Problematic standard errors and confidence intervals for skewness and kurtosis

Daniel B. Wright · Joshua A. Herrington

Published online: 7 February 2011
© Psychonomic Society, Inc. 2011

Abstract Many statistics packages print skewness and kurtosis statistics with estimates of their standard errors. The function most often used for the standard errors (e.g., in SPSS) assumes that the data are drawn from a normal distribution, an unlikely situation. Some textbooks suggest that if the statistic is more than about 2 standard errors from the hypothesized value (i.e., an approximate value for the critical value from the t distribution for moderate or large sample sizes when $\alpha = 5\%$), the hypothesized value can be rejected. This is an inappropriate practice unless the standard error estimate is accurate and the sampling distribution is approximately normal. We show distributions where the traditional standard errors provided by the function underestimate the actual values, often being 5 times too small, and distributions where the function overestimates the true values. Bootstrap standard errors and confidence intervals are more accurate than the traditional approach, although still imperfect. The reasons for this are discussed. We recommend that if you are using skewness and kurtosis statistics based on the 3rd and 4th moments, bootstrapping should be used to calculate standard errors and confidence intervals, rather than using

the traditional standard. Software in the freeware R for this article provides these estimates.

Keywords Skewness · Kurtosis · Confidence intervals

Every modern introductory statistics book discusses the importance of examining a variable's distribution as part of exploratory data analysis (Tukey, 1977), both to examine the distribution with respect to the assumptions of the inferential statistics that may be used and to get a feel for the data, what Robert Rosenthal described as becoming friends with your data (Wright, 2003). There are several verbal phrases that can be used to describe a distribution, including "most of the data are around here," "the data are spread out," "the distribution has one really long tail," and "the distribution has heavy tails and a high peak." These vague concepts can be operationalized in many ways (Balanda & MacGillivray, 1988; Mosteller & Tukey, 1977). The first two phrases are often operationalized with the mean and standard deviation, but there is nothing in the definitions of these statistics that make them uniquely appropriate for these concepts, although they fit well mathematically with a large family of well-known statistics.

The latter two phrases are often described with two statistical concepts Karl Pearson introduced: skewness and kurtosis. Skewness is where one of the tails is longer than the other. Pearson (1895) used this phrase as a way to describe asymmetry in distributions and created several statistics—for example, $[\text{mean}(x) - \text{mode}(x)]/\text{sd}(x)$ —to measure asymmetry. According to this statistic, if the mean is greater than the mode, it is positively skewed. Variables like reaction time and income are usually positively skewed, because there are a few trials on which people are really slow and a few people with large incomes, but most of the data are at the lower ends of these

The code for all graphs, functions, and simulations are available at www2.fiu.edu/~dwright/skewkurt. The functions are also available as part of the R package **mrt** 0.4.

D. B. Wright · J. A. Herrington
Florida International University,
Miami, FL, USA

D. B. Wright (✉)
Psychology Department, Florida International University,
11200 S.W. 8th Street,
Miami, FL 33199, USA
e-mail: dwright@fiu.edu

distributions. Most researchers and students seem to understand the concept of skewness fairly well.

Kurtosis is a less well-understood concept and therefore deserves further explanation (see DeCarlo, 1997, for an excellent review aimed at psychologists). Pearson (1905) described kurtosis in comparison with the normal distribution and used the phrases *leptokurtic*, *platykurtic*, and *mesokurtic* to describe different distributions. If the distribution has more values in the tails of the distribution, less in the shoulders of the distribution, and a peak, it is leptokurtic. If there are fewer values in the tails, more in the shoulders, and less in the peak, it is platykurtic. If the values are about the same as the normal distribution, it is mesokurtic. These phrases are frequently confused (see Fiori & Zenga, 2009, for a historical analysis; see also Đorić, Nikolić-Đorić, Jevremović, & Mališić, 2009). “Student” [Gosset] (1927, p. 160) offered a handy memory aid, depicted in Fig. 1. A platykurtic curve, like a platypus, has a short tail (and a head that looks like a tail) and is flat-topped, whereas a leptokurtic curve, like two leaping kangaroos, has long tails and is peaked up in the center. These descriptions only work well if the distributions you are comparing have similar means, variances, and skewness values.

