The Stata Journal (2016)
16, Number 4, pp. 880–899

Distributional estimates for the comparison of proportions of a dichotomized continuous outcome

Odile Sauzet
Department of Epidemiology and International Public Health
Bielefeld School of Public Health
Bielefeld University
Bielefeld, Germany
odile.sauzet@uni-bielefeld.de

Maren Kleine
Department of Epidemiology and International Public Health
Bielefeld School of Public Health
Bielefeld University
Bielefeld, Germany
maren.kleine@uni-bielefeld.de

Abstract. We present the package `distdicho`, which contains a range of commands covering the present development of the distributional method for the dichotomization of continuous outcomes. The method provides estimates with standard error of a comparison of proportions (difference, odds ratio, and risk ratio) derived, with similar precision, from a comparison of means.

Keywords: st0459, distdicho, distdichoi, reg_distdicho, sk_distdicho, sk_distdichoi, distributional method, dichotomization, continuous outcomes, skew-normal distribution

1 Introduction and motivation

Dichotomization of continuous outcomes is a common practice despite numerous arguments against it (Ragland 1992; Royston, Altman, and Sauerbrei 2006). A reason for this lies in the interpretation of results in terms of population at risk or patients who require a treatment. The distributional method for the dichotomization of continuous outcomes has been developed to allow comparisons of proportions to complement a comparison of means with equal precision. The original work was developed for the comparison of two groups for outcomes normally distributed with equal variance in the two groups (Peacock et al. 2012). Because of the restrictive nature of the equal variance hypothesis, the method has been further developed to provide a correction for unequal variances (Sauzet and Peacock 2014). Sauzet, Ofuya, and Peacock (2015) addressed the question of the robustness to deviations from normality and showed that for small deviations the method worked well. In case of perturbation to the normal distribution (for example, because of an excess of patients with high blood pressure or preterm babies having much lower birth weights), a method based on the skew-normal distribution (Azzalini 2005) has also been proposed in Sauzet, Ofuya, and Peacock (2015).

Because most analyses comparing continuous outcomes between two groups are not performed with a t test but with potentially complex regression models, comparisons of proportions can also be obtained, after adjusting the distributional method, to reflect the results of linear possibly mixed models (Sauzet et al. 2016).

We have developed a package of commands to cover all the applications of the distributional methods that have been developed so far. In the following, we illustrate the usage of the various commands and options in the package.

2 Distributional estimates for the comparison of proportions

2.1 The normal method

In this section, we review the basic principle of the distributional method as published in Peacock et al. (2012) and Sauzet and Peacock (2014).

The distributional method is a large-sample approximation method for the estimation of proportions and their standard errors (SEs) assuming a normal distribution for the data. It is based on the delta method and uses estimates for the mean and variance from the data. We recall here the formula obtained to compute estimates and SEs for proportions, difference in proportions, risk ratios (RRs), and odds ratios (ORs) derived from the normal distribution.

Let $\overline{X}_n$ be the sample mean of n independent, identically normally distributed random variables $X_i,\ i = 1, \dots, n$. Let x_0 be a real number. The random variable $p(\overline{X}_n)$ for the proportion of the population with an outcome value under the threshold (cutpoint) x_0 is defined as

$$p\left(\overline{X}_n\right) = \int_{-\infty}^{x_0} f_{N(\overline{X}_n,\sigma^2)}(t)dt$$

where $f_{N(\mu,\sigma^2)}$ is the density function of the normal distribution with mean μ and variance σ^2. It is a function of the sample mean with variance σ^2. According to the delta method, $p(\overline{X}_n)$ is asymptotically normally distributed with mean $p(\overline{x}_n)$ (mean sample estimate) and standard deviation (SD)

$$\text{SD}\left\{p\left(\overline{X}_n\right)\right\} = \frac{s}{\sqrt{n}} f_{N(\overline{x}_n,s^2)}(x_0)$$

So, the estimate for the proportion under quantile x_0 is estimated by $\int_{-\infty}^{x_0} f_{N(\overline{x}_n,s^2)}(t)dt$ with $\text{SE} = (s/\sqrt{n}) f_{N(\overline{x}_n,s^2)}(x_0)$, where s is the sample estimate for the SD assumed to be the known SD in the population.

Therefore, for two groups, if the variance is assumed to be the same in both groups, we obtain estimates for the difference in proportion d as the difference between the estimated proportions with SE, using for the common SD the pooled estimate from the data,

$$s_{\text{pooled}} = \sqrt{\frac{(n_t-1)s_t^2 + (n_c-1)s_c^2}{(n_t+n_c-2)}}$$

$$\text{SE}(d) = \sqrt{\frac{s_{\text{pooled}}^2}{n_t} f^2_{N(\overline{x}_{t,n_t}, s^2_{\text{pooled}})}(x_0) + \frac{s_{\text{pooled}}^2}{n_c} f^2_{N(\overline{x}_{c,n_c}, s^2_{\text{pooled}})}(x_0)}$$

