Journal of Experimental Psychopathology
JEP Volume 2 (2011), Issue 2, 252–270
ISSN 2043-8087 / DOI: 10.5127/jep.013611

Using Bootstrap Estimation and the Plug-in Principle for Clinical Psychology Data

Daniel B. Wright[a], Kamala London[b], Andy P. Field[c]

[a] *Psychology Department, Florida International University*

[b] *Department of Psychology, University of Toledo*

[c] *School of Psychology, University of Sussex*

Abstract

Psychologists estimate the precision of their statistics both to conduct hypothesis tests and to construct confidence intervals. The methods traditionally used for this are available only for a small set of statistics (e.g., the mean and transformations of it) and often make unrealistic assumptions about the variables' distributions. These assumptions are often particularly unrealistic in data derived from clinical samples, or when looking at groups responding at the extreme end of clinical constructs. Bootstrap estimation is a computer intensive procedure that offers a flexible and automatic alternative. The computer takes thousands of bootstrap samples from the observed data and from these bootstrap samples estimates the precision of the statistic. High-speed personal computers make the bootstrap a viable and appealing technique throughout the sciences. This article offers a tutorial on the theory and practice of applying bootstrap estimation to data from clinical samples and measures relevant to experimental psychopathology.

Keywords: Bootstrap, Robust methods

Correspondence to: Daniel B. Wright, Psychology Department, Florida International University, 11200 S.W. 8th Street, Miami, FL, 33199. Email: dwright@fiu.edu

1. Psychology Department, Florida International University, 11200 S.W. 8th Street, Miami, FL, 33199.

2. Department of Psychology, 2801 W Bancroft St. Mailstop 948, Toledo, OH 43606

3. School of Psychology, University of Sussex, Falmer, Brighton, East Sussex, BN1 9QH UK

Received 26-Nov-10; received in revised form 19-Jan-11; accepted 21-Jan-11

Table of Contents

Using Bootstrap Estimation and the Plug-in Principle for Clinical Psychology Data

At the heart of statistical inference is estimating the precision of an estimate. This can be done by calculating the standard error (*SE*), which is the basis of hypothesis testing, or by calculating the confidence interval (*CI*), which is the method urged by many psychology journals and societies (e.g., for the APA, see Wilkinson et al., 1999; for the BPS, see Wright, 2003). For over a century mathematical statisticians have created formulae for estimating SEs and CIs. They have solved this difficult task for only a small number of statistics, like the mean. Unfortunately, even these formulae are often only appropriate if the user is willing to make certain assumptions; for example, the residuals being normally distributed. In a survey of psychology research, Micceri (1989) found that most real data deviate greatly from the normal distribution, implying that the traditional methods for calculating SEs and CIs are seldom appropriate. While traditional tests often protect against falsely rejecting a hypothesis in the presence of outliers, they tend to overestimate the standard error and width of the confidence interval, thus decreasing the power of studies (Wilcox, 1998).

Data from clinical scales and clinical populations are often particularly prone to having non-normal distributions. Substantial skew has been shown in outcome measures relevant to clinical trials (Delucchi & Bostrom, 2004; Tang et al., 2005). This is true of both general measures such as economic and cost indicators (Barber & Thompson, 2000; Hlatky, Boothroyd, & Johnstone, 2002), quality of life (Arostegui, Nunez-Anton, & Quintana, 2007) and social functioning (Tyrer et al., 2005), but also measures of disorder specific constructs such as mania (Picardi et al., 2008), suicidal ideation (Binks et al., 2006) and depression (Rutter, & Miglioretti, 2003; Zimmerman, Chelminski, & Posternak, 2004). Perhaps most problematic for experimental psychopathology research is the fact that measures of clinical constructs are often heavily skewed in normal populations (e.g., Rutter & Miglioretti, 2003; Tyrer et al., 2005; Zimmerman et al., 2004). Experimental psychopathology research relies heavily on investigating clinical constructs in analogue populations, and therefore, data are very likely to be non-normal and especially skewed.

Bootstrapping offers a flexible and general alternative that can be used to find SEs and CIs for any statistic. Bootstrapping makes fewer assumptions than the traditional approaches, and in its most

popular form is mathematically simple. Introduced to the statistics community in 1977 by Bradley Efron (1979), bootstrapping builds upon other forms of computer intensive procedures (sometimes called *Monte Carlo* methods). Casella (2003) describes the bootstrap as a paradigm shift in the mindset of modern statistics. The difficult and sometimes impossible task of solving long complex equations has been replaced by making the computer work a little harder. Researchers do not need to try to fit their analyses into the small set of statistics for which traditional approaches have provided estimates for the SE and CI. Most important, bootstrapping is usually more accurate than traditional approaches (Efron & Tibshirani, 1993). However, as with any form of data analysis some care is needed when applying bootstrapping.

The value of bootstrapping is in the ease with which it can be applied and the diversity of statistics to which it can be applied (e.g., medians, correlations, regression coefficients). It is widely used by statisticians and is available in many computer packages. In a popular psychology statistics book, Howell (2007, p. 636) states that bootstrapping "will overtake what are now the more common nonparametric tests, and may eventually overtake traditional parametric tests." He says the question is not *if* bootstrapping will take over, but *when*.

