snp16.*y*	Robust confidence intervals for median (and other percentile) differences between two groups

Author: Roger Newson, Imperial College London, UK. Email: r.newson@imperial.ac.uk Date: 28 February 2012.

Abstract: A program is presented for calculating robust confidence intervals for Hodges–Lehmann median (and other percentile) differences (and ratios) between values of a variable in two subpopulations. The median difference is usually the same as that produced by the programs `cid` and `npshift`, using the Lehmann method. However, the confidence limits are typically different, being robust to the possibility that the two population distributions differ in ways other than location, such as having unequal variances. The program uses the program `somersd`, and is part of the `somersd` package.

Keywords: robust, confidence interval, median, percentile, difference, ratio, rank–sum, sign test, Wilcoxon, Hodges, Lehmann, paired, two–sample.

1 Syntax

`cendif` *depvar* [`using` *filename*][*weight*][`if` *exp*][`in` *range*], `by(`*groupvar*`)` [`centile(`*numlist*`)` `level(`#`)` `eform` `ystargenerate(`*newvarlist*`)` `cluster(`*varname*`)` `cfweight(`*expression*`)` `funtype(`*functional_type*`)` `tdist` `transf(`*transformation_name*`)` `saving(`*filename*[`,replace`]`)` `nohold`]

where *transformation_name* is one of

`iden | z | asin`

and *functional_type* is one of

`wcluster | bcluster | vonmises`

`fweights`, `iweights` and `pweights` are allowed; see help for `weight`. They are treated as described in **Methods and Formulas** below.

`bootstrap`, `by`, `jackknife`, and `statsby` are allowed; see help for `prefix`.

1.1 Description

`cendif` calculates confidence intervals for Hodges–Lehmann median differences, and other percentile differences, between values of a Y–variable in *depvar* for a pair of observations chosen at random from two groups A and B, defined by the *groupvar* in the `by` option. These confidence intervals are robust to the possibility that the population distributions in the two groups are different in ways other than location. This might happen if, for example, the two populations had different variances. For positive–valued variables, `cendif` can be used to calculate confidence intervals for median ratios or other percentile ratios. `cendif` is part of the `somersd` package (Newson, 2006a), and requires the program `somersd` in order to work.

1.2 Options

`by(`*groupvar*`)` is not optional. It specifies the name of the grouping variable. This variable must have exactly two possible values. The lower value indicates Group A, and the higher value indicates Group B.

`centile(`*numlist*`)` specifies a list of percentile differences to be reported, and defaults to `centile(50)` (median only) if not specified. Specifying `centile(25 50 75)` will produce the 25th, 50th and 75th percentile differences.

`level(`# `)` specifies the confidence level (percent) for confidence intervals; see help for `level`.

`eform` specifies that exponentiated percentile differences are to be given. This option is used if *depvar* is the log of a positive–valued variable. In this case, confidence intervals are calculated for percentile ratios between values of the original positive variable, instead of for percentile differences.

`ystargenerate(`*newvarlist*`)` specifies a list of variables to be generated, corresponding to the percentile differences, containing the differences $Y^*(\theta) = Y - group_1 * \theta$, where $group_1$ is a binary variable indicating membership of Group 1, and θ is the percentile difference. The variable names in the *newvarlist* are matched to the list of percentiles specified by the `centile()` option, sorted in ascending order of percent. If the two lists have different lengths, then `cendif` generates a number *nmin* of new variables equal to the minimum length of the two lists, matching the first *nmin* percentiles with the first *nmin* new variable names. Usually, there is only one percentile difference (the median difference), and one new `ystargenerate()` variable.

`cluster(`*varname*`)` specifies the variable which defines sampling clusters. If `cluster` is defined, then the confidence intervals are calculated assuming that the data are a sample of clusters from a population of clusters, rather than a sample of observations from a population of observations.

`cfweight(`*expression*`)` specifies an expression giving the cluster frequency weights. These cluster frequency weights must have the same value for all observations in a cluster. If `cfweight()` and `cluster()` are both specified, then each cluster in the dataset is assumed to represent a number of identical clusters equal to the cluster frequency weight for that cluster. If `cfweight()` is specified and `cluster()` is unspecified, then each observation in the dataset is treated as a cluster, and assumed to represent a number of identical one–observation clusters equal to the cluster frequency weight. For more details on the interpretation of weights, see **Interpretation of weights** in the manual `somersd.pdf`. Note that the observation frequency weights are used by `cendif` for tabulating the group frequencies.