The statistics that are usually used to operationalize skewness and kurtosis are variations of the third and fourth central moments. The r^{th} central moment, denoted m_r , is defined by $\sum (x_i - \bar{x})^r / n$, with the stipulation that m_0 is 1 and m_1 is 0. The word *moment* often refers to a standardized moment that takes into account lower order moments and is corrected to take bias into account. Because calculating the 3rd and 4th moments involves taking the residual to the 3rd and 4th power, respectively, they are greatly influenced by extreme values. There are many other ways to operationalize skewness and kurtosis—for example, on the basis of the values of L moments (e.g., Hosking, 1992) or different quartiles (e.g., Bowley's method; see Kotz & Seier, 2009, for a recent discussion). Although these more robust statistics can be useful and they are discussed much in econometrics, they are far less common in psychology than the statistics based on the moments, so they will not be considered here.

Fig. 1 Memory aides for descriptions of kurtosis. (Adapted from “Student,” 1927)

Joanes and Gill (1998) compared the skewness and kurtosis statistics that are produced by several of the large commercial statistics packages. Table 1 shows the equations. Let n be the sample size and m_r be the r^{th} moment as defined above. These are equivalent to the equations in Joanes and Gill (pp. 184–185), although the terms have been rearranged to stress the relationships among the different statistics. The earliest statistics are g_1 and g_2 . The other equations are used by different packages to adjust for bias when estimating the population values. As $n \rightarrow \infty$, these estimates converge to g_1 and g_2 . The alternative statistics for skewness are g_1 multiplied by a function of n so all provide estimates that behave similarly. The -3 in g_2 was placed there so that the normal distribution has $g_2 = 0$. Values above 0 show kurtosis in excess of the normal distribution, so sometimes $g_2 > 0$ is called *excess kurtosis*. The other kurtosis statistics are not found simply by multiplying g_2 by a function of n . Joanes and Gill show that these statistics can produce different estimates.

Because the phrases *leptokurtic* and *platykurtic* are easily confused, we will follow DeCarlo (1997) and use the phrase *positive kurtosis* to mean that the kurtosis value is higher than that for the normal distribution (leptokurtic) and *negative kurtosis* to mean that the kurtosis value is less than that of the normal distribution (platykurtic). DeCarlo (1997, p. 294) stressed that a proper interpretation of kurtosis for a unimodal symmetric distribution involves considering the peakedness of the center, the thickness of the tails, and the amount in the distribution's shoulders. Positive kurtosis implies a movement of part of the distribution's mass from the shoulders to the peak and the tails. Negative kurtosis implies data being moved from the peak and tails into the shoulders (Balanda & MacGillivray, 1988).

Precision of estimates

The focus of this article is to compare different estimates of precision for skewness and kurtosis statistics. It is important to know the precision of any estimate. Calculating the standard error of any estimate allows you to conduct a t test on the observed statistic against some hypothesized value. Thus, if we observed a kurtosis value of .5 and we wanted to test whether this was significantly different from 0 (the value for the normal distribution), we could divide .5 by the estimated standard error and look up this ratio in a t table. With the sample sizes typical in psychology ($20 < n < 80$), the necessary t value to be significant at $\alpha = .05$ is about 2 (which is similar enough to the critical value for the normal distribution, 1.96, that sometimes the normal distribution

Table 1 Statistics for skewness and kurtosis and the equations for their traditional standard errors

	$g1$	$G1$	$b1$
Skewness	$m_3/m_2^{3/2}$	$g1 \frac{\sqrt{n(n-1)}}{n-2}$	$g1 \left(\frac{n-1}{n}\right)^{3/2}$
SE traditional	$\sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$	$SEg1 \frac{\sqrt{n(n-1)}}{(n-2)}$	$SEg1 \left(\frac{n-1}{n}\right)^{3/2}$
	$g2$	$G2$	$b2$
Kurtosis	$\frac{m_4}{m_2^2} - 3$	$\frac{(n-1)((n+1)g2+6)}{(n-2)(n-3)}$	$(g2+3)\left(\frac{n-1}{n}\right)^2 - 3$
SE traditional	$\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}$	$SEg2 \frac{(n-1)(n+1)}{(n-2)(n-3)}$	$SEg2 \left(\frac{n-1}{n}\right)^2$

value is used). Furthermore, the standard error can be used to calculate confidence intervals around the estimate. Many society guidelines recommend reporting confidence intervals when reporting any test statistic (Wilkinson and the Task Force on Statistical Inference, 1999). Of course if the estimate of standard error is poor, this procedure should not be followed.

