on integer or string values.

`popstand(`*varname*`)` specifies the variable in the using file that contains the standard population weights. If `popstand()` is not specified, `distrate` assumes that it is named as *popvar* in the study population.

`list(`*varlist*`)` specifies the variables to be listed. (To list the population variable, use `N`.)

`sepby(`*varlist*`)` draws a separator line whenever *varlist* values change.

`format(%`*fmt*`)` specifies the format for variables containing the rate estimates.

`formatn(`*#*`)` specifies the number of digits for the format of the population variable `N`.

`mult(`*#*`)` specifies the units to be used in reported results. For example, if the analysis time is in years, specifying `mult(1000)` results in rates per 1,000 person-years.

`level(`*#*`)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`.

`dobson` specifies to also display the confidence limits of Dobson et al. (1991).

`saving(`*filename*[`, replace`]`)` allows for saving the estimates in a file.

`prefix(`*string*`)` or `postfix(`*string*`)` adds *string* as a prefix or a suffix to the names of variables containing rates and confidence limits when the estimates are saved.

2.3 Saved results

`distrate` saves the following in `r()`:

Scalars

`r(k)`	number of groups identified by distinct values of the `by()` variables

Matrices

`r(Nobs)`	$1 \times k$ vector of study population
`r(NDeath)`	$1 \times k$ vector of number of events
`r(crude)`	$1 \times k$ vector of crude rates
`r(adj)`	$1 \times k$ vector of adjusted rates
`r(lb_G)`	$1 \times k$ vector of lower bound of Tiwari adjusted rates
`r(ub_G)`	$1 \times k$ vector of upper bound of Tiwari adjusted rates
`r(se_gam)`	$1 \times k$ vector of standard error of adjusted rates
`r(lb_D)`	$1 \times k$ vector of lower bound of Dobson adjusted rates
`r(ub_D)`	$1 \times k$ vector of upper bound of Dobson adjusted rates

3 Example

Sinonasal cancers are rare tumors. Recognized causes include wood and leather dusts and nickel compounds. In Italy in recent years, a nationwide network of regional registries[3] (ReNaTuNS) has been established and merged with the national registry of mesotheliomas[4] (ReNaM), with the aims of monitoring incidence and mortality and providing legal assistance to the affected workers. `sinonasal.dta` contains data extracted from the Lombardy Mesothelioma and Sinonasal Cancer Registry in Northern Italy, established in 2008. All newly diagnosed (incident) sinonasal cancer cases in 2008–2009 were subdivided in the 16 regional local health units (ASLs) and by gender and age (18 categories) for a total of 576 records. The resident population had previously been multiplied by 2 (years) to obtain person-years (the variable `pop`).

```
. use sinonasal.dta, clear
. sort asl_code age sex
. list in 1/10, separator(0)
```

	asl_code	sex	age_grp	cases	pop
1.	BG	M	00-04	0	50162
2.	BG	F	00-04	0	47998
3.	BG	M	05-09	0	47778
4.	BG	F	05-09	0	45682
5.	BG	M	10-14	0	47980
6.	BG	F	10-14	0	45044
7.	BG	M	15-19	0	51624
8.	BG	F	15-19	0	48806
9.	BG	M	20-24	0	60336
10.	BG	F	20-24	0	58220

3. http://www.ispesl.it/dml/leo/Renatuns.asp
4. http://www.ispesl.it/renam/Index.asp

We first calculate age-standardized rates (×100,000) for the whole region by gender using Segi's world population (18 age categories) as the standard, contained in `world_pop.dta` and reproduced in table 1.

Table 1. Standard world (Segi's) population

Age (years)	Population
00–04	12,000
05–09	10,000
10–14	9,000
15–19	9,000
20–24	8,000
25–29	8,000
30–34	6,000
35–39	6,000
40–44	6,000
45–49	6,000
50–54	5,000
55–59	4,000
60–64	4,000
65–69	3,000
70–74	2,000
75–79	1,000
80–84	500
85+	500

```
. *Standardized rates by gender
. distrate cases pop using world_pop.dta, standstrata(age_grp) popstand(pop)
> by(sex) mult(100000) format(%8.1f) formatn(7)
Directly standardized rates (per 100000)
CI based on the gamma distribution (Fay and Feuer, 1997. Tiwari and al., 2006)
```

sex	cases	N	crude	rateadj	lb_gam	ub_gam	se_gam
M	79	8866488	0.9	0.5	0.4	0.7	0.1
F	45	9376798	0.5	0.2	0.2	0.3	0.0