`funtype(`*functional_type*`)` specifies whether the percentile differences estimated are between–cluster, within–cluster or Von Mises percentile differences. These three functional types are specified by the options `funtype(bcluster)`, `funtype(wcluster)` or `funtype(vonmises)`, respectively, and correspond to the functional types of the same names used by `somersd`. If `funtype()` is not specified, then `funtype(bcluster)` is assumed, and between–cluster percentile differences are estimated. If the clusters are pairs of observations, and if the `by()` option specifies an indicator variable indicating whether the observation is the first or second member of its pair, then the within–cluster median difference is the parameter corresponding to the sign test, and the Von Mises median difference is the conventional Hodges–Lehmann median difference between the group of first members and the group of second members, with confidence limits adjusted for clustering.

`tdist` specifies that the standardized Somers' D estimates are assumed to be sampled from a t–distribution with $n-1$ degrees of freedom, where n is the number of clusters, or the number of observations if `cluster()` is not specified. If `tdist` is not specified, then the standardized Somers' D estimates are assumed to be sampled from a standard Normal distribution. Simulation study data suggest that the `tdist` option should be recommended.

`transf(`*transformation_name*`)` specifies that the Somers' D estimates are to be transformed, defining a standard error for the transformed population value, from which the confidence limits for the percentile differences are calculated. `z` (the default) specifies Fisher's z (the hyperbolic arctangent), `asin` specifies Daniels' arcsine, and `iden` specifies identity or untransformed.

`saving(`*filename*`[, replace])` specifies a dataset, to be created, whose observations correspond to the observed values of differences between a value of *depvar* in Group A and a value of *depvar* in Group B. `replace` instructs Stata to replace any existing dataset of the same name. The saved dataset can then be re–used if `cendif` is called later, with `using`, to save the large amounts of processing time used to calculate the set of observed differences. The `saving()` option and the `using` utility are provided mainly for programmers to use, at their own risk.

`nohold` indicates that any existing estimation results are to be overwritten with a new set of estimation results, for the use of programmers. In default, any existing estimation results are restored after execution of `cendif`.

1.3 Saved results

`cendif` saves in `r()`:

Scalars

`r(N)`	number of observations	`r(N_clust)`	number of clusters
`r(N_1)`	first sample size N_1	`r(N_2)`	second sample size N_2
`r(df_r)`	residual degrees of freedom (if `tdist` present)	`r(level)`	confidence level

Macros

`r(depvar)`	name of *Y*–variable	`r(by)`	name of `by()` variable defining groups
`r(clustvar)`	name of cluster variable	`r(cfweight)`	`cfweight()` expression
`r(funtype)`	`funtype()` option	`r(tdist)`	`tdist` if specified
`r(wtype)`	weight type	`r(wexp)`	weight expression
`r(centiles)`	list of percents for percentiles	`r(Dslist)`	list of D^*–values for percentiles
`r(transf)`	transformation specified by `transf()`	`r(tranlab)`	transformation label in output
`r(eform)`	`eform` if specified		

Matrices

`r(cimat)`	confidence intervals for differences or ratios	`r(Dsmat)`	upper and lower limits for $D^*(\theta)$

The mathematical notation is specified in **Methods and Formulas** below.

2 Methods and Formulas

Suppose that a population contains two disjoint subpopulations A and B, and a random variable Y is defined for individuals from both subpopulations. For $0 < q < 1$, a $100q$th percentile difference in Y between Populations A

and B is defined as a value θ satisfying

$$D[Y^*(\theta)|X] = 1 - 2q, \tag{1}$$

where X is a binary variable equal to 1 for Population A and 0 for Population B, $Y^*(\theta)$ is defined as Y if $X = 1$ and $Y + \theta$ if $X = 0$, and $D[\cdot\,|\,\cdot]$ denotes Somers' D (Somers, 1962; Newson, 2006a). Somers' D is defined as

$$D[V|W] = E\left[\,\mathrm{sign}(V_1 - V_2)\,\mathrm{sign}(W_1 - W_2)\,\right] / E\left[\mathrm{sign}(W_1 - W_2)^2\right], \tag{2}$$

where (W_1, V_1) and (W_2, V_2) are bivariate data points sampled independently from the same population, and $E[\cdot]$ denotes expectation. In the case of (1), where $W = X$ and $V = Y^*(\theta)$, Somers' D is the difference between two conditional probabilities. Given an individual sampled from Population A and an individual sampled from Population B, these are the probability that the individual from Population A has the higher Y^*–value and the probability that the individual from Population B has the higher Y^*–value. Somers' D is therefore the parameter equal to zero under the null hypothesis tested by the "non–parametric" Wilcoxon rank–sum test on $Y^*(\theta)$. In the case where $q = 0.5$ (and therefore $1 - 2q = 0$), a $100q$th percentile difference is known as a Hodges–Lehmann median percentile difference, and is zero under the null hypothesis tested by a Wilcoxon rank–sum test on Y. The median percentile difference was introduced explicitly by Hodges and Lehmann (1963). However, it is also a special case of the Theil median slope, introduced by Theil (1950) and discussed by Sen (1968). The Hodges–Lehmann median difference is simply a Theil–Sen median slope where the X–variable is binary.