The traditional standard errors from Table 1 are functions only of n , so they do not depend on skewness, kurtosis, or any other aspect of the distribution's shape. Figure 2 shows the relationship between traditional standard errors and sample size for the different statistics in Table 1. The curves for $g1$, $g2$, $b1$, and $b2$ increase for small n s (e.g., the maximum for $b1$ is 9 and for $b2$ is 18), and all the curves decrease as sample size gets larger. $G1$ and $G2$ steadily decrease for all values greater than 5.

Our interest is in how close the curves in Fig. 2 are to the estimated values from bootstrap procedures and the actual values. We made a function called **sktable** in R that calculates all the statistics in Table 1 for a variable and is available in the R package **mrt** (Wright, 2010).¹ It also calculates bootstrap standard errors and confidence intervals and is described in more detail in the Appendix. Here, we focus on the standard errors; in the next section, the confidence intervals are discussed. The function uses a simple nonparametric bootstrap: taking B samples of size n with replacement from the original sample, calculating the value of each for the six statistics in Table 1, and then taking the standard deviation of the B values to estimate the standard error (see Efron & Tibshirani, 1993, Chap. 6). The function has the default $B = 2,000$, but the user can change this. The 95% bias-corrected and accelerated (BCa) intervals are reported. The BCa intervals are the most recommended of the main types of bootstrap confidence intervals (Carpenter & Bithell, 2000). The calculations for BCa intervals work by calculating the bias between the sample statistic and the mean of the statistic for all the bootstrap samples and adjusting the interval accordingly.

¹ R is a free statistical environment and package (R Development Core Team, 2010). The Web site for this article contains links for downloading the package and other information about the package.

In order to compare the traditional and bootstrap standard error estimates, Monte Carlo simulation methods were used for the bootstrap estimates. The number of replications in both statistical simulations and replications for bootstrap estimates is important. We used $k = 1,000$ trials for each of several sample sizes to create curves for the bootstrap estimates that can be compared with those in Fig. 2. This number was enough to yield smooth curves. For the number of bootstrap replications, it is generally necessary to have more replications for estimating confidence intervals than for estimating standard errors. Efron and Tibshirani (1993, p. 52) advise that "very seldom are more than $B = 200$ replications needed for estimating a standard error," so this will be the number used in our simulations. The R package **boot** (Canty & Ripley, 2010), which is based on Davison and Hinkley (1997), was used.

The data were created from six different distributions: normal (symmetric and kurtosis of zero); Student's t with $df = 5$ (symmetric and positive kurtosis); mixed 80% normal with a mean of 0 and a standard deviation of 1 and 20% normal with a mean of 0 and a standard deviation of 10 (symmetric and positive kurtosis); sometimes called a scaled contaminated normal distribution); chi-squared with $df = 1$ (positively skewed and positive kurtosis); Bernoulli (i.e., binomial with one trial) with a probability of $.5 + \sqrt{1/12}$ (negatively skewed and kurtosis of zero); and uniform (symmetric and negative kurtosis). For these simulations, sample size was varied from 5 to 100 in steps of 5. Some of the Bernoulli

Fig. 2 The relationship between the traditional standard errors for skewness and kurtosis and sample size

samples had all 0 s or all 1 s and, so, had no variance. These samples were excluded from the analyses both when there was no variance for one of the 1,000 trials and within the bootstrap procedure when a replication had no variance.

Because the distributions are known, the standard deviation of large samples from the distribution can be used to estimate the actual standard errors. These estimates will be more valid (i.e., unbiased) than bootstrap estimates, because of idiosyncratic aspects of the sample distribution, but may still be errant because of the idiosyncrasies caused by the effect of taking third and fourth powers with extreme values. Although bootstrap standard error estimates tend to be fairly good, they can have problems particularly with extreme distributions (Chernick, 2008, Chap. 9). Of course, in practice the population distribution is not known, which means that, in many cases, the observed distribution will be the best estimate available for the population distribution (and if some a priori knowledge of the distribution exists, parametric bootstrap replications can be used).

