

**Inskew0** — Find zero-skewness log or Box–Cox transform[Description](#)[Quick start](#)[Menu](#)[Syntax](#)[Options](#)[Remarks and examples](#)[Stored results](#)[Methods and formulas](#)[Acknowledgment](#)[Reference](#)[Also see](#)

## Description

`lnskew0` creates  $newvar = \ln(\pm exp - k)$ , choosing  $k$  and the sign of  $exp$  so that the skewness of  $newvar$  is zero.

`bcskew0` creates  $newvar = (exp^\lambda - 1)/\lambda$ , the Box–Cox power transformation (Box and Cox 1964), choosing  $\lambda$  so that the skewness of  $newvar$  is zero.  $exp$  must be strictly positive.

## Quick start

Generate `newv1`, the zero-skewness log transform of continuous variable `v1`

```
lnskew0 newv1 = v1
```

Same as above, but transform ratio of `v1` to `v2`

```
lnskew0 newv1 = v1/v2
```

Zero-skewness Box–Cox transform, `newv2`, of `v2`

```
bcskew0 newv2 = v2
```

Same as above, and change the value for convergence to 0.0001 from the default 0.001

```
bcskew0 newv2 = v2, zero(.0001)
```

## Menu

### **Inskew0**

Data > Create or change data > Other variable-creation commands > Zero-skewness log transform

### **bcskew0**

Data > Create or change data > Other variable-creation commands > Box–Cox transform

## Syntax

*Zero-skewness log transform*

```
lnskew0 newvar = exp [if] [in] [, options]
```

*Zero-skewness Box–Cox transform*

```
bcskew0 newvar = exp [if] [in] [, options]
```

*options*

Description

Main

<code>delta(#)</code>	increment for derivative of skewness function; default is <code>delta(0.02)</code> for <code>lnskew0</code> and <code>delta(0.01)</code> for <code>bcskew0</code>
<code>zero(#)</code>	value for determining convergence; default is <code>zero(0.001)</code>
<code>level(#)</code>	compute the confidence interval at confidence level #; by default, no confidence interval is calculated

collect is allowed with `lnskew0` and `bcskew0`; see [U] 11.1.10 Prefix commands.

## Options

Main

`delta(#)` specifies the increment used for calculating the derivative of the skewness function with respect to  $k$  (`lnskew0`) or  $\lambda$  (`bcskew0`). The default values are 0.02 for `lnskew0` and 0.01 for `bcskew0`.

`zero(#)` specifies a value for skewness to determine convergence that is small enough to be considered zero and is, by default, 0.001.

`level(#)` specifies the confidence level for the confidence interval for  $k$  (`lnskew0`) or  $\lambda$  (`bcskew0`). The confidence interval is calculated only if `level()` is specified. # is specified as an integer; 95 means 95% confidence intervals. The `level()` option is honored only if the number of observations exceeds 7.

## Remarks and examples

[stata.com](https://www.stata.com)

### ► Example 1: Inskew0

Using our automobile dataset (see [U] 1.2.2 Example datasets), we want to generate a new variable equal to  $\ln(\text{mpg} - k)$  to be approximately normally distributed. `mpg` records the miles per gallon for each of our cars. One feature of the normal distribution is that it has skewness 0.

```
. use https://www.stata-press.com/data/r18/auto
(1978 automobile data)
```

```
. lnskew0 lnmpg = mpg
```

Transform	k	[95% conf. interval]	Skewness
ln(mpg-k)	5.383659	(not calculated)	-7.05e-06

This created the new variable `lnmpg = ln(mpg - 5.384)`:

```
. describe lnmpg
```

Variable name	Storage type	Display format	Value label	Variable label
lnmpg	float	%9.0g		ln(mpg-5.383659)

Because we did not specify the `level()` option, no confidence interval was calculated. At the outset, we could have typed

```
. use https://www.stata-press.com/data/r18/auto, clear
(1978 automobile data)
. lnskew0 lnmpg = mpg, level(95)
```

Transform	k	[95% conf. interval]		Skewness
ln(mpg-k)	5.383659	-17.12339	9.892416	-7.05e-06

The confidence interval is calculated under the assumption that  $\ln(\text{mpg} - k)$  really does have a normal distribution. It would be perfectly reasonable to use `lnskew0`, even if we did not believe that the transformed variable would have a normal distribution—if we literally wanted the zero-skewness transform—although, then the confidence interval would be an approximation of unknown quality to the true confidence interval. If we now wanted to test the believability of the confidence interval, we could also test our new variable `lnmpg` by using `swilk` (see [R] [swilk](#)) with the `lnnormal` option.