Estimates for the SE for the log(RR) are obtained through the function $h(\overline{X}_n) = \log\{p(\overline{X}_n)\}$. The SE for the log(RR) is

$$\text{SE}\{\log(\text{RR})\} = \sqrt{\frac{s_{\text{pooled}}^2}{n_t} \frac{f^2_{N(\overline{x}_{t,n_t}, s^2_{\text{pooled}})}(x_0)}{p_t^2} + \frac{s_{\text{pooled}}^2}{n_c} \frac{f^2_{N(\overline{x}_{c,n_c}, s^2_{\text{pooled}})}(x_0)}{p_c^2}}$$

Estimates for the SE for the log(OR) are obtained through the function $g(\overline{X}_n) = \log[\{p(\overline{X}_n)\}/\{1-p(\overline{X}_n)\}]$. The SE for the log(OR) is

$$\text{SE}\{\log(\text{OR})\} = \sqrt{\frac{s_{\text{pooled}}^2}{n_c} \frac{f^2_{N(\overline{x}_{c,n_c}, s^2_{\text{pooled}})}(x_0)}{p_c^2(1-p_c)^2} + \frac{s_{\text{pooled}}^2}{n_t} \frac{f^2_{N(\overline{x}_{t,n_t}, s^2_{\text{pooled}})}(x_0)}{p_t^2(1-p_t)^2}}$$

The equal variance condition can be relaxed either by providing a known ratio of variances between the two groups or, when this is not possible, by adding a correction factor to the SE which otherwise would be underestimated when the variances are not assumed known. Moreover, this correction factor can also be used to correct the SEs for large effects (see Sauzet and Peacock [2014]), because the variability due to using the observed pooled SD needs to be accounted for in the SE whether the variances are assumed equal or not.

2.2 The skew-normal method

The principle of the skew-normal method is the same as for the normal method but using the skew-normal distribution defined by Azzalini (2005). This distribution is a generalization of the normal distribution that works by adding a third parameter, α, defining the skewness ($\alpha = 0$ gives the normal distribution). We briefly recall how the formula for the SEs is obtained (Sauzet, Ofuya, and Peacock 2015).

Let $\overline{X}_n$ be the sample mean of n independent, identically skew-normal distributed random variables $X_i,\ i = 1, \ldots, n$, with mean μ, variance σ^2, and skewness parameter α. Let x_0 be a threshold of interest. The random variable $p(\overline{X}_n)$ for the proportion of the population with an outcome value under the threshold x_0 is defined as

$$p(\overline{X}_n) = \int_{-\infty}^{x_0} 2\frac{e^{\frac{-1}{2w^2}\left\{t-\left(\overline{X}_n+\alpha'\right)\right\}^2}}{\sqrt{2\pi w^2}} \left[\int_{-\infty}^{\alpha\left\{t-\left(\overline{X}_n+\alpha'\right)\right\}/w} \frac{e^{\frac{-1}{2}r^2}}{\sqrt{2\pi}} dr\right] dt$$

where $\alpha' = \mu - w\mu_z$, and $w^2 = \sigma^2/(1-\mu_z^2)$ with $\mu_z^2 = (2/\pi)\{\alpha^2/(1+\alpha^2)\}$ (see Azzalini [2005]).

From the delta method, we obtain that $p(\overline{X}_n)$ is approximately normally distributed with SD

$$\frac{w^2}{\sqrt{n}}\left(1-\mu_z^2\right)p'(\mu)^2$$

The formula for $p'(\mu)$ was derived in Sauzet, Ofuya, and Peacock (2015), where we obtained

$$p'\left(\overline{X}_n\right) = -2\frac{e^{\frac{-1}{2w^2}\left\{x_0-\left(\overline{X}_n+\alpha'\right)\right\}^2}}{\sqrt{2\pi w^2}}\Phi\left[\alpha\left\{x_0-\left(\overline{X}_n-\alpha'\right)\right\}/w\right]$$

with Φ being the standard normal cumulative distribution function.

The formulas for the SEs for the difference in proportions d, log(RR), and log(OR) are as follows:

$$\begin{aligned}
\text{SE}(d)^2 = {} & \frac{w_1^2}{\sqrt{n_1}}\left(1-\mu_z^2\right)\left[\frac{2e^{\frac{-1}{2w_1^2}\left\{x_0-\left(\mu_1+\alpha_1'\right)\right\}^2}}{\sqrt{2\pi w_1^2}}\Phi\left\{\alpha\frac{x_0-\left(\mu_1-\alpha_1'\right)}{w_1}\right\}\right]^2 \\
& + \frac{w_2^2}{\sqrt{n_2}}\left(1-\mu_z^2\right)\left[\frac{2e^{\frac{-1}{2w_2^2}\left\{x_0-\left(\mu_2+\alpha_2'\right)\right\}^2}}{\sqrt{2\pi w_2^2}}\Phi\left\{\alpha\frac{x_0-\left(\mu_2-\alpha_2'\right)}{w_2}\right\}\right]^2 \\
\text{SE}\left\{\log(\text{RR})\right\}^2 = {} & \frac{1}{p_1^2}\frac{w_1^2}{\sqrt{n_1}}\left(1-\mu_z^2\right)\left[\frac{2e^{\frac{-1}{2w_1^2}\left\{x_0-\left(\mu_1+\alpha_1'\right)\right\}^2}}{\sqrt{2\pi w_1^2}}\Phi\left\{\alpha\frac{x_0-\left(\mu_1-\alpha_1'\right)}{w_1}\right\}\right]^2 \\
& + \frac{1}{p_2^2}\frac{w_2^2}{\sqrt{n_2}}\left(1-\mu_z^2\right)\left[\frac{2e^{\frac{-1}{2w_2^2}\left\{x_0-\left(\mu_2+\alpha_2'\right)\right\}^2}}{\sqrt{2\pi w_2^2}}\Phi\left\{\alpha\frac{x_0-\left(\mu_2-\alpha_2'\right)}{w_2}\right\}\right]^2 \\
\text{SE}\left\{\log\left(\text{OR}\right)\right\}^2 = {} & \frac{1}{\left\{p_1\left(1-p_1\right)\right\}^2}\frac{w_1^2}{\sqrt{n_1}}\left(1-\mu_z^2\right) \\
& \left[\frac{2e^{\frac{-1}{2w_1^2}\left\{x_0-\left(\mu_1+\alpha_1'\right)\right\}^2}}{\sqrt{2\pi w_1^2}}\Phi\left\{\alpha\frac{x_0-\left(\mu_1-\alpha_1'\right)}{w_1}\right\}\right]^2 \\
& + \frac{1}{\left\{p_2\left(1-p_2\right)\right\}^2}\frac{w_2^2}{\sqrt{n_2}}\left(1-\mu_z^2\right) \\
& \left[\frac{2e^{\frac{-1}{2w_2^2}\left\{x_0-\left(\mu_2+\alpha_2'\right)\right\}^2}}{\sqrt{2\pi w_2^2}}\Phi\left\{\alpha\frac{x_0-\left(\mu_2-\alpha_2'\right)}{w_2}\right\}\right]^2
\end{aligned}$$

2.3 The distributional method for adjusted distributions

Distributional estimates also can be obtained to describe an adjusted difference in means, that is, following a linear regression model of the form

$$Y_i = \beta_0 + \beta_{r_i} + \beta X_i + \epsilon_i$$

where Y is a random variable and ϵ_i is the error term for observation i following a normal distribution with a mean of 0 and a variance of σ_e^2. An exposure is defined by a categorical variable R with $k+1$ levels, for example, not smoking during pregnancy, smoking regularly, or smoking occasionally. We recall how the distributional method can be used in the context of a regression model (see also Sauzet et al. [2016]).

Then, using the marginal means $E(Y|R=r)$ for the $k+1$ levels of exposures, we obtain $k+1$ adjusted distributional probabilities for each level of the exposure $r = 0, 1, \ldots, k$,

$$\begin{aligned} p_r = P(Y < a|R=r) &= P\left\{\epsilon + E(Y|R=r) < a\right\} \\ &= \Phi\left\{\frac{a - E(Y|R=r)}{\sigma_e}\right\} \end{aligned}$$

for a linear regression.

The method can be generalized to mixed models, for example, with the simple random-intercept model with two levels,

$$Y_i = \beta_0 + \beta_{r_i} + \boldsymbol{\beta} X_i + \mu_i + \epsilon_i$$

where $\boldsymbol{\beta}$ is a vector of fixed effects, μ_i is a random element with a mean of 0 and a variance of σ_r^2, and the error term ϵ_i has a variance of σ_e^2. Then,

$$\begin{aligned} p_r = P(Y < a|R=r) &= P\left\{\mu + \epsilon + E(Y|R=r) < a\right\} \\ &= \Phi\left\{\frac{a - E(Y|R=r)}{\sqrt{\sigma_e^2 + \sigma_r^2}}\right\} \end{aligned}$$

The SEs are obtained as seen in section 2.1.

3 The distdicho and distdichoi commands

Because the distributional method is a complement to a comparison of means, the `distdicho` command and its immediate form, `distdichoi`, first return the results of a t test followed by a table containing the relevant information for each group and the distributional estimates for the difference in proportions, the RR, the OR, their SEs, and a confidence interval. The confidence interval is based on the assumption of a normal distribution of the estimate. For small sample sizes, the confidence interval might be too narrow (see Sauzet and Peacock [2014]). Confidence intervals are returned using the current level in the system, which can be modified using the `set level` command.

3.1 Syntax

distdicho *varname1 varname2* [*if*] [*in*], cp(*#*) [twovar tail(lower|upper)
varr(*#*) unequal correction bootci nrep(*#*)]

distdichoi *#obs1 #mean1 #sd1 #obs2 #mean2 #sd2 #cp* [*#varr*
{lower|upper}]

3.2 Options for distdicho

cp(*#*) specifies the cutpoint under which the distributional proportions among the exposed and the nonexposed (reference) are computed using the distributional method described in Peacock et al. (2012) and Sauzet and Peacock (2014). cp() requires a real number. cp() is required.

twovar must be specified if the two variables provided are the outcome values for each group. By default, the first variable provides the outcome values, and the second provides the group categories of exposed and unexposed.

tail(lower|upper) provides the tail of the distribution in which the proportions are to be computed. The default is tail(lower); tail(upper) will provide estimates in the upper tail.

varr(*#*) specifies the number of exposed or unexposed ratios of variances. The default is varr(1).

unequal specifies whether to use a correction for an unknown variance ratio if no assumption can be made about the variance ratio. For the immediate command, this is specified by giving the value 0 for the ratio of variances.

correction specifies for large effect sizes (> 0.7) that a correction factor can be used (valid for difference in proportions only). See Sauzet and Peacock (2014).

bootci specifies to calculate bootstrap bias-corrected confidence intervals instead of distributional ones by using the command bootstrap with 2,000 replications (the default) under the hypotheses that the data are normally distributed and that the variance is known and equal to the data variance for the default version or that the ratio of variances is known for the unequal variance case.

nrep(*#*) specifies the number of bootstrap replications. The default is nrep(2000).

