The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–18

WORKING PAPER Distributional estimates for the comparison of proportions of a dichotomised continuous outcome

Odile Sauzet
Department of Epidemiology & International Public Health,
Bielefeld School of Public Health (BiSPH),
Bielefeld University,
Bielefeld, Germany
odile.sauzet@uni-bielefeld.de
Maren Kleine
Department of Epidemiology & International Public Health,
Bielefeld School of Public Health (BiSPH),
Bielefeld University,
Bielefeld, Germany
maren.kleine@uni-bielefeld.de

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

Keywords: st0001, distdicho, reg_distdicho, sk_distdicho, distributional method, dichotomisation, continuous outcomes

1 Introduction and motivation

Dichotomisation of continuous outcomes is a common practice despite numerous arguments against it (Ragland (1992); Royston et al. (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 dichotomisation 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)). Due to 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)). In Sauzet et al. (2015b) the question of the robustness to deviations from normality has been addressed and showed that for small deviations the method worked well. In case of perturbation to the normal distribution (e.g. because of an excess of patients with hight blood pressure or preterm babies having much lower birthweights) a method based on the skew-normal distribution (Azzalini (2005)) has

also been proposed in Sauzet et al. (2015b).

Because most analysis comparing continuous outcomes between two groups are not performed with a t-test but with potentially complex regression models, with the distributional method adjusted comparisons of proportions can also be be obtained to reflect the results of linear possibly mixed models (Sauzet et al. (2015a)).

A module of Stata commands has been developed to cover all the applications of the distributional methods which have been developed so far. In the following we use examples to illustrate the usage of the various commands and options in the module.

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 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 formulae obtained to compute estimates and standard errors for proportions, difference in proportions, risk ratios and odds ratios derived from the normal distribution.

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

$$p(\overline{X}_n) = \int_{-\infty}^{x_0} f_{N(\overline{X}_n,\sigma^2)}(t)dt \tag{1}$$

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

$$\mathrm{sd}(p(\overline{X}_n)) = \frac{s}{\sqrt{n}} f_{N(\overline{x}_n,s^2)}(x_0)$$

so the estimate for the proportion under the quantile x_o is estimated by $\int_{-\infty}^{x_0} f_{N(\overline{x}_n,s^2)}(t)dt$ with standard error $\frac{s}{\sqrt{n}} f_{N(\overline{x}_n,s^2)}(x_0)$ where s the sample estimate for the standard deviation assumed to be the known standard deviation 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 standard error using for the common standard deviation the pooled estimate from the data :

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

$$\text{se}(d) = \sqrt{\frac{s^2_{pooled}}{n_t} f^2_{N(\overline{x}_{t,n_t}, s^2_{pooled})}(x_0) + \frac{s^2_{pooled}}{n_c} f^2_{N(\overline{x}_{c,n_c}, s^2_{pooled})}(x_0)} \tag{3}$$

Estimates for the standard error for the log risk ratio $\log(rr)$ is obtained through the function $h(\overline{X}_n) = \log(p(\overline{X}_n))$. The standard error for the log risk ratio is

$$\text{se}(\log(rr)) = \sqrt{\frac{s^2_{pooled}}{n_t} \frac{f^2_{N(\overline{x}_{t,n_t} s^2_{pooled})}(x_o)}{p_t^2} + \frac{s^2_{pooled}}{n_c} \frac{f^2_{N(\overline{x}_{c,n_c} s^2_{pooled})}(x_o)}{p_c^2}} \tag{4}$$

Estimates for the standard error for the log odds ratio is obtained through the function $g(\overline{X}_n)) = \log(\frac{p(\overline{X}_n)}{1-p(\overline{X}_n)})$. The standard error for the log odds ratio is

$$\text{se}(\log(or)) = \sqrt{\frac{s^2_{pooled}}{n_c} \frac{f^2_{N(\overline{x}_{c,n_c}, s^2_{pooled})}(x_o)}{p_c^2(1-p_c)^2} + \frac{s^2_{pooled}}{n_t} \frac{f^2_{N(\overline{x}_{t,n_t}, s^2_{pooled})}(x_o)}{p_t^2(1-p_t)^2}} \tag{5}$$