Note that a value of θ satisfying (1) is not always unique. If Y has a discrete distribution, then there may be no solution, or a wide interval of solutions. However, the method used here is intended to produce a confidence interval containing any given θ satisfying (1), with a probability at least equal to the confidence level, if such a θ exists.

We will assume that there are N_1 observations sampled from Population A and N_2 observations sampled from Population B, giving a total of $N_1 + N_2 = N$ observations. These observations will be identified by double subscripts, so that Y_{ij} is the Y–value for the jth observation sampled from the ith population (where i=1 for Population A and $i = 2$ for Population B). The corresponding X–values (ones and zeros) will be denoted X_{ij}. The observations will be assumed to have importance weights (`iweights` or `pweights`) denoted w_{ij}, and cluster sequence numbers denoted c_{ij}. `cendif` follows the usual Stata practice of assuming an `fweight` to stand for multiple observations, with the same values for all other variables. The clusters may be nested within the two groups or contain observations from each of the two groups. If clusters are present, then the confidence intervals will be calculated assuming that the sample was generated by sampling clusters independently from a population of clusters, rather than by sampling N observations independently from the total population of observations, or by sampling N_1 and N_2 observations from Populations A and B, respectively. (In default, all the w_{ij} will be ones, and the c_{ij} will be in sequence from 1 to N, so the difference between these three alternatives will not matter.) We will denote by M the number of distinct values of a difference $Y_{1j} - Y_{2k}$ observed between Y–values in the two samples. The difference values themselves will be denoted $t_1, \ldots, t_M$. For each h from 1 to M, we define the sum of product weights of differences equal to t_h as

$$W_h = \sum_{j,k:\, Y_{1j} - Y_{2k} = t_h} \delta(c_j, c_k) w_{1j} w_{2k}, \tag{3}$$

where $\delta(\cdot,\cdot)$ is a function specified by the `funtype()` option. If `funtype(bcluster)` is specified (or if `funtype()` is unspecified), then $\delta(a,b)$ is 0 if $a = b$ and 1 if $a \neq b$. If `funtype(wcluster)` is specified, then $\delta(a,b)$ is 1 if $a = b$ and 0 if $a \neq b$. If `funtype(vonmises)` is specified, then $\delta(a,b)$ is 1 for all a and b. Therefore, W_h is a sum of between–cluster product weights if `funtype(bcluster)` is specified, a sum of within–cluster product weights `funtype(wcluster)` is specified, and a total sum of product weights if `funtype(vonmises)` is specified.

Given a value of θ expressed in units of Y, we can define $Y^*_{ij}(\theta)$ to be Y_{ij} if $i = 1$, and $Y_{ij} + \theta$ if $i = 2$. The sample Somers' D of $Y^*(\theta)$ with respect to X is defined as

$$D^*(\theta) = \hat{D}[Y^*(\theta)|X] = \frac{\sum_{j=1}^{N_1}\sum_{k=1}^{N_2} \delta(c_{1j}, c_{2k})\, w_{1j} w_{2k}\, \mathrm{sign}(Y_{1j} - Y_{2k} - \theta)}{\sum_{j=1}^{N_1}\sum_{k=1}^{N_2} \delta(c_{1j}, c_{2k})\, w_{1j} w_{2k}} = \frac{\sum_{h:\, t_h > \theta} W_h - \sum_{h:\, t_h < \theta} W_h}{\sum_{h=1}^{M} W_h}, \tag{4}$$

where $\hat{D}[\cdot\,|\,\cdot]$ denotes the sample Somers' D, defined by the methods of Newson (2006a). Clearly, given a sample, $D^*(\theta)$ is a nonincreasing function of θ. Figure 1 shows $D^*(\theta)$ as a function of θ for differences between trunk capacities of US and non–US cars (expressed in cubic feet) in the `auto` data. The squares represent the values $D^*(t_h)$ for the observed differences t_h. Note that $D^*(\theta)$ is discontinuous at the observed differences, and constant in each open interval between two successive observed differences.