3.3 Options for distdichoi

#varr specifies the number of exposed or unexposed ratios of variances.

lower|upper provides the tail of the distribution in which the proportions are to be computed. The default is lower.

3.4 Examples

Birth weight, body-mass index (BMI), and gestational age are outcomes taken from the St. George's Hospital birth weight study (Peacock, Bland, and Anderson 1995). We consider various group comparisons, including smoking status during pregnancy, first pregnancy (primipari) or second/subsequent pregnancy (multipari), and employment status.

Example 1

This dataset contains the birth weight of 1,772 babies, of which 1,599 were live term births [gestational age (`gest`) greater than or equal to 37 weeks and variable `babycon` equal to 1]. For 1,458 of these births, information about the smoking status of the mother during pregnancy is available.

Live term births are known to be normally distributed (Wilcox 2001), but we can check that is the case here by plotting the outcomes in the two groups of smoking and nonsmoking mothers (see figure 1). We perform the analysis to those births by using the `if` qualifier. The threshold of interest is 2,500 grams, defining low birth weight babies.

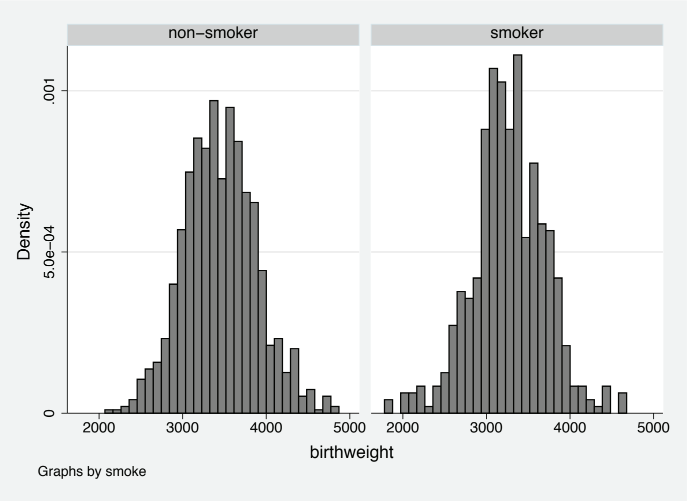

Figure 1. Histogram of birth weights by smokers (`1`) and nonsmokers (`0`)

There is no evidence of unequal variances between smokers and nonsmokers. Therefore, we can apply the simplest form of the distributional method using the cutpoint 2,500 grams to obtain the comparison of proportions of babies whose birth weight is under the cutpoint.

```
. use bwsmoke
. distdicho birthwt smoke if babycon==1 & gest>=37 & gest!=., cp(2500)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
non-smok |     975    3452.728    13.97786    436.4585    3425.298    3480.158
  smoker |     483    3266.965    19.91754     437.733    3227.829    3306.101
---------+--------------------------------------------------------------------
combined |   1,458    3391.189    11.66472    445.4029    3368.308    3414.071
---------+--------------------------------------------------------------------
    diff |            185.7634    24.30893                138.0791    233.4477
------------------------------------------------------------------------------
    diff = mean(non-smok) - mean(smoker)                          t =   7.6418
Ho: diff = 0                                     degrees of freedom =     1456

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000

Distributional estimates for the comparison of proportions
below the cut-point 2500
Standard error computed under the hypothesis that
the ratio of variances is equal to 1
----------------------------------------------------------------
Group      |     Obs      Mean      Std dev.    Dist. prop.
-----------+----------------------------------------------------
non-smok   |     975    3452.728    436.4585     .0146009
smoker     |     483    3266.965     437.733     .0395829
-----------+----------------------------------------------------
Stat       |    Estimate  Std error     [95% Conf. Interval]
-----------+----------------------------------------------------
Diff. prop |    .024982   .0040644       .017016     .032948
Risk ratio |   2.710985   .3496464      2.111901    3.480013
Odds ratio |   2.781502   .3699933      2.150348    3.597909
----------------------------------------------------------------
```

The results show that mothers who smoke have on average babies weighing 185.76 grams less than mothers who do not smoke during pregnancy. This difference, assuming the normality of the outcome, corresponds to a difference in proportions of low birth weight babies of almost 2.5 percentage points (difference in proportions: 0.025) between smoking and nonsmoking mothers with a confidence interval of $[0.017, 0.033]$. The precision of this estimate reflects the precision of the difference in means.