Figure 3 shows the relationship between the traditional and bootstrap standard errors when the data are sampled from a normal distribution. The curves are fairly close to each other. For example, for the $G2$ statistic with $n = 40$, the traditional standard error estimate is .73, while the bootstrap estimate is .61, whereas with $n = 80$, the values are a traditional standard error estimate of .53 and a bootstrap standard error estimate of .43. The bootstrap values are slightly lower than the traditional values for most sample sizes, but the difference is relatively small, as compared with the differences found for the remaining distributions. For the normal distribution, the actual and traditional curves are the same.

Figure 4 compares the bootstrap and traditional standard errors for a t distribution with 5 degrees of freedom. The $t5$ distribution is symmetric, with skewness equal to zero. There is more weight in the tails and a higher peak than in the normal distribution, so it has positive kurtosis. The deviation from the normal distribution makes the bootstrap standard errors larger than the traditional estimates (which assume normality), because extreme values, which will be more frequent than with a normal distribution, have large effects on skewness and kurtosis and, thus, increase their variability. The curve for the actual standard errors is above the bootstrap estimates. For the kurtosis statistics, the standard error *increases* as the sample size increases. The pattern for the mixture distribution is similar (Fig. 5), except that the estimates for the actual standard errors do not increase with sample size. Figure 6 for the $\chi^2(1)$ distribution shows a pattern similar to that in Fig. 4, with the estimated actual standard errors being the highest (and increasing with sample size for kurtosis), followed by the bootstrap estimates, and with the traditional estimates being the smallest. The Bernoulli curves (Fig. 7) show the actual and bootstrap estimates as being very similar but the traditional estimates as being too low. For the uniform distribution, the bootstrap and actual estimates are again similar, but the traditional estimates are too high (Fig. 8).

Estimates of the standard errors for skewness and kurtosis that take into account other moments are available. Kendall, Stuart, and Ord (1987, p. 344) provide equations for these for $g1$ and $g2$, and these estimates can be scaled to adjust for bias in the same way as the traditional estimates in Table 1. These statistics sometimes cannot be calculated because the estimate

Fig. 3 The actual, traditional, and bootstrap standard errors for skewness and kurtosis for sample sizes up to $n = 100$ for the normal distribution

Fig. 4 The actual, traditional, and bootstrap standard errors for skewness and kurtosis for sample sizes up to $n = 100$ for the *Student t*₅ distribution

they use for the variance of the sampling distribution can be negative. These estimates are included as part of the **sktable** function, but we do not formally compare them with the other approaches. When their values exist, they appear similar to the bootstrap estimates for the distributions we examined, but they are calculated much more quickly. Anscombe and Glynn (1983) described transforming the data to lessen the impact of outliers. Their test has been implemented in the function **anscombe.test** in the package **moments** (Komsta & Novomestky, 2007). A function **anscombe.boot** was written for this article and

is part of package **mrt** (Wright, 2010). It gives the BCa bootstrap confidence intervals for the z values produced from **anscombe.test** (see the Appendix for more details).

Using the standard errors for hypothesis testing and confidence intervals

Many researchers use the skewness and kurtosis standard errors to test the hypothesis that the statistics are equal to zero. We looked through several psychology statistics

Fig. 5 The actual, traditional, and bootstrap standard errors for skewness and kurtosis for sample sizes up to $n = 100$ for a mixture of two normal distributions

Fig. 6 The actual, traditional, and bootstrap standard errors for skewness and kurtosis for sample sizes up to $n = 100$ for the χ^2 distribution with $df = 1$

textbooks and found that many did not mention skewness and kurtosis or just provided qualitative definitions. DeCarlo (1997) has reported how, when mentioned, the kurtosis statistic has often been incorrectly described. Of those textbooks that did mention these statistics, most did not discuss the precision of the estimates. Of those that did, they all used traditional standard error estimates, or something very similar. Consider the advice from two excellent textbooks. Tabachnick and Fidell (1996, p. 72) told readers to use the approximations of standard errors for

skewness of $\sqrt{6/n}$ and for kurtosis $\sqrt{24/n}$, which produce roughly the traditional values in Table 1, and to calculate a z value for these. Field (2009, pp. 138–139) says to use the estimates SPSS produces (for $G1$ and $G2$) to allow testing particular hypotheses and comparing different distributions. Because we have drawn samples from known distributions, we can test how well the different methods are at correctly rejecting the null hypothesis when it is true. The normal, $t5$, mixture, and uniform distributions are unskewed, and so should have $G1$ values, in the population, of zero.