The equal variance condition can be relaxed by either providing a known ratio of variances between the two groups or when this is not possible by adding a correction factor to the standard error which otherwise would be underestimated when the variances are not assumed known. Moreover this correction factor can also be used to correct the standard errors for large effects (see Sauzet and Peacock (2014)) as the variability due to using the observed pooled standard deviation need to be accounted for in the standard error whether the variances are assumed equal of 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 generalisation of the normal distribution which works by adding a third parameter α defining the skewness ($\alpha = 0$ gives the normal distribution). We briefly recall how the formula for the standard errors are obtained (Sauzet et al. (2015b)).

Lets $\overline{X}_n$ be the sample mean of n independent identically skew-normal distributed random variables $X_i,\ i = 1...n$ with mean μ, variance σ^2 and skewness parameter α.

Lets x_0 be a threshold of interest. The random variable $p(\overline{X}_n)$ for the proportion of the population with outcome value under the threshold x_0 is defined as

$$p(\overline{X}_n) = \int_{-\infty}^{x_0} 2\frac{e^{\frac{-1}{2w^2}(t-(\overline{X}_n+\alpha'))^2}}{\sqrt{2\pi w^2}} \left(\int_{-\infty}^{\alpha(t-(\overline{X}_n+\alpha'))/w} \frac{e^{\frac{-1}{2}r^2}}{\sqrt{2\pi}} dr \right) dt \tag{6}$$

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

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

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

The formula for $p'(\mu)$ was derived in Sauzet et al. (2015b) where we obtained, Φ being the standard normal cumulative distribution function:

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

The formulae for the standard error for the difference in proportions d, log risk ration $log(rr)$ and log odds ration $log(or)$ follow:

$$se(d)^2 = \frac{w_1^2}{\sqrt{n_1}}\left(1-\mu_z^2\right)\left(\frac{2e^{\frac{-1}{2w_1^2}(x_0-(\mu_1+\alpha_1'))^2}}{\sqrt{2\pi w_1^2}}\Phi\left(\alpha\frac{x_0-(\mu_1-\alpha_1')}{w_1}\right)\right)^2 +$$
$$\frac{w_2^2}{\sqrt{n_2}}\left(1-\mu_z^2\right)\left(\frac{2e^{\frac{-1}{2w_2^2}(x_0-(\mu_2+\alpha_2'))^2}}{\sqrt{2\pi w_2^2}}\Phi\left(\alpha\frac{x_0-(\mu_2-\alpha_2')}{w_2}\right)\right)^2$$

$$se(log(rr))^2 = \frac{1}{p_1^2}\frac{w_1^2}{\sqrt{n_1}}\left(1-\mu_z^2\right)\left(\frac{2e^{\frac{-1}{2w_1^2}(x_0-(\mu_1+\alpha_1'))^2}}{\sqrt{2\pi w_1^2}}\Phi\left(\alpha\frac{x_0-(\mu_1-\alpha_1')}{w_1}\right)\right)^2 +$$
$$\frac{1}{p_2^2}\frac{w_2^2}{\sqrt{n_2}}\left(1-\mu_z^2\right)\left(\frac{2e^{\frac{-1}{2w_2^2}(x_0-(\mu_2+\alpha_2'))^2}}{\sqrt{2\pi w_2^2}}\Phi\left(\alpha\frac{x_0-(\mu_2-\alpha_2')}{w_2}\right)\right)^2$$

$$se(log(or))^2 = \frac{1}{(p_1(1-p_1))^2}\frac{w_1^2}{\sqrt{n_1}}\left(1-\mu_z^2\right)\left(\frac{2e^{\frac{-1}{2w_1^2}(x_0-(\mu_1+\alpha_1'))^2}}{\sqrt{2\pi w_1^2}}\Phi\left(\alpha\frac{x_0-(\mu_1-\alpha_1')}{w_1}\right)\right)^2 +$$
$$\frac{1}{(p_2(1-p_2))^2}\frac{w_2^2}{\sqrt{n_2}}\left(1-\mu_z^2\right)\left(\frac{2e^{\frac{-1}{2w_2^2}(x_0-(\mu_2+\alpha_2'))^2}}{\sqrt{2\pi w_2^2}}\Phi\left(\alpha\frac{x_0-(\mu_2-\alpha_2')}{w_2}\right)\right)^2$$