Example 2

The outcome BMI is skewed, but this can be corrected by a transformation. Inverse BMI is reasonably normally distributed. Therefore, we can use the distributional method to compare the proportion of obese mothers at the beginning of pregnancy between primipari and multipari. The proportion of interest is in the upper tail of the distribution of BMIs, but it is in the lower tail of the inverse BMI because inverse is a decreasing function on positive values. The cutpoint must also be transformed and is equal to $1/30 \simeq 0.033$.

```
. use bmi
. distdicho inv_bmi parity, cp(0.033)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   primi |     891    .0443954    .0001971    .0058843    .0440085    .0447823
   multi |     890    .0429524    .0002084    .0062174    .0425434    .0433614
---------+--------------------------------------------------------------------
combined |   1,781    .0436743    .0001444    .0060942    .0433911    .0439575
---------+--------------------------------------------------------------------
    diff |             .001443    .0002869                .0008804    .0020057
------------------------------------------------------------------------------
    diff = mean(primi) - mean(multi)                              t =   5.0304
Ho: diff = 0                                     degrees of freedom =     1779

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000

Distributional estimates for the comparison of proportions
below the cut-point .033
Standard error computed under the hypothesis that
the ratio of variances is equal to 1
--------------------------------------------------------------
Group      |    Obs      Mean       Std dev.    Dist. prop.
-----------+--------------------------------------------------
primi      |     891    .0443954    .0058843     .0298778
multi      |     890    .0429524    .0062174     .0500682
-----------+--------------------------------------------------
Stat       |    Estimate  Std error     [95% Conf. Interval]
-----------+--------------------------------------------------
Diff. prop |  .0201903   .0041399     .0120763    .0283044
Risk ratio |  1.675764     .17357     1.370073    2.049659
Odds ratio |  1.711381   .1846074     1.387752    2.110482
--------------------------------------------------------------
```

While the mean values are difficult to interpret in the original scale, the proportions are not. The distributional method for the dichotomization of normally distributed outcomes shows that the difference in proportions of obesity among multipari mothers is 2 percentage points higher than among primipari mothers. We also can see that the risk of obesity is 1.68 times higher among multipari mothers than among primipari, and the odds of obesity are 1.71 times higher.

Example 3

The proportion of obese mothers can be compared between those who are employed and those who are not. However, the SDs of the inverse BMI cannot be assumed to be equal for employed and unemployed mothers (see Sauzet and Peacock [2014]). If we fail to have any theoretical bases to provide a known variance ratio, we use a correction factor with the `unequal` option.

```
. use bmi2

. distdicho inv_bmi employ, cp(0.033) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
employed |     851    .0438576    .0001936    .0056465    .0434777    .0442375
unemploy |     709    .0433858    .0002427    .0064623    .0429093    .0438623
---------+--------------------------------------------------------------------
combined |   1,560    .0436431    .0001528    .0060336    .0433435    .0439428
---------+--------------------------------------------------------------------
    diff |            .0004718    .0003104               -.0001371    .0010808
------------------------------------------------------------------------------
    diff = mean(employed) - mean(unemploy)                        t =   1.5199
Ho: diff = 0                     Satterthwaite´s degrees of freedom =  1417.45

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9356         Pr(|T| > |t|) = 0.1288          Pr(T > t) = 0.0644

Distributional estimates for the comparison of proportions
below the cut-point .033
Standard error computed with correction for
unknown variance ratio
---------------------------------------------------------------
Group      |      Obs      Mean      Std dev.    Dist. prop.
-----------+---------------------------------------------------
employed   |       851    .0438576    .0056465      .027248
unemploy   |       709    .0433858    .0064623     .0540131
-----------+---------------------------------------------------
Stat       |     Estimate  Std error     [95% Conf. Interval]
-----------+---------------------------------------------------
Diff. prop |   .0267651   .0076436      .011784    .0417462
Risk ratio |   1.982276   .3341296     1.434148    2.739898
Odds ratio |   2.038361   .3536209     1.461411    2.843085
---------------------------------------------------------------
```

The distributional method for the dichotomization of normally distributed outcomes shows that the difference in proportions of obesity among unemployed mothers is 2.7 percentage points higher than among employed mothers. It also shows that the risk of obesity (RR) is almost twice as high among unemployed versus employed mothers, who have an RR almost equal to the odds of obesity (OR).

Example 4

If, on the contrary, we have reason to assume that the ratio of the variance of unemployed to employed is 1.3, then the comparisons of proportions are obtained using this value and no correction factor is needed:

```
. use bmi2

. distdicho inv_bmi employ, cp(0.033) varr(1.3)

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
employed |     851    .0438576    .0001936    .0056465    .0434777    .0442375
unemploy |     709    .0433858    .0002427    .0064623    .0429093    .0438623
---------+--------------------------------------------------------------------
combined |   1,560    .0436431    .0001528    .0060336    .0433435    .0439428
---------+--------------------------------------------------------------------
    diff |            .0004718    .0003104               -.0001371    .0010808
------------------------------------------------------------------------------
    diff = mean(employed) - mean(unemploy)                        t =   1.5199
Ho: diff = 0                     Satterthwaite´s degrees of freedom =  1417.45

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9356         Pr(|T| > |t|) = 0.1288          Pr(T > t) = 0.0644

Distributional estimates for the comparison of proportions
below the cut-point .033
Standard error computed under the hypothesis that
the ratio of variances is equal to 1.3
--------------------------------------------------------------
Group      |    Obs      Mean      Std dev.    Dist. prop.
-----------+--------------------------------------------------
employed   |     851    .0438576    .0056465     .0274554
unemploy   |     709    .0433858    .0064623     .0536526
-----------+--------------------------------------------------
Stat       |   Estimate  Std error     [95% Conf. Interval]
-----------+--------------------------------------------------
Diff. prop |  .0261972   .0046343      .0171142    .0352803
Risk ratio |  1.954174   .2165354      1.575774    2.423441
Odds ratio |   2.00827   .2320762      1.604792    2.513191
--------------------------------------------------------------
```

The known value for the ratio of variances we used is the observed one. Therefore, the estimates obtained in examples 3 and 4 are similar. However, because we have been more conservative when we did not assume we knew the variance ratio, the SEs are larger in example 3 than in example 4.