Fig. 7 The actual, traditional, and bootstrap standard errors for skewness and kurtosis for sample sizes up to $n = 100$ for the Bernoulli distribution with $p = .5 + \sqrt{1/12}$

Fig. 8 The actual, traditional, and bootstrap standard errors for skewness and kurtosis for sample sizes up to $n = 100$ for the uniform distribution

Therefore, hypothesis tests should reject this hypothesis approximately 5% of the time, using the $\alpha = 5\%$ level. The normal and Bernoulli ($p = .5 + \sqrt{1/12}$) distributions have kurtosis values of zero, so they should have $G2$ values, in the population, of zero. Their values should also be rejected approximately 5% of the time.

Table 2 shows the percentage of null hypotheses rejected from 2,000 trials with $n = 100$ and $\alpha = .05$. A trial was significant if the absolute value of the statistic divided by the standard error was greater than 1.98 or if the 95% BCa confidence interval (with 2,000 replications) did not include zero. When the null hypothesis is true in the population, rejecting it is a Type 1 error and should occur at a rate near the nominal value—here, 5%. For the skewness measures, the Type 1 error rate is as high as 76% for the traditional method. Thus, if you are sampling 100 cases from a distribution composed of 80% from a normal distribution with $\sigma = 1$ and 20% from a normal distribution with $\sigma = 10$ (i.e., a scale contaminated normal distribution), you will errantly reject the null hypothesis three quarters of the time if you use the approach advocated in many textbooks! If using the bootstrap methods, the proportion of rejections still exceeds the nominal level, but the discrepancy is not as large. If the distribution is uniform, the traditional method produces far fewer errors than is appropriate. While having fewer errors is good, there is a cost in the power. This occurs only because the traditional standard errors are too large. Overall, the traditional approach works if the distribution is normal, but it fails badly with all the other symmetric distributions. The bootstrap methods also produce too many rejections for the $t5$ and mixture

distributions, but the number of these errors is less. The uniform distribution is different from the others because, for it, the traditional standard error estimates were larger than both the bootstrap standard errors and the actual standard errors.

The traditional approach for the $G2$ hypothesis test is good when the data are normal, but the Type 1 error rate goes to 26% for the Bernoulli distribution. The bootstrap methods have a similar error rate to the $G1$ tests for the normal distribution, producing too high an error rate. For the Bernoulli distribution, the bootstrap methods produce approximately the nominal level of Type 1 errors, but overall, we are not encouraged by any set of hypothesis tests.

Thus, the traditional approach is very poor, but the bootstrap t approach and the BCa confidence interval approach are not ideal. In general, statisticians have observed that the bootstrap t approach can produce "erratic results," and while it can be used for estimating measures of central tendency, such as the median and mean, it "cannot be trusted for more general problems" (Efron & Tibshirani, 1993, pp. 160, 162). Therefore, Efron and Tibshirani state that the BCa interval should be used for hypothesis testing. However, the error rates are similar for these intervals. Bootstrap estimation tends to have difficulties with distributions with extreme values, particularly when those extreme points are highly influential (Chernick, 2008, Chap. 9). The skewness and kurtosis statistics in Table 1 require taking a residual term to the third and fourth power, respectively. This will mean that extreme points can have a very large effect. It is important to stress that the instability is due to the actual standard

Table 2 The Type 1 error rate: The percentages of times the null hypothesis of either $G1 = 0$ or $G2 = 0$ was incorrectly rejected, from 2,000 trials with $n = 100$ for three approaches ($\alpha = .05$)

	Traditional <i>SE</i>	Bootstrap <i>SE</i>	BCa CIs
<i>G1</i> significance testing			
Normal	4.9%	12.3%	18.1%
t ($df = 5$)	40.7%	14.3%	17.7%
Mixture of normals	76.4%	16.4%	14.3%
Uniform	0.3%	3.8%	4.1%
<i>G2</i> significance testing			
Normal	4.4%	11.5%	14.3%
Bernoulli ($p = .5 + \sqrt{1/12}$)	26.0%	4.0%	4.9%

deviation of the statistics not converging quickly (or at all) to small values.