2.3 The distributional method for adjusted distributions

Distributional estimates can also be obtained to describe an adjusted difference in means, i.e. 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 variance σ_e^2. An exposure is defined by a categorical variable R with $k+1$ levels, e.g. not smoking during pregnancy, smoking regularly, smoking occasionally. We recall how the distributional method can be used in the context of a regression model (see also Sauzet et al. (2015a))

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 $r = 0, 1, .., k$,

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

for a linear regression.

The method can be generalised to mixed models for example with simple random intercept model with two levels

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

where β is a vector of fixed effects and μ_i a random element with mean zero and a variance σ_r^2 and the error term ϵ_i with variance σ_e^2. Then:

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

The standard errors 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 returns the results of a t-test followed by a table containing the relevant information for each groups and the distributional estimates for difference in proportions, risk ratio and odds ratio, their standard error 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 command `set level`.

3.1 Syntax

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

distdichoi #obs1 #mean1 #sd1 #obs2 #mean2 #sd2 #cp [ upper, #varr]
```

3.2 Options

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

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

`tail(`*upper*`)` provides the tail of the distribution in which the proportions are to be computed. The default is the lower tail, `tail(upper)` will provide estimates in the upper tail. For the immediate command *upper* must be specified to obtain estimate in the upper tail.

`varr(#)` by default the the ratio of variances exposed/unexposed is assumed to be one. Other assumed ratios can be specified with the option `varr`.

`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 the this is specified by giving the value 0 for the ratio of variances.

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

`bootci` bootstrap bias corrected confidence intervals are calculated instead of distributional ones using the command `bootstrap` with 2000 (default) replicate under the hypotheses for the variance specified by the command line.

`nrep(#)` the number of bootstrap replicates can be altered. The default value is 2000.

3.3 Examples

Birthweight, body-mass index and gestational age are outcomes taken from the St George's Birthweight Study (Peacock et al. (1995)). We consider various group comparisons including the smoking status during pregnancy yes (1)/no (0), parity first (primipari) (0) /second or subsequent (multipari) pregnancy (1), employed (1)/unemployed (2).

Example 1

This first dataset contains the birthweight of 1772 births of which 1599 are live term birth (gestational age (*gest*) greater or equal to 37 weeks and the variable *babycon* is equal to 1). For 1458 of this birth, information about the smoking status of the mother during pregnancy is available. Live term birth are known to be normally distributed (Wilcox (2001)) but we can check that it is the case here by plotting the outcomes in the two groups of smoking and non smoking mothers (see Fig. 1). We are therefore performing the analysis to those birth by using the `if` qualifier . The threshold of interest is 2500 g defining low-birthweight babies.

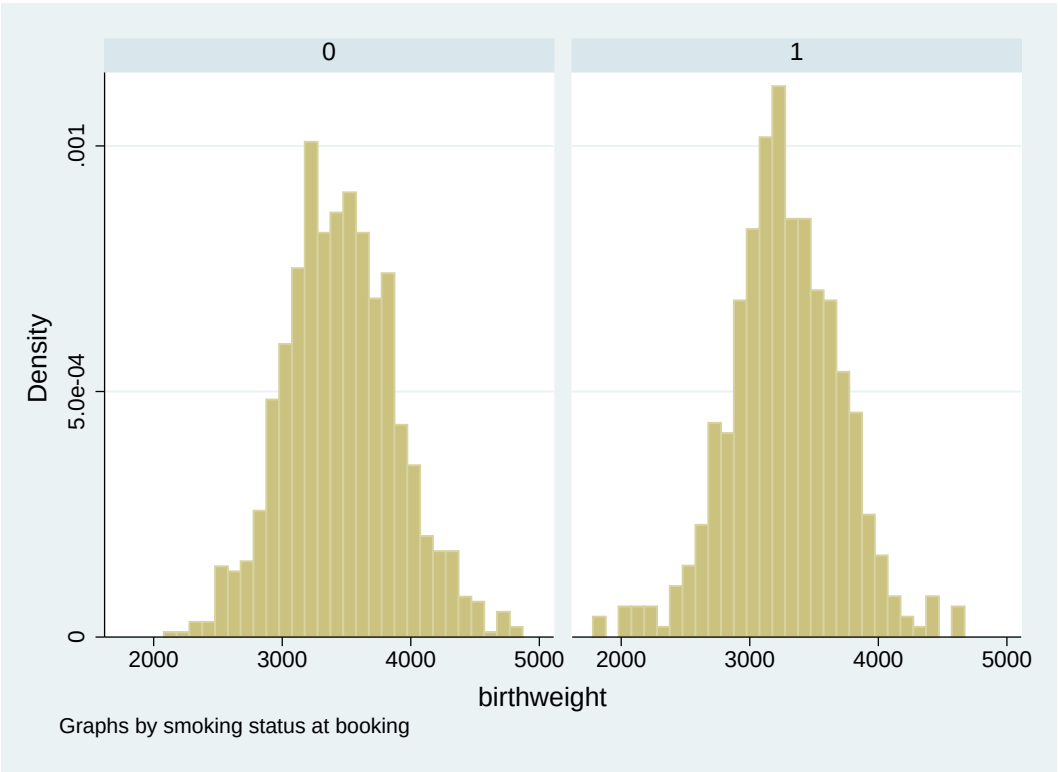

Figure 1: Histogram of birthweights by smokers(=1) and non-smokers(=0)

There is no evidence of unequal variances between smokers and non-smokers, therefore we can apply the simplest form of the distributional method using the cut-point 2500g to obtain the comparison of proportions of babies whose birthweight is under the cut-point.