3.5 Stored results

`distdicho` and `distdichoi` store the following in `r()`:

Scalars

`r(prop1)`	distributional proportion estimate for group at risk
`r(prop2)`	distributional proportion estimate for reference group
`r(propdiff)`	distributional estimate for difference in proportions between group at risk and reference group
`r(distrr)`	distributional estimate for RR between group at risk and reference group
`r(distor)`	distributional estimate for OR between group at risk and reference group
`r(sediff)`	SE for distributional estimate of the difference in proportion
`r(serr)`	SE for distributional estimate of the RR
`r(seor)`	SE for distributional estimate of the OR
`r(ciinf)`	difference in proportions: lower limit of confidence interval
`r(cisup)`	difference in proportions: upper limit of confidence interval
`r(ciinfrr)`	RR: lower limit of confidence interval
`r(cisuprr)`	RR: upper limit of confidence interval
`r(ciinfor)`	OR: lower limit of confidence interval
`r(cisupor)`	OR: upper limit of confidence interval

4 The sk_distdicho and sk_distdichoi commands

We now discuss the skew normal version of `distdicho`. The `sk_distdicho` command has the same syntax as the `distdicho` command but without a method for unequal variance.

4.1 Syntax

`sk_distdicho` *varname1 varname2* [*if*] [*in*], `cp(`*#*`)` [`twovar`
`tail(lower|upper)` `bootci` `nrep(`*#*`)`]

`sk_distdichoi` *#obs1 #mean1 #sd1 #obs2 #mean2 #sd2 #cp* [{`lower`|`upper`}
#alpha]

4.2 Options for sk_distdicho

`cp(`*#*`)` specifies the cutpoint under which the distributional proportions among the exposed and the nonexposed (reference) are computed using the distributional method described in Peacock et al. (2012). `cp()` requires a real number. `cp()` is required.

`twovar` must be specified if the two variables provided are the outcome values for each group. By default, the first variable provides the outcome values, and the second provides the group categories of exposed and unexposed.

`tail(lower|upper)` provides the tail of the distribution in which the proportions are to be computed. The default is `tail(lower)`; `tail(upper)` will provide estimates in the upper tail.

`bootci` specifies to calculate bootstrap bias-corrected confidence intervals instead of distributional ones by using the command `bootstrap` with 2,000 replications (the default) under the hypotheses that the data are normally distributed and that the variance is known and equal to the data variance for the default version or that the ratio of variances is known for the unequal variance case.

`nrep(#)` specifies the number of bootstrap replications. The default is `nrep(2000)`.

4.3 Options for sk_distdicho

`lower|upper` provides the tail of the distribution in which the proportions are to be computed. The default is `lower`.

#alpha specifies the skew-normal alpha coefficient.

4.4 Examples

Example 5

In the following example, we show that the two commands `sk_distdicho` and `distdicho` give similar results for the difference in proportions when the data are approximately normally distributed. We reproduce example 1 but with the command `sk_distdicho` instead of `distdicho`.

```
. use bwsmoke
. sk_distdicho birthwt smoke if babycon==1 &gest>=37 & gest!=., cp(2500)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
non-smok |     975    3452.728    13.97786    436.4585    3425.298    3480.158
  smoker |     483    3266.965    19.91754     437.733    3227.829    3306.101
---------+--------------------------------------------------------------------
combined |   1,458    3391.189    11.66472    445.4029    3368.308    3414.071
---------+--------------------------------------------------------------------
    diff |            185.7634    24.30893                138.0791    233.4477
------------------------------------------------------------------------------
    diff = mean(non-smok) - mean(smoker)                          t =   7.6418
Ho: diff = 0                                     degrees of freedom =     1456

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
Distributional estimates for the comparison of proportions
below the cut-point 2500
Alpha:  .86689235
--------------------------------------------------------------
Group      |    Obs      Mean       Std dev.    Dist. prop.
-----------+--------------------------------------------------
smoker     |     483    3266.965     437.733     .0365651
non-smok   |     975    3452.728    436.4585     .0124953
-----------+--------------------------------------------------
Stat       |    Estimate  Std error     [95% Conf. Interval]
-----------+--------------------------------------------------
Diff. prop |   .0240698   .0041428        .01595    .0321896
Risk ratio |   2.926313   .4872841      2.125134    4.029538
Odds ratio |   2.999422   .5115186       2.16213    4.160959
--------------------------------------------------------------
```

The estimates and SEs obtained here and in example 1 are almost identical for the difference in proportions, even if the estimated skew parameter α is not close to 0. This shows that the distributional method is robust to small variations to normality. However, because the estimated proportions for each group vary between example 1 and example 5, the RR and OR also vary between these two examples.