We aim to include θ in a confidence interval for a $100q$th percentile difference if, and only if, the sample $D^*(\theta)$ is compatible with a *population* $D[Y^*(\theta)|X]$ equal to $1-2q$. The methods of Newson (2006a), used by the program `somersd` and described in the manual `somersd.pdf`, typically use a transformation $\zeta(\,\cdot\,)$, which, for present purposes, may either be the identity, the arscine or Fishers' z (the hyperbolic arctangent). The transformed sample statistic $\hat{\zeta}(\theta) = \zeta[D^*(\theta)]$ is assumed to be Normally distributed around the population parameter $\zeta\{D[Y^*(\theta)|X]\}$. In the present application, we assume that, if $D[Y^*(\theta)|X] = 1-2q$, then the quantity

$$[\,\hat{\zeta}(\theta) - \zeta(1-2q)\,] \,/\, \mathrm{SE}[\hat{\zeta}(\theta)] \tag{5}$$

has a standard Normal distribution, where $\mathrm{SE}[\hat{\zeta}(\theta)]$ is the sampling standard deviation (or standard error) of $\zeta[D^*(\theta)]$. If we knew the value of $\mathrm{SE}[\hat{\zeta}(\theta)]$, then a $100(1-\alpha)\%$ confidence interval for a $100q$th percentile difference might be the interval of values of θ for which

$$\zeta^{-1}\big\{\,\zeta(1-2q) - z_\alpha\,\mathrm{SE}[\hat{\zeta}(\theta)]\,\big\} \le D^*(\theta) \le \zeta^{-1}\big\{\,\zeta(1-2q) + z_\alpha\,\mathrm{SE}[\hat{\zeta}(\theta)]\,\big\}, \tag{6}$$

where z_α is the $100(1-\frac{1}{2}\alpha)$th percentile of the standard Normal distribution.

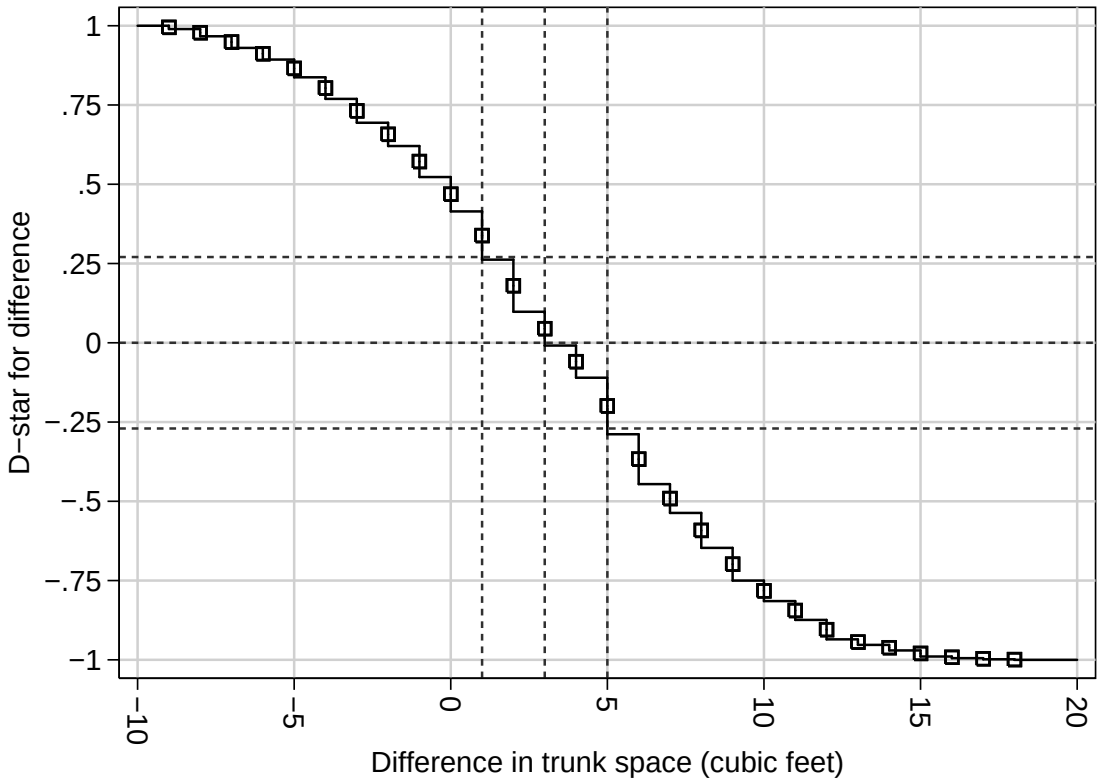

Figure 1. $D^*(\theta)$ plotted against the difference θ in trunk space between US and non–US cars

To construct such a confidence interval, we proceed as follows. Given a value of D such that $-1 < D < 1$, define

$$B_{\mathrm{L}}(D) = \inf\{\theta : D^*(\theta) \le D\}, \quad B_{\mathrm{R}}(D) = \sup\{\theta : D^*(\theta) \ge D\},$$

$$B_{\mathrm{C}}(D) = \begin{cases} B_{\mathrm{L}}(D), & \text{if } B_{\mathrm{R}}(D) = \infty \\ B_{\mathrm{R}}(D), & \text{if } B_{\mathrm{L}}(D) = -\infty \\ [\,B_{\mathrm{L}}(D) + B_{\mathrm{R}}(D)\,]/2, & \text{otherwise.} \end{cases} \tag{7}$$