◀

## □ Technical note

`lnskew0` and `bcskew0` report the resulting skewness of the variable merely to reassure you of the accuracy of its results. In our example above, `lnskew0` found  $k$  such that the resulting skewness was  $-7 \times 10^{-6}$ , a number close enough to zero for all practical purposes. If we wanted to make it even smaller, we could specify the `zero()` option. Typing `lnskew0 new=mpg, zero(1e-8)` changes the estimated  $k$  to 5.383552 from 5.383659 and reduces the calculated skewness to  $-2 \times 10^{-11}$ .

When you request a confidence interval, `lnskew0` may report the lower confidence interval as `.'`, which should be taken as indicating the lower confidence limit  $k_L = -\infty$ . (This cannot happen with `bcskew0`.)

As an example, consider a sample of size  $n$  on  $x$  and assume that the skewness of  $x$  is positive, but not significantly so, at the desired significance level—say, 5%. Then, no matter how large and negative you make  $k_L$ , there is no value extreme enough to make the skewness of  $\ln(x - k_L)$  equal the corresponding percentile (97.5 for a 95% confidence interval) of the distribution of skewness in a normal distribution of the same sample size. You cannot do this because the distribution of  $\ln(x - k_L)$  tends to that of  $x$ —apart from location and scale shift—as  $x \rightarrow \infty$ . This “problem” never applies to the upper confidence limit,  $k_U$ , because the skewness of  $\ln(x - k_U)$  tends to  $-\infty$  as  $k$  tends upward to the minimum value of  $x$ .

□

## ▷ Example 2: `bcskew0`

In [example 1](#), using `lnskew0` with a variable such as `mpg` is probably undesirable. `mpg` has a natural zero, and we are shifting that zero arbitrarily. On the other hand, use of `lnskew0` with a variable such as temperature measured in Fahrenheit or Celsius would be more appropriate because the zero is indeed arbitrary.

For a variable like `mpg`, it makes more sense to use the Box–Cox power transform (Box and Cox 1964):

$$y^{(\lambda)} = \frac{y^\lambda - 1}{\lambda}$$

$\lambda$  is free to take on any value, but  $y^{(1)} = y - 1$ ,  $y^{(0)} = \ln(y)$ , and  $y^{(-1)} = 1 - 1/y$ .

`bcskew0` works like `lnskew0`:

```
. bcskew0 bcmpg = mpg, level(95)
```

Transform	L	[95% conf. interval]		Skewness
(mpg^L-1)/L	-0.3673283	-1.212752	.4339645	.0001898

The 95% confidence interval includes  $\lambda = -1$  ( $\lambda$  is labeled L in the output), which has a rather more pleasing interpretation—gallons per mile—than  $(\text{mpg}^{-0.3673} - 1)/(-0.3673)$ . The confidence interval, however, is calculated assuming that the power transformed variable is normally distributed. It makes perfect sense to use `bcskew0`, even when you do not believe that the transformed variable will be normally distributed, but then the confidence interval is an approximation of unknown quality. If you believe that the transformed data are normally distributed, you can alternatively use `boxcox` to estimate  $\lambda$ ; see [R] [boxcox](#).

◀

## Stored results

`lnskew0` and `bcskew0` store the following in `r()`:

Scalars

<code>r(gamma)</code>	$k$ ( <code>lnskew0</code> )
<code>r(lambda)</code>	$\lambda$ ( <code>bcskew0</code> )
<code>r(lb)</code>	lower bound of confidence interval
<code>r(ub)</code>	upper bound of confidence interval
<code>r(skewness)</code>	resulting skewness of transformed variable

## Methods and formulas

Skewness is as calculated by `summarize`; see [R] [summarize](#). Newton’s method with numeric, uncentered derivatives is used to estimate  $k$  (`lnskew0`) and  $\lambda$  (`bcskew0`). For `lnskew0`, the initial value is chosen so that the minimum of  $x - k$  is 1, and thus  $\ln(x - k)$  is 0. `bcskew0` starts with  $\lambda = 1$ .

## Acknowledgment

`lnskew0` and `bcskew0` were written by Patrick Royston of the MRC Clinical Trials Unit, London, and coauthor of the Stata Press book *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*.

## Reference

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26: 211–252.

## Also see

[R] **boxcox** — Box–Cox regression models

[R] **ladder** — Ladder of powers

[R] **swilk** — Shapiro–Wilk and Shapiro–Francia tests for normality

Stata, Stata Press, and Mata are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. StataNow and NetCourseNow are trademarks of StataCorp LLC. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2023 StataCorp LLC, College Station, TX, USA. All rights reserved.



For suggested citations, see the FAQ on [citing Stata documentation](#).