Example 6

In example 2, we used a transformation to obtain a normally distributed outcome. We use the same data to compare the skew-normal approach with the transformation approach. Note that now the proportion of interest (obesity) is in the upper tail of the distribution.

```
. use bmi
. sk_distdicho bmi parity, cp(30) tail(upper)
Two-sample t test with equal variances
───────────────────────────────────────────────────────────────────────────────
   Group │     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
─────────┼─────────────────────────────────────────────────────────────────────
   primi │     891    22.96176    .1135206    3.388547    22.73896    23.18456
   multi │     890    23.84148    .1345053    4.012678    23.57749    24.10546
─────────┼─────────────────────────────────────────────────────────────────────
combined │   1,781    23.40137    .0885863    3.738509    23.22763    23.57512
─────────┼─────────────────────────────────────────────────────────────────────
    diff │           -.8797151    .1759908               -1.224886   -.5345447
───────────────────────────────────────────────────────────────────────────────
    diff = mean(primi) - mean(multi)                              t =  -4.9986
Ho: diff = 0                                     degrees of freedom =     1779

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
Distributional estimates for the comparison of proportions
above the cut-point 30
Alpha:  4.1193072
───────────────────────────────────────────────────────────────
Group      │   Obs      Mean       Std dev.    Dist. prop.
───────────┼───────────────────────────────────────────────────
multi      │    890    23.84148    4.012678     .0683555
primi      │    891    22.96176    3.388547     .0485803
───────────┼───────────────────────────────────────────────────
Stat       │   Estimate  Std error     [95% Conf. Interval]
───────────┼───────────────────────────────────────────────────
Diff. prop │  .0197752   .0040157     .0119046    .0276457
Risk ratio │  1.407061   .0965391     1.230598    1.608829
Odds ratio │  1.436928   .1047111     1.246382    1.656604
───────────────────────────────────────────────────────────────
```

The estimates obtained here and in example 2 are very close because the transformation used in example 2 was quite successful in providing an approximately normal distribution. We still have, for example, a difference of about 2 percentage points in proportions of obesity between multipari and primipari mothers. However, the SEs for these estimates are smaller using the skew-normal method.

4.5 Stored results

sk_distdicho stores the same results as those stored by distdicho with the following also stored in r():

Scalars

r(alpha)	estimated skew-normal alpha coefficient

5 The reg_distdicho command

The command reg_distdicho uses the stored results of the command regress, mixed, or xtreg to provide distributional estimates of adjusted comparisons of proportion between the reference level of a factor and the other levels of this factor. The reference level must be coded with the lowest value.

5.1 Syntax

reg_distdicho *varname*, cp(#) [tail(lower|upper) dist(sk)]

5.2 Options

Only the following option is specific to the reg_distdicho command. For the other options, see the distdicho command. Because reg_distdicho uses stored results from a regression model, there is no option for bootstrap confidence intervals.

dist(sk) specifies to use the skew-normal method if there remains a perturbation to the normal distribution. The default is that the residuals are assumed normally distributed.

5.3 Examples

Example 7

Example 1 is revisited again, but this time we would like an estimate of proportion comparison adjusted for gestational age.

```
. use bwsmoke
. regress birthwt i.smoke gest if babycon==1
      Source |       SS           df       MS      Number of obs   =     1,578
-------------+----------------------------------   F(2, 1575)      =    502.13
       Model |   175438127         2  87719063.6   Prob > F        =    0.0000
    Residual |   275142224     1,575  174693.476   R-squared       =    0.3894
-------------+----------------------------------   Adj R-squared   =    0.3886
       Total |   450580352     1,577  285719.944   Root MSE        =    417.96

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |
     smoker  |  -164.5144   22.40716    -7.34   0.000    -208.4654   -120.5634
        gest |   155.4258   5.051078    30.77   0.000     145.5182    165.3333
       _cons |  -2760.235     199.78   -13.82   0.000    -3152.098   -2368.373
------------------------------------------------------------------------------
```

```
. reg_distdicho smoke, cp(2500)
Comparisons of proportions based on marginal effects of
regress birthwt i.smoke gest if babycon==1
Distributional estimates for the comparison of proportions below the cut-point 2500
-----------------------------------------------------------
Group      |    Obs      Mean      Std dev.    Dist. prop.
-----------------------------------------------------------
1          |    1060    3372.722    417.9635     .0183974
2          |     518    3208.208    417.9635     .0450923
-----------------------------------------------------------
Stat       |    Estimate  Std error     [95% Conf. Interval]
-----------------------------------------------------------
Diff. prop |  .0266949   .0043955     .0194649     .033925
Risk ratio |  2.451019   .2954949     2.014368    2.982321
Odds ratio |  2.519538   .3149271      2.05612    3.087403
-----------------------------------------------------------
```

The adjusted difference in means of low birth weight babies between smoking and nonsmoking mothers is smaller than in example 1, but the corresponding difference in proportions (2.7% compared with 2.5%) is larger due to a different position of the proportions of the two groups.