The default output includes the `by()` variable, the number of cases and person-years (`N`), the crude rate, the standardized rate (`rateadj`), the lower and upper confidence bounds of the standardized rate (`lb_gam` and `ub_gam`), and its standard error. We then calculate the age-standardized rates (×100,000) by ASL and output results in an external Stata dataset.

```
. *Standardized rates by gender and ASL
. distrate cases pop using world_pop.dta, standstrata(age_grp) popstand(pop)
> by(sex asl_code) sepby(sex) mult(100000) format(%8.1f) formatn(7) prefix(SN)
> saving(sinonasal_rates.dta, replace)
Directly standardized rates (per 100000)
CI based on the gamma distribution (Fay and Feuer, 1997. Tiwari and al., 2006)
```

sex	asl_code	cases	N	crude	rateadj	lb_gam	ub_gam	se_gam
M	BG	7	960586	0.7	0.5	0.2	1.1	0.2
M	BS	10	1000536	1.0	0.7	0.3	1.3	0.2
M	CO	4	527920	0.8	0.5	0.1	1.4	0.3
M	CR	2	325702	0.6	0.3	0.0	1.7	0.4
M	LC	3	304622	1.0	0.6	0.1	2.3	0.5
M	LO	1	200010	0.5	0.3	0.0	2.5	0.7
M	MB	8	729144	1.1	0.6	0.3	1.4	0.3
M	MI	3	264232	1.1	0.7	0.1	2.7	0.7
M	MI1	7	857956	0.8	0.5	0.2	1.2	0.2
M	MI2	5	564436	0.9	0.6	0.2	1.5	0.3
M	MIC	5	1224050	0.4	0.2	0.1	0.6	0.1
M	MN	2	364332	0.5	0.2	0.0	1.5	0.4
M	PV	9	479708	1.9	0.8	0.4	1.9	0.4
M	SO	3	173724	1.7	0.9	0.2	3.6	0.8
M	VA	10	795066	1.3	0.7	0.3	1.4	0.3
M	VS	0	94464	0.0	0.0	0.0	4.5	1.2
F	BG	2	988190	0.2	0.1	0.0	0.5	0.1
F	BS	7	1032316	0.7	0.4	0.1	1.0	0.2
F	CO	2	557378	0.4	0.1	0.0	0.9	0.2
F	CR	1	345698	0.3	0.2	0.0	1.6	0.4
F	LC	2	318726	0.6	0.5	0.1	2.1	0.5
F	LO	0	209174	0.0	0.0	0.0	2.2	0.6
F	MB	3	764296	0.4	0.2	0.0	0.9	0.2
F	MI	1	277886	0.4	0.3	0.0	2.1	0.6
F	MI1	3	891514	0.3	0.2	0.0	0.7	0.2
F	MI2	0	580618	0.0	0.0	0.0	0.8	0.2
F	MIC	14	1379052	1.0	0.4	0.2	0.8	0.2
F	MN	1	388036	0.3	0.1	0.0	1.4	0.4
F	PV	3	518686	0.6	0.2	0.0	1.2	0.3
F	SO	4	181432	2.2	0.9	0.2	3.6	0.9
F	VA	2	845856	0.2	0.1	0.0	0.7	0.2
F	VS	0	97940	0.0	0.0	0.0	4.6	1.3

```
file sinonasal_rates.dta saved
```

The output dataset can be used for further processing, for instance, to produce graphs of rates by gender and ASL (figures 1 and 2). Note that the rates and confidence bounds have been prefixed by `SN`.

```
. use sinonasal_rates.dta
(Directly Standardized Rates (per 100000))
. set scheme sj
. twoway (rcap SNlb_gam SNub_gam asl_code if sex == 1, lcolor(black)
> lwidth(medthick))
> (scatter SNrateadj asl_code if sex == 1, mcolor(black) msize(medium)),
> title("Lombardy Sinonasal Cancer Registry, 2008-09") subtitle("Men")
> xtitle("ASL") ytitle("Rate X 100,000 (95% CI)")
> xlabel(0.5 " " 1 "BG" 2 "BS" 3 "CO" 4 "CR" 5 "LC" 6 "LO" 7 "MB"
> 8 "MI" 9 "MI1" 10 "MI2" 11 "MIC" 12 "MN" 13 "PV" 14 "SO" 15 "VA" 16 "VS"
> 16.5 " ", labsize(2)) xtick(1(1)12)
> yline(0.5, lpattern(dash) lcolor(black) lwidth(thin))
> ylabel(0(1)5, labsize(2) grid)
> caption("Standard: World (Segi´s) population  ---
> Lombardy Region standardized rate")
> note("Confidence intervals calculated with the Tiwari et al. (2006) formula")
> legend(off)
```