(By convention, the supremum (or infimum) of a set unbounded to the right (or left) are defined as ∞ (or $-\infty$), respectively.) Clearly, $B_{\mathrm{L}}(D) \le B_{\mathrm{C}}(D) \le B_{\mathrm{R}}(D)$, and the values of $B_{\mathrm{L}}(D)$ and $B_{\mathrm{R}}(D)$ (if finite) can be either the same t_h, or two successive ones. The confidence interval for a $100q$th percentile difference is centered on the sample $100q$th percentile difference, defined as

$$\hat{\xi}_q = B_{\mathrm{C}}(1-2q). \tag{8}$$

`cendif` then calls `somersd`, with the X_{ij} as the predictor variable, and the $Y^*_{ij}(\hat{\xi}_q)$, for the values of q implied by the `centile()` option, as the predicted variables. The standard errors generated by `somersd` are used as estimates $\widehat{\mathrm{SE}}[\hat{\zeta}(\hat{\xi}_q)]$ of the standard error of $\hat{\zeta}(\theta)$ where θ satisfies (1). The lower and upper confidence limits for a $100q$th percentile difference are, respectively,

$$\hat{\xi}_q^{(\min)} = B_{\mathrm{L}}\Big(\zeta^{-1}\big\{\,\zeta(1-2q) - z_\alpha\,\widehat{\mathrm{SE}}[\hat{\zeta}(\hat{\xi}_q)]\,\big\}\Big), \quad \hat{\xi}_q^{(\max)} = B_{\mathrm{R}}\Big(\zeta^{-1}\big\{\,\zeta(1-2q) + z_\alpha\,\widehat{\mathrm{SE}}[\hat{\zeta}(\hat{\xi}_q)]\,\big\}\Big). \tag{9}$$

If `tdist` is specified, then `cendif` uses the t–distribution with $\nu = N-1$ degrees of freedom (or $\nu = N_{\mathrm{clust}}-1$ degrees of freedom if there are N_{clust} clusters) instead of the Normal distribution, so $t_{\nu,\alpha}$ replaces z_α in (9). Note that the upper and lower confidence limits may occasionally be infinite, in the case of extreme percentiles and/or very small sample numbers. `cendif` codes these infinite limits as plus or minus the Stata `creturn` value `c(maxdouble)`, which is the system maximim double precision value (see on–line help for `creturn`). Figure 1 shows the median difference in trunk capacity, and its confidence limits, as reference lines on the horizontal axis. The estimated median difference is 3 cubic feet, with 95% confidence limits from 1 to 5 cubic feet. The reference lines on the vertical axis are the optimum, minimum and maximum values of $D^*(\theta)$ required for θ to be in the confidence interval. These values of $D^*(\theta)$ are saved by `cendif` in the matrix `r(Dsmat)`. If the option `saving()` is specified, then `cendif` also saves an output dataset with M observations, corresponding to the ordered differences t_h. The variables are `diff` (containing the t_h), `weight` (containing the W_h), `Dstar` (containing the $D^*(t_h)$), and `Dstar_r`, which contains the right–hand limiting value of $D^*(\theta)$,

$$D^*_R(t_h) = \lim_{\theta \to t_h +} D^*(\theta), \tag{10}$$

which is the value of $D^*(\theta)$ in the open interval (t_h, t_{h+1}) for $h < M$.

Lehmann (1963) presents a method which, for large samples, is essentially equivalent to (6), in the special case where $q = 0.5$ and $\zeta(D) = D$. (This is the method for calculating confidence intervals for median differences popularized by Conover (1999), Campbell and Gardner (1988) and Gardner and Altman *et al.* (2000), and available in Stata using Duolao Wang's `npshift` routine (Wang, 1999) or Patrick Royston's `cid` routine, downloadable from SSC (Royston, 1998).) However, Lehmann's method uses the assumption that the two population distributions are different only in location. This assumption (essentially) enables the calculation of $\mathrm{SE}[\hat{\zeta}(\theta)]$ for large samples, and of the exact distribution of $D^*(\theta)$ for small samples. It also implies that the median difference is the difference between medians. In the present case, we are not making this assumption, as the confidence interval is intended to be robust to the possibility that the two populations are different in ways other than location. (For instance, the two populations might be unequally variable.) The median difference is therefore not necessarily the difference between medians. Also, we have to estimate $\mathrm{SE}[\hat{\zeta}(\theta)]$, and this estimate is itself subject to some amount of sampling error. The method of `cendif` contrasts to Lehmann's method as the unequal–variance t–test contrasts to the equal–variance t–test. Lehmann's method, like the equal–variance t–test, assumes that you can use data from the larger of two samples to estimate the population variability of the smaller sample.