Example 8

The final example uses `smoking.dta` from the book *Multilevel and Longitudinal Modeling Using Stata* (Rabe-Hesketh and Skrondal 2012). In the multilevel model, babies are the first level and mothers are the second level.

```
. use http://www.stata-press.com/data/mlmus3/smoking
. mixed birwt i.smoke mage year || momid:
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0:   log likelihood = -65291.849
Iteration 1:   log likelihood = -65291.845
Computing standard errors:
Mixed-effects ML regression                     Number of obs     =      8,604
Group variable: momid                           Number of groups  =      3,978
                                                Obs per group:
                                                              min =          2
                                                              avg =        2.2
                                                              max =          3
                                                Wald chi2(3)      =     381.36
Log likelihood = -65291.845                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |
     Smoker  |  -254.4345   17.51951   -14.52   0.000    -288.7721   -220.0969
        mage |   10.39172   1.279693     8.12   0.000     7.883567    12.89987
        year |   12.96842   3.073012     4.22   0.000     6.945428    18.99141
       _cons |   3178.528   35.82147    88.73   0.000     3108.319    3248.736
------------------------------------------------------------------------------
```

```
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Identity              |
                  var(_cons) |   120155.2   4429.523      111779.7    129158.2
-----------------------------+------------------------------------------------
               var(Residual) |   141423.3   2949.447      135759.1    147323.9
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 1134.56       Prob >= chibar2 = 0.0000
. reg_distdicho smoke, cp(2500)
Comparisons of proportions based on marginal effects of
mixed birwt i.smoke mage year || momid:
Distributional estimates for the comparison of proportions below the cut-point 2500
------------------------------------------------------------
Group      |    Obs      Mean      Std dev.    Dist. prop.
-----------+------------------------------------------------
0          |    7400    3504.997    511.4475     .0247069
1          |    1204    3250.562    511.4475     .0711166
-----------+------------------------------------------------
Stat       |    Estimate  Std error     [95% Conf. Interval]
-----------+------------------------------------------------
Diff. prop |   .0464098   .0039742     .0398727    .0529468
Risk ratio |   2.878416   .1773509      2.60174    3.184514
Odds ratio |    3.02223   .1987087      2.71338    3.366234
------------------------------------------------------------
```

In this dataset, the mean difference in birth weight between smoking and nonsmoking mothers (254 grams) adjusted for age of mother and year of birth as well as the nonindependence of siblings in multiple births is much larger than the one obtained in the dataset used in the previous examples. There was no adjustment for gestational age because that information is not available. This mean difference corresponds to 4.6 percentage points more low birth weight babies among the smoking mothers than among the nonsmoking mothers (95% confidence interval $[0.040, 0.053]$).

5.4 Stored results

reg_distdicho stores the same results as those stored by distdicho. Results are stored only if there are two levels of risks.

6 Conclusion

The commands available in the package distdicho make the distributional method for the dichotomization of continuous outcomes easily accessible either for simple comparison following a t test or to obtain adjusted comparisons. Thus effects obtained on mean comparison can also be presented as comparison of proportions to increase the understanding of the study results in terms of population at risk.

7 Acknowledgments

We acknowledge the financial contribution granted for this work from the Research Centre for Mathematical Modelling at Bielefeld University.

We would like to thank the reviewer for his or her helpful comments and recommendations, which helped improve this article as well as the commands.

8 References

Azzalini, A. 2005. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics* 32: 159–188.

Peacock, J. L., J. M. Bland, and H. R. Anderson. 1995. Preterm delivery: Effects of socioeconomic factors, psychological stress, smoking, alcohol, and caffeine. *British Medical Journal* 311: 531–536.

Peacock, J. L., O. Sauzet, S. M. Ewings, and S. M. Kerry. 2012. Dichotomising continuous data while retaining statistical power using a distributional approach. *Statistics in Medicine* 31: 3089–3103.

Rabe-Hesketh, S., and A. Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata. Volume II: Categorical Responses, Counts, and Survival.* 3rd ed. College Station, TX: Stata Press.

Ragland, D. R. 1992. Dichotomizing continuous outcome variables: Dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology* 3: 434–440.

Royston, P., D. G. Altman, and W. Sauerbrei. 2006. Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine* 25: 127–141.

Sauzet, O., J. Breckenkamp, T. Borde, S. Brenne, M. David, O. Razum, and J. L. Peacock. 2016. A distributional approach to obtain adjusted comparisons of proportions of a population at risk. *Emerging Themes in Epidemiology* 13(8): 1–10.

Sauzet, O., M. Ofuya, and J. L. Peacock. 2015. Dichotomisation using a distributional approach when the outcome is skewed. *BMC Medical Research Methodology* 15: 40.

Sauzet, O., and J. L. Peacock. 2014. Estimating dichotomised outcomes in two groups with unequal variances: A distributional approach. *Statistics in Medicine* 33: 4547–4559.

Wilcox, A. J. 2001. On the importance—and the unimportance—of birthweight. *International Journal of Epidemiology* 30: 1233–1241.

About the authors

Odile Sauzet is a senior research fellow in biostatistics in the Bielefeld School of Public Health at Bielefeld University in Germany. Her research interests concern the development of statistical methods in epidemiology and public health.

Maren Klein is a statistician still pursuing her studies in biological informatics.