Discussion

The traditional standard errors for skewness and kurtosis printed by many statistics packages are very poor. While they are appropriate for normal distributions, deviations from normality, like a t distribution with $df = 5$ or a mixture of two normal curves with different standard deviations, produce standard errors that can be 5 times too small. Because normal distributions are rare in psychology (Micceri, 1989), the practice of encouraging the use of standard errors that are grossly in error with deviations from normality by printing them in statistics packages should be stopped. The function **sktable**, written for this article, produces the bootstrap standard errors, BCa confidence intervals, and standard errors, which take into account characteristics of the distributions.

One question is whether the traditional standard errors can be used if you are using this only to test whether the distribution is normal. This approach is not recommended for at least two reasons. First, there are several better tests that are designed to test normality, such as the Shapiro and Wilk test (1965), and are available in many statistics packages. Testing skewness and kurtosis, individually, is problematic because they are dependent on each other, so rather than simply following the textbook approach, tests like D'Agostino's test (D'Agostino, Belanger, & D'Agostino, 1990) could be used. Second, because the traditional standard errors both under- and overestimate the true standard errors, you would not know whether the statistical test was liberal or conservative without guessing about the distribution. One method for guessing about the distribution is to take thousands of resamples of the observed distribution, which is what bootstrapping does.

The problems with the standard errors shown in this article are due to the use of the third and fourth moments to calculate skewness and kurtosis. These are

very sensitive to deviations in the tails of the distributions and are not sensitive to deviations in the peaks of the distributions (Lindsay & Basak, 2000). Seier and Bonett (2003) discussed formulae that allow the user to vary the relative influence of deviations in the tails and the peak of a distribution, but these are not commonly used. The transformation described by Anscombe and Glynn (1983) can also be used. A function is described in the Appendix that uses this, but it is also not widely used. If skewness and kurtosis are operationalized in other ways, the impact of extreme points can be lessened. Balanda and MacGillivray (1988) discussed several ways in which quartiles could be used for operationalizing these statistics that would be less influenced by extreme points. One of the most promising is L -moments (Hosking, 1992). Hosking shows how skewness and kurtosis measures based on these are more consistent with the Shapiro–Wilk test of normality. These can be adjusted, for example, by trimming, which further increases their robustness (Elamir & Seheult, 2003). In time, these alternative methods may become more popular, but dramatically changing the statistics for measuring them would also change the meaning of skewness and kurtosis. Therefore, it is unlikely that these alternatives will become widespread in the near future.

The main recommendations from this article are that researchers should be careful when using traditional measures of standard error for skewness and kurtosis. Bootstrap confidence intervals can be used, but researchers should be aware that these may still be in error. For testing whether a distribution is of a certain shape, tests designed specifically for this purpose should be used. Methods based on transformations of the moments and L -moments should be considered.

Appendix

The **sktable** function produces all the statistics in Table 1 plus bootstrap standard errors, standard errors that take into account other moments from Kendall et al. (1987), and BCa bootstrap confidence intervals. The function calls individual functions for each of these

statistics. All of these functions (and a function for running simulations for any distribution and a function for testing hypotheses) are part of the **mrt** package (Wright, 2010). That function calls the **boot** package, which is automatically installed when installing **mrt**.

A variable can be created to illustrate the function:

```
set.seed(305)
sktable(rt(200,10))
```

and the output is:

	g1	G1	b1	g2	G2	b2
estimate	-0.1309629	-0.1319546	-0.1299819	0.5923677	0.6380586	0.5565338
trad SE	0.1706327	0.1719248	0.1693545	0.3337070	0.3422024	0.3303783
w/mom SE	0.1692780	0.1705598	0.1680100	0.5122262	0.5252662	0.5071167
boot SE	0.2435268	0.2322226	0.2267685	0.4344167	0.4410719	0.4285649
lb BCa	-0.6063754	-0.5981438	-0.5867463	-0.1167060	-0.1338628	-0.1181461
ub BCa	0.3170932	0.3234357	0.2894199	1.6940856	1.6848059	1.7126058

The w/mom SE estimates (those from Kendall et al., 1987) for the skewness values will often produce **NaN** for "not a number," because the function tries to take the square root of a negative variance estimate. Within those functions, **0i** could be added to negative variance estimates to make them complex numbers and allow the square root to be calculated, but the resulting standard errors would be difficult to interpret.