(*output omitted*)

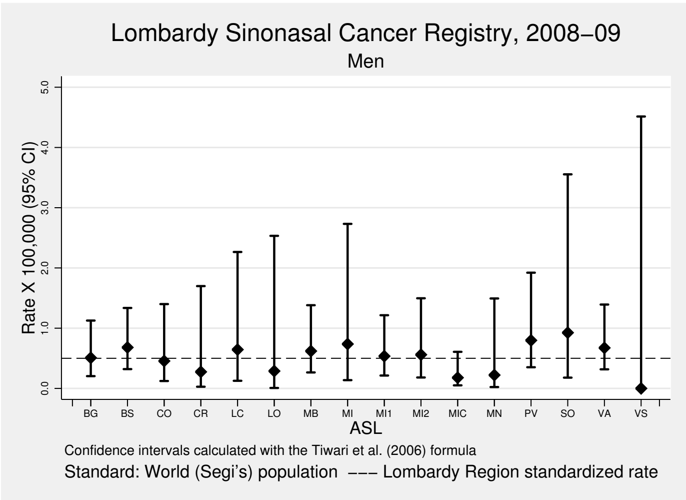

Figure 1. Graph of standardized rates by ASL, men

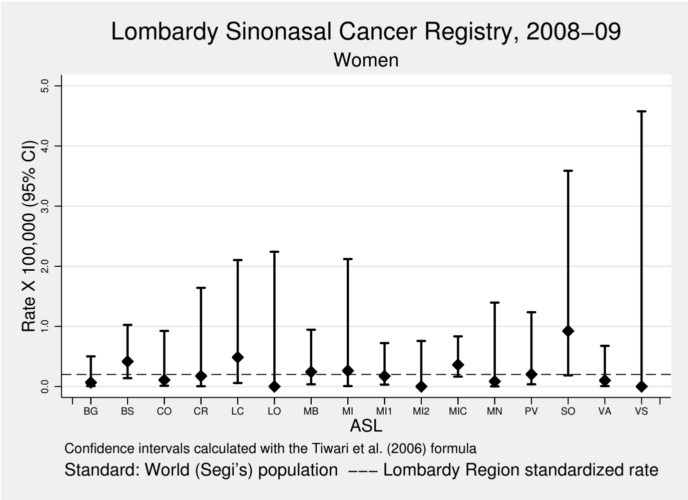

Figure 2. Graph of standardized rates by ASL, women

We then calculate standardized rates by gender and ASL using the Cochran (1977) formula implemented by `dstdize` (equivalent to the Keyfitz [1966] formula).

```
. use sinonasal.dta, clear
. dstdize cases pop age_grp, by(sex asl_code) using(world_pop.dta)
  (output omitted)
```

The lower confidence limits obtained with the Tiwari, Clegg, and Zou (2006) formula (`distrate`) are larger than (or at most equal to) those obtained with the Cochran (1977) formula (`dstdize`) (table 2). Also, the upper bounds calculated with `distrate` are usually larger, and the discrepancy increases when the number of cases is very low. When there are no cases, `distrate` calculates an upper bound while `dstdize` does not.

Table 2. Comparison of confidence limits for age-standardized rates obtained with `distrate` (Tiwari) and `dstdize` (Cochran)

			Tiwari		Cochran	
	ASL	Cases	Lower	Upper	Lower	Upper
Men	BG	7	0.2	1.1	0.1	0.9
	BS	10	0.3	1.3	0.3	1.1
	CO	4	0.1	1.4	0.0	0.9
	CR	2	0.0	1.7	0.0	0.7
	LC	3	0.1	2.3	0.0	1.4
	LO	1	0.0	2.5	0.0	0.9
	MB	8	0.3	1.4	0.2	1.1
	MI	3	0.1	2.7	0.0	1.6
	MI1	7	0.2	1.2	0.1	0.9
	MI2	5	0.2	1.5	0.1	1.1
	MIC	5	0.1	0.6	0.0	0.3
	MN	2	0.0	1.5	0.0	0.5
	PV	9	0.4	1.9	0.3	1.3
	SO	3	0.2	3.6	0.0	2.0
	VA	10	0.3	1.4	0.2	1.1
	VS	0	0.0	4.5	0.0	0.0
Women	BG	2	0.0	0.5	0.0	0.2
	BS	7	0.1	1.0	0.1	0.8
	CO	2	0.0	0.9	0.0	0.3
	CR	1	0.0	1.6	0.0	0.5
	LC	2	0.1	2.1	0.0	1.2
	LO	0	0.0	2.2	0.0	0.0
	MB	3	0.0	0.9	0.0	0.6
	MI	1	0.0	2.1	0.0	0.8
	MI1	3	0.0	0.7	0.0	0.4
	MI2	0	0.0	0.8	0.0	0.0
	MIC	14	0.2	0.8	0.1	0.6
	MN	1	0.0	1.4	0.0	0.3
	PV	3	0.0	1.2	0.0	0.4
	SO	4	0.2	3.6	0.0	2.0
	VA	2	0.0	0.7	0.0	0.3
	VS	0	0.0	4.6	0.0	0.0