```
. 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	1458	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 g less than mothers who don't smoke during pregnancy. This difference, assuming the normality of the outcome, corresponds to a difference in proportions of low birthweight babies of almost 2.5 percentage points (difference in proportions: 0.025) between smoking and non-smoking mothers with a confidence interval of [0.017, 0.033]. The precision of this estimates 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 begining of pregnancy between primi 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 cut-point also need to 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 |      1781    .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. Err.       [95% Conf. Interval]
----------------+-------------------------------------------------------------
primi           |      891    .0443954    .0058843      .0298778    .0500682
multi           |      890    .0429524    .0062174      .0500682    .0298778
----------------+-------------------------------------------------------------
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 dichotomisation of normally distributed outcomes shows that the difference in proportions of obesity among multipari mothers is 2 percentage points higher than among primipari mother. Also we 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 higher.

Example 3

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

```
. use  bmi2
. distdicho inv_bmi employ, cp(0.033) uneq
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	1560	.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 dichotomisation of normally distributed outcomes shows that the difference in proportions of obesity among unemployed mothers is 2.7 percentage points higher than among employed mother. It shows also that the risk of obesity (risk ratio) is almost twice among unemployed than among employed mothers almost equal to the odds of obesity (odds ratio)

Example 4

If on the contrary we have reasons to assume that the ratio of variance unemployed/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	1560	.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 example 3 and 4 are the similar. However because we have been more conservative when we did not assume we knew the variance ratio, the standard errors are larger in example 3 than in example 4.

3.4 Saved results

`distdicho` and `distdichoi` save in `r()`

Scalars

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

4 The sk_distdicho and sk_distdichoi commands

The `sk_distdicho` command has the same syntax as the `distdicho` command also with an option to obtain bootstrap confidence intervals but has no method for unequal variance.

4.1 Syntax

`sk_distdicho` *varname*1 *varname*2 [*if*] [*in*] [, twovar tail(lower upper) bootci nrep(#)] cp(#)

`sk_distdichoi` #obs1 #mean1 #sd1 #obs2 #mean2 #sd2 #cp (*upper lower*)tail #α

4.2 Options

`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 the group categories: exposed, unexposed.

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

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

`bootci` bootstrap bias corrected confidence intervals are calculated instead of distributional ones using the command `bootstrap` with 2000 (default) replicate under the hypotheses for the variance specified by the command line.

`nrep(#)` The number of bootstrap replicates can be altered. The default value is 2000.

4.3 Examples

Example 5

In the following example we show that two commands `sk_distdicho` and `distdicho` give similar results for the difference in proportions when the data is approximatively normally distributed. We reproduce Example 1 using 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 |     1458    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 standard errors 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 groups vary between Example 1 and Example 5, the risk ratios and odds ratios also vary between this 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 to the transformation approach. Note that now the proportions 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	1781	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:  4.1193066
```

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 approximatively normal distribution. We still have for example a difference from about 2 percentage points in proportions of obesity between multi and primipari mothers. However the standard error for these estimates are smaller using the skew-normal method.