There are two options for the function. The first is not running any bootstrap methods in order

to make the function faster. For this, type **sktable(variable,bootm=FALSE)**. The second option is allowing the number of bootstrap replications to vary. The default is set to 2,000 but can be changed by typing, for example, **sktable(variable, reps=200)**. The function will often produce warnings for the bootstrap estimates because of extreme values. The function **anscombe.boot** has the following form:

```
anscombe.boot(varx, reps=2000, printtest=FALSE, clevel=.95, method="bca", ...)
```

The user needs to include the variable of interest. The optional arguments are, in order, for the number of bootstrap replications, whether to print the output of the Anscombe–Glynn test, the confidence interval, and whether to use BCa or percentile confidence intervals. In testing the function, there were occasions when the BCa method had computational difficulties. If this occurs, we recommend trying again, in case it was due to some oddity in the bootstrap samples, and then typing in anything other than **"bca"** for the final argument, which will cause the percentile method to be used. The kurtosis statistic produced is the original Pearson statistic (without the -3 adjustment by Fisher). The z score bounds will need to be back-transformed into kurtosis (g2) values. Here is a call to the function:

```
anscombe.boot(runif(30))
```

Kurtosis	z	lb of z	ub of z
2.112667	-1.197220	-3.262169	0.750317

References

- Anscombe, F. J., & Glynn, W. J. (1983). Distribution of kurtosis statistic for normal statistics. *Biometrika*, 70, 227–234.
- Balanda, K. P., & MacGillivray, H. L. (1988). Kurtosis: A critical review. *American Statistician*, 42, 111–119.
- Canty, A., & Ripley, B. (2010). *boot: Bootstrap R (S-Plus) functions*. R package version 1.2-42.
- Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19, 1141–1164.
- Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed.). Hoboken: Wiley.
- D'Agostino, R. B., Belanger, A., & D'Agostino, R. B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *American Statistician*, 44, 316–321.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press.
- DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, 2, 292–307.
- Dorić, D., Nikolić-Dorić, E., Jevremović, V., & Mališić, J. (2009). On measuring skewness and kurtosis. *Quality & Quantity*, 43, 481–493.
- Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton: Chapman & Hall/CRC.

- Elamir, E. A. H., & Seheult, A. H. (2003). Trimmed L-moments. *Computational Statistics and Data Analysis*, 43, 299–314.
- Field, A. P. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.
- Fiori, A. M., & Zenga, M. (2009). Karl Pearson and the origin of kurtosis. *International Statistical Review*, 77, 40–50.
- Hosking, J. R. M. (1992). Moments or *L* moments? An example comparing two measures of distributional shape. *American Statistician*, 46, 186–189.
- Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society D*, 47, 183–189.
- Kendall, M., Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics: Vol. 1. Distribution theory*. New York: Oxford University Press.
- Komsta, L., & Novomestky, F. (2007). Moments: Moments, cumulants, skewness, kurtosis and related tests. R package version 0.11. Retrieved from <http://www.r-project.org>, <http://www.komsta.net/>
- Kotz, S., & Seier, E. (2009). Analysis of quantile kurtosis: Center and tails. *Statistical Papers*, 50, 553–568.
- Lindsay, B. G., & Basak, P. (2000). Moments determine the tail of a distribution (but not much else). *American Statistician*, 54, 1–4.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156–166.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading: Addison-Wesley.
- Pearson, K. (1895). Contributions to the mathematical theory of evolution: II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London*, 186, 343–414.
- Pearson, K. (1905). Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A Rejoinder. *Biometrika*, 4, 169–212.
- R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>
- Seier, E., & Bonett, D. G. (2003). Two families of kurtosis measures. *Metrika*, 58, 59–70.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.
- “Student” [Gosset, W. S.] (1927). Errors of routine analysis. *Biometrika*, 19, 151–164.
- Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.
- Wilkinson, L., The Task Force on Statistical Inference, & APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist*, 54, 594–604.
- Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *The British Journal of Educational Psychology*, 73, 123–136. doi:10.1348/000709903762869950
- Wright, D. B. (2010). mrt: Wright & London's Modern Regression Techniques. R package version 0.4. Available at <http://CRAN.R-project.org/package=mrt>