4 Conclusion

In this article, we illustrated the command `distrate`, which calculates confidence intervals for standardized rates using the formulas proposed by Tiwari, Clegg, and Zou (2006) as a modification of the method proposed by Fay and Feuer (1997). The method used by `distrate` is recommended in the case of rare diseases and is currently used in the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute in Bethesda, Maryland; the Italian Association of Cancer Registries (Associazione Italiana Registri Tumori, AIRTUM); and the Lombardy Mesothelioma and Sinonasal Cancer Registry. Useful characteristics of the program are the compact output, the possibility of specifying a desired multiplier for rates, and the easy output of rates and confidence intervals to an external file for further processing. These characteristics make `distrate` a useful addition to the official Stata command `dstdize`.

5 References

Breslow, N. E., and N. E. Day. 1987. *Statistical Methods in Cancer Research: Vol. 2—The Design and Analysis of Cohort Studies*. Lyon: IARC.

Checkoway, H., N. E. Pearce, and D. Kriebel. 2004. *Research Methods in Occupational Epidemiology*. 2nd ed. New York: Oxford University Press.

Clayton, D., and M. Hills. 1993. *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. New York: Wiley.

Curado, M. P., B. Edwards, H. R. Shin, H. Storm, J. Ferlay, M. Heanue, and P. Boyle, ed. 2007. *Cancer Incidence in Five Continents: Vol.ĨX*. Lyon: IARC.

Dobson, A. J., K. Kuulasmaa, E. Eberle, and J. Scherer. 1991. Confidence intervals for weighted sums of Poisson parameters. *Statistics in Medicine* 10: 457–462.

Fay, M. P., and E. J. Feuer. 1997. Confidence intervals for directly standardized rates: A method based on the gamma distribution. *Statistics in Medicine* 16: 791–801.

Keyfitz, N. 1966. Sampling variance of age standardised mortality rates. *Human Biology* 38: 309–317.

Miettinen, O. S. 1972a. Components of the crude risk ratio. *American Journal of Epidemiology* 96: 168–172.

———. 1972b. Standardization of risk ratios. *American Journal of Epidemiology* 96: 383–388.

———. 1985. *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. New York: Wiley.

———. 2011. *Epidemiological Research: Terms and Concepts*. Dordrecht: Springer.

Rothman, K. J. 1986. *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins.

———. 2002. *Epidemiology: An Introduction*. Oxford: Oxford University Press.

Rothman, K. J., S. Greenland, and T. L. Lash. 2008. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins.

Tiwari, R. C., L. X. Clegg, and Z. Zou. 2006. Efficient interval estimation for age-adjusted cancer rates. *Statistical Methods in Medical Research* 15: 547–569.

About the authors

Dario Consonni is an epidemiologist in the Epidemiology Unit at the Fondazione IRCCS Ca' Granda—Ospedale Maggiore Policlinico in Milan, Italy. His main activities are the design, conduct, and analysis of occupational, environmental, and clinical epidemiology studies. He teaches epidemiology and Stata at the Master in Epidemiology at the University of Turin, organized by the Associazione Italiana di Epidemiologia.

Enzo Coviello is an epidemiologist in the Statistics and Epidemiology Unit at ASL BT in Barletta, Italy. He is a longtime Stata user and enthusiast as well as the author of several popular Stata commands, including `stcascoh`, `stcompet`, and `strs`. His main interest is in the analysis of population-based cancer registries data.

Carlotta Buzzoni is a statistician working in the Clinical and Descriptive Epidemiology Unit at ISPO in Firenze, Italy. Her main interest is data management and statistical analyses of data for the Italian network of cancer registries (AIRTUM).

Carolina Mensi is an epidemiologist in the Department of Preventive Medicine at the Fondazione IRCCS Ca' Granda—Ospedale Maggiore Policlinico in Milan, Italy. Her main interest is data collection and management for the Lombardy Mesothelioma and Sinonasal Cancer Registry.