4.4 Saved results

The results saved by the the command `sk_distdicho` are the same as the ones saved by the `distdicho` command with the following also stored in `r()`

Scalars
`r(alpha)` the estimate skew-normal alpha coefficient

5 The `reg_distdicho` command

The command `reg_distdicho` uses the stored result of the commands `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 needs to be coded with the lowest value.

`reg_distdicho` *varname*`1` [*if*] [*in*] [`, tail(upper) dist(sk)`] `cp(`#`)`

5.1 Options

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

`dist` The default is that the errors in the regression model are assumed normally

distributed. If there remain a perturbation to the normal distribution, the skew-normal method can be used with the option dist(sk)

5.2 Examples

Example 7

Example 1 is revisited again but we would like 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	= 1578
				F(2, 1575)	= 502.13
Model	175438127	2	87719063.6	Prob > F	= 0.0000
Residual	275142224	1575	174693.476	R-squared	= 0.3894
				Adj R-squared	= 0.3886
Total	450580352	1577	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 birthweight babies between smoking and non-smoking mothers is smaller than in Example 1 but the corresponding difference in proportions (2.7 % compared to 2.5%) is larger due to a different position of the proportions of the two groups.

Example 8

The last example uses the dataset smoking.dta which include multiple births from

the book Multilevel and Longitudinal Modeling Using Stata Rabe-Hesketh and Skrondal (2008). In the multilevel model babies are the first level and the mother the second level.

```
.use  http://www.stata-press.com/data/mlmus2/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      =      8604
Group variable: momid                           Number of groups   =      3978
                                                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 regression: 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 || momi d:
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 birthweight between smoking and non-smoking mothers (254 g) adjusted for age of mother and year of birth as well as the non-

independence of siblings in multiple birth is much larger that the one obtained in the dataset used in the previous examples. There was no adjustment for gestational age because the information is not available. This mean difference corresponds to 4.6 percentage points more low birthweight babie among the smoking mothers than among the non-smoking mothers (95% confidence inteval [0.040, 0.053]).

5.3 Saved results

The results saved by the the command `reg_distdicho` are the identical to the ones saved by the `distdicho` command. There are saved results only if there are two levels of risks.

6 Conclusion

The functions available in the package `DistDicho` make the distributional method for the dichotomisation 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 proportion to increase the understanding of the study results in terms of population at risk.

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

The authors would like to thanks the reviewer for his/her very helpful comments and recommendations which helped improving this paper as well as the Stata commands.

7 References

Azzalini, A. 2005. The Skew-normal Distribution and Related Multivariate Families. *Scandinavian Journal of Statistics* 32(2): 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. *BMJ* 311(7004): 531–535.

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(26): 3089–3103.

Rabe-Hesketh, S., and A. Skrondal. 2008. *Multilevel and longitudinal modeling using stata*. 2nd ed. College Station and Tex.: Stata Press Publication.

Ragland, D. R. 1992. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology (Cambridge, Mass.)* 3(5): 434–440.

Royston, P., D. G. Altman, and W. Sauerbrei. 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 25(1): 127–141.

Sauzet, O., J. Breckenkamp, T. Borde, S. Brenne, M. David, R. Razum, and J. L. Peacock. 2015a. A distributional approach to obtain adjusted comparisons of proportions of a population at risk. *Under review* .

Sauzet, O., M. Ofuya, and J. L. Peacock. 2015b. 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(26): 4547–4559.

Wilcox, A. J. 2001. On the importance - and the unimportance - of birthweight. *International Journal of Epidemiology* 30(6): 1233–1241.