I have been carrying out some simulations of sampling from two Normal populations, with a view to finding the coverage probabilities and median lengths of the confidence intervals for the median difference generated by `npshift` and by `cendif` with the `tdist` option. So far, I find that, even with small sample sizes, the `cendif` method *usually* gives coverage probabilities closer to the nominal value than the Lehmann method when variances are unequal, in which case `npshift` produces confidence intervals either too wide or too narrow, depending on whether the larger or smaller sample has the greater population variance. Usually, the difference in coverage probability is small (1% or 2%), so the Lehmann method performs fairly well, in spite of false assumptions. However, if a sample of 20 is compared to a sample of 10, and the population standard deviation of the smaller sample is three times that of the larger sample, then the nominal 95% confidence interval has a true coverage probability of only 90% under the Lehmann method, compared to 94% under the `cendif` method. I plan to report the results of these simulations in detail elsewhere.

3 Examples

3.1 Car weights in the `auto` data

In the `auto` data, we compare weights of US cars and non–US cars. We use `cid` and `cendif` to estimate the median difference:

```
. cid weight, by(foreign) median unpaired
Rank-based confidence interval for difference in  medians by foreign
Variable |     Obs     Estimate           K        [95% Conf. Interval]
```

```
---------+-------------------------------------------------------------
  weight |      74         1095         406             720        1350
. cendif weight, tdist by(foreign)
Y-variable: weight (Weight (lbs.))
Grouped by: foreign (Car type)
Group numbers:
   Car type |      Freq.     Percent        Cum.
------------+-----------------------------------
   Domestic |         52       70.27       70.27
    Foreign |         22       29.73      100.00
------------+-----------------------------------
      Total |         74      100.00
Transformation: Fisher's z
Degrees of freedom: 73
95% confidence interval(s) for percentile difference(s)
between values of weight in first and second groups:
   Percent    Pctl_Dif     Minimum     Maximum
        50        1095         750        1330
```

We note that the median difference in weight is 1095 pounds, according to both `cid` and `cendif`. However, the confidence limits given by `cendif` are 750 and 1330 pounds, whereas the confidence limits given by `cid` are 720 and 1350 pounds. This is because non–US cars are fewer in number, and less variable in weight, than US cars, and `cid` assumes equal variances, whereas `cendif` allows for unequal variances. If we carry out equal–variance and unequal–variance t–tests (not shown), then we find a similar difference in the width of the confidence limits for the mean difference.

`cendif` can also calculate confidence intervals for percentiles other than medians. These contain information about the degree of overlap between the two populations. Here, we estimate the 0th, 25th, 50th, 75th and 100th percentile differences, using the `centile` option:

```
. cendif weight, tdist by(foreign) ce(0(25)100)
Y-variable: weight (Weight (lbs.))
Grouped by: foreign (Car type)
Group numbers:
   Car type |      Freq.     Percent        Cum.
------------+-----------------------------------
   Domestic |         52       70.27       70.27
    Foreign |         22       29.73      100.00
------------+-----------------------------------
      Total |         74      100.00
Transformation: Fisher's z
Degrees of freedom: 73
95% confidence interval(s) for percentile difference(s)
between values of weight in first and second groups:
   Percent    Pctl_Dif     Minimum     Maximum
         0       -1620  -8.99e+307       -1620
        25         485         90         820
        50        1095         750        1330
        75        1555        1320        1790
       100        3080        3080   8.99e+307
```

The 0th and 100th percentile differences are the minimum and maximum differences, respectively, and their confidence limits extend from their respective sample values to the system limits `c(mindouble)` and `c(maxdouble)`, representing true values of $-\infty$ and $+\infty$, respectively.

If we want to estimate percentile ratios of weight, rather than percentile differences, then we simply take logs and use the `eform` option:

```
. gene logwt=log(weight)
. cendif logwt, tdist by(foreign) ce(0(25)100) eform
Y-variable: logwt
Grouped by: foreign (Car type)
Group numbers:
   Car type |      Freq.     Percent        Cum.
------------+-----------------------------------
   Domestic |         52       70.27       70.27
    Foreign |         22       29.73      100.00
------------+-----------------------------------
      Total |         74      100.00
```

```
Transformation: Fisher's z
Degrees of freedom: 73
95% confidence interval(s) for percentile ratio(s)
between values of exp(logwt) in first and second groups:
   Percent    Pctl_Rat     Minimum     Maximum
         0   .52631583           0   .52631583
        25   1.1935375   1.0328637    1.358491
        50   1.4806389   1.3090908   1.6323524
        75    1.744916   1.6071432   1.8774505
       100   2.7499989   2.7499989   8.99e+307
```

We note that, typically, US cars are 148% as heavy as non–US cars, with confidence limits ranging from 131% to 163% as heavy. The 25th percentile ratio (103% to 136%) shows that the two car types do not overlap to a great extent. The 0th and 100th percentile differences have confidence intervals extending from their sample values to 0 and "$+\infty$", respectively.

3.2 Paired blood pressures in the `bplong` data

The `bplong` dataset is distributed with Stata. It has one observation per blood pressure measurement per patient in a fictional medical study in which the blood pressure of each patient is measured twice, once before treatment and once after treatment. It can be accessed by the `sysuse` command as follows:

```
. sysuse bplong, clear
(fictional blood-pressure data)
. describe, simple
patient  sex      agegrp   when     bp
```

The variable `patient` identifies the patient, the variable `when` is 1 for before–treatment measurements and 2 for after–treatment measurements, and the variable `bp` contains the blood pressure, presumably expressed in millimetres of mercury (mm Hg). Note that the dataset `bpwide`, also distributed with Stata, contains the same data in wide format, with one observation per patient and two blood pressure variables, representing the patient's blood pressure before and after treatment, respectively. However, `cendif` can only use these data in long format, with one observation per blood pressure per patient.

We might wish to compare untreated and treated blood pressures. One way to do this is to estimate the median of all paired differences between untreated and treated blood pressures from the same patient. We can do this using `cendif`, with the options `cluster(patient)` and `funtype(wcluster)`:

```
. cendif bp, by(when) tdist cluster(patient) funtype(wcluster)
Y-variable: bp (Blood pressure)
Grouped by: when (Status)
Group numbers:
     Status |      Freq.     Percent        Cum.
------------+-----------------------------------
     Before |         120       50.00       50.00
      After |         120       50.00      100.00
------------+-----------------------------------
      Total |         240      100.00
Transformation: Fisher's z
Degrees of freedom: 119
Number of clusters (patient) = 120
95% confidence interval(s) for within-cluster percentile difference(s)
between values of bp in first and second groups:
   Percent    Pctl_Dif     Minimum     Maximum
        50         3.5           1           8
```

The dataset contains 240 measurements on 120 patients, of which 120 measurements were taken before treatment, and 120 measurements were taken from the same patients after treatment. The sample within–cluster median difference between untreated and treated blood pressures is 3.5 mm Hg, with 95% confidence limits from 1 to 8 mm Hg. Therefore, a typical patient in the population experiences a lowering of blood pressure between 1 mm Hg and 8 mm Hg after treatment. This within–cluster median difference is a rank–based parameter, whose value is zero under the null hypothesis tested by `signtest` (see [R] `signrank`). Note that `signtest` can only be used on wide–format data, and produces a P–value, whereas `cendif` can only be used on long–format data, and produces a confidence interval.

The sign test may have low power to detect a difference, compared to alternatives such as the signed–rank test and the paired t–test. Another possibly more powerful alternative, which is rank–based and produces a confidence

interval, is to estimate the median difference between all pairs of untreated and treated measurements, whether or not they belong to the same patient. `cendif` can do this, using the options `cluster(patient)` and `funtype(vonmises)`, as follows:

```
. cendif bp, by(when) tdist cluster(patient) funtype(vonmises)
Y-variable: bp (Blood pressure)
Grouped by: when (Status)
Group numbers:
     Status |      Freq.     Percent        Cum.
------------+-----------------------------------
     Before |         120       50.00       50.00
      After |         120       50.00      100.00
------------+-----------------------------------
      Total |         240      100.00
Transformation: Fisher's z
Degrees of freedom: 119
Number of clusters (patient) = 120
95% confidence interval(s) for Von Mises percentile difference(s)
between values of bp in first and second groups:
   Percent    Pctl_Dif     Minimum     Maximum
        50           6           2           9
```

This time, the median difference is 6 mm Hg, with 95% confidence limits from 2 to 9 mm Hg. Therefore, the confidence interval for the Von Mises percentile difference (for all pairs of untreated and treated blood pressures) is as wide as the confidence interval for the within–cluster median difference (between untreated and treated blood pressures from the same patient). However, Rosner *et al.* (2006) report the results of a simulation study, in which the power of a clustered rank–sum test was greater than that of the alternative signed–rank test when testing for non–zero median differences between sets of paired data comparable to `bplong`. The Von Mises median difference is based on the equivalent of a clustered rank–sum test, and the signed–rank test is usually more powerful than the sign test. We might therefore expect the confidence interval for the Von Mises median difference to be narrower than that for the within–cluster median difference in most samples. However, more work is required to assess the relative power of these methods.

4 Acknowledgements

I would like to thank Nicholas J. Cox of Durham University, UK, and William Gould of StataCorp, for some very helpful advice on the coding of infinite confidence limits, such as those occasionally resulting from Equation (9). I would also like to thank David Airey of Vanderbilt University, TN, USA for drawing my attention to some recent studies on rank–sum tests for clustered data, including Rosner *et al.* (2006).

5 Historical note

This document is a post–publication update of an article which appeared in the Stata Technical Bulletin (STB) as Newson (2000d). The `somersd` package appeared in Newson (2000a), and a post–publication update of that STB article is distributed with this document as part of the documentation of the `somersd` package. The `somersd` package was later revised in Newson (2000b), Newson (2000c), Newson (2000d), Newson (2001a), Newson (2001b) and Newson (2006a). After 2001, STB was replaced by The Stata Journal (SJ), and most subsequent updates to the `somersd` package only appeared on SSC and on Roger Newson's homepage at *http://www.imperial.ac.uk/nhli/r.newson/*, which is accessible from within net–aware Stata as of 28 February 2012. However, Newson (2002) gives a comprehensive review of Somers' D, Kendall's τ_a, median differences, and their estimation in Stata using the `somersd` package, and Newson (2006a) and Newson (2006b) describe the update of `somersd` to Version 9 of Stata.

6 References

Altman, D. G., D. Machin, T. N Bryant and M. J Gardner. 2000. *Statistics with Confidence*. London, UK: British Medical Journal.

Campbell, M. J. and M. J. Gardner. 1988. Calculating confidence intervals for some non–parametric analyses. *British Medical Journal* 296: 1454–1456.

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

Hodges, J. L. and E. L. Lehmann. 1963. Estimates of location based on rank tests. *The Annals of Mathematical Statistics* 34(2): 598–611.

Lehmann, E. L. 1963. Nonparametric confidence intervals for a shift parameter. *The Annals of Mathematical Statistics* 34(4): 1507–1512.

Newson, R. 2000a. snp15: `somersd` – Confidence intervals for nonparametric statistics and their differences. *Stata Technical Bulletin* 55: 47–55. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 312–322.

Newson, R. 2000b. snp15.1: Update to `somersd`. *Stata Technical Bulletin* 57: 35. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 322–323.

Newson, R. 2000c. snp15.2: Update to somersd. *Stata Technical Bulletin* 58: 30. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, p. 323.

Newson, R. 2000d. snp16: Robust confidence intervals for median and other percentile differences between groups. *Stata Technical Bulletin* 58: 30–35. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 324–331.

Newson, R. 2001a. snp15.3: Update to somersd. *Stata Technical Bulletin* 61: 22. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, p. 33x.

Newson, R. 2001b. snp16.1: Update to cendif. *Stata Technical Bulletin* 61: 22. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, p. 33x.

Newson, R. 2002. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal* 2: 45–64. A pre–publication draft can be downloaded from Roger Newson's website at http://www.imperial.ac.uk/nhli/r.newson/ as of 28 February 2012, using the net command in Stata.

Newson, R. 2006a. Confidence intervals for rank statistics: Somers' D and extensions. *Stata Journal* 6(3): 309–334. A pre–publication draft can be downloaded from Roger Newson's website at http://www.imperial.ac.uk/nhli/r.newson/ as of 28 February 2012, using the net command in Stata.

Newson, R. 2006b. Confidence intervals for rank statistics: Percentile slopes, differences, and ratios. *Stata Journal* 6(4): 497–520. A pre–publication draft can be downloaded from Roger Newson's website at http://www.imperial.ac.uk/nhli/r.newson/ as of 28 February 2012, using the net command in Stata.

Rosner, B., R. J. Glynn and M–L. T. Lee. 2006. Extension of the rank–sum test for clustered data: Two–group comparisons with group membership defined at the subunit level. *Biometrics* 62(4): 1251–1259.

Royston, P. 1998. CID: Stata module to calculate confidence intervals for means or differences. On the SSC–Ideas list at http://ideas.uqam.ca/ideas/data/Softwares/bocbocodeS338001.html as of 28 February 2012.

Sen, P. K. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* 63(324): 1379–1389.

Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. *American Sociological Review* 27: 799–811.

Theil, H. 1950. A rank invariant method of linear and polynomial regression analysis, I, II, III. *Proceedings of the Koninklijke Nederlandse Akademie Wetenschappen, Series A – Mathematical Sciences* 53: 386–392, 521–525, 1397–1412.

Wang, D. 1999. sg123: Hodges–Lehmann estimation of a shift in location between two populations. *Stata Technical Bulletin* 52: 52–53. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 255–257.