The purpose of this article is to explain conceptually what bootstrap estimation is and to illustrate it with recent examples of relevance to experimental psychopathology. We hope to familiarize readers with bootstrapping so that they feel comfortable using this modern statistical procedure. Readers interested in the mathematical and computational details should consult the textbooks listed in the references (e.g., Chernick, 2008; Efron, & Tibshirani, 1993; Good, 2006; Lunneborg, 2000).

This article begins with discussion of the *plug-in principle* (Efron & Tibshirani, 1993). Next, the mechanics of creating bootstrap samples and calculating statistics on these samples are described. These methods are illustrated with recent data from two clinical trials, and finally, we stress the value of teaching bootstrapping to the next generation of psychopathology researchers.

The Plug-in Principle

The plug-in principle is simple, but is the driving force behind both traditional and bootstrap methods for inferential statistics. Suppose you want to estimate the mean of a variable in a population. Ideally you would take all the data in the population and calculate the mean. This is what you are trying to estimate, often denoted μ, but usually obtaining values for the whole population is impossible. Instead, a sample is taken, a statistic is calculated, and then that value (e.g., the mean) is used to estimate the population value. We write $E(\mu) = \bar{x}$, in which $\bar{x}$ is the sample mean and E( ) means 'the expected value'. Statisticians talk about the sample mean being a *statistic* that is used to estimate a population *parameter*. Here it is a *point estimate* meaning it estimates a single point for the value of the statistic in the population. This sample mean is a plug-in estimate for the population mean. The sample mean is a good plug-in estimate because it usually is a fairly accurate estimate of the population mean.

Some sample statistics perform less well as plug-in estimates of the corresponding population values. The sample range is a poor plug-in estimate of the population range because it greatly underestimates the population value. The sample variance is a pretty good plug-in estimate for the population variance, but it slightly underestimates the population value. Sometimes it is prudent to slightly tweak a statistic to make it a better estimate. The textbook solution for correcting the bias of the sample variance is to divide the sum of squared deviations by $n-1$ rather than n (Howell, 2007). Bootstrapping is not used to improve point estimates; it is used to estimate how accurate point estimates are.

The bootstrap uses the plug-in principle by estimating the population's distribution with the sample distribution. The left panel of Figure 1 shows a sample's distribution for 25 people answering a 1–7 rating scale for a response from a psychometric test. In the most common form of bootstrap estimation, the sample distribution is used as a plug-in estimate of the population distribution. In traditional statistics researchers often assume that the population distribution is normally distributed and just plug-in the sample mean and standard deviation. As Micceri (1989) has shown, this assumption is usually invalid in psychology. The bootstrap uses the information from your data rather than making this assumption. Just like how we tweak the sample variance (by dividing by $n-1$ rather than n), sometimes you may want to tweak your estimate for the population distribution. Nobody in the sample distribution in Figure 1 circled '5' (left panel). If the researchers think this is just a sampling fluke, they may want to smooth the distribution as has been done in the right panel of Figure 1. In practice, smoothing data like this usually makes little difference on the resulting estimates for most statistics. With this in mind we will just use the sample distribution as a plug-in estimate for the population distribution.

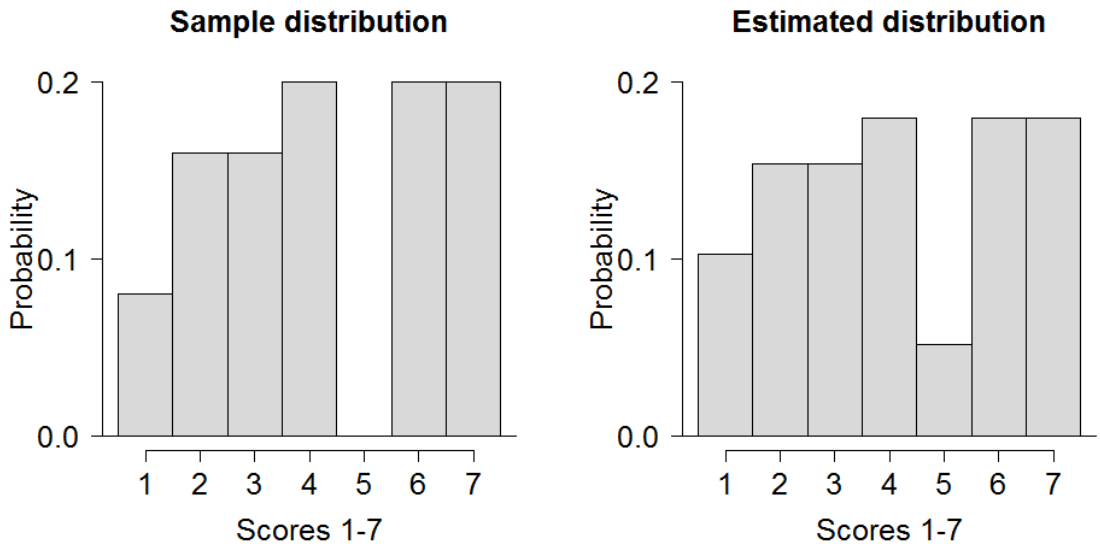

Figure 1: The left panel shows 25 people's responses on a 1–7 rating scale. No one responded with "5". The right panel has smoothed these data slightly.

Bootstrap Sampling and Bootstrap Estimation: Examples

Bootstrap Estimates for the Median

Figure 2 shows how the bootstrap procedure works for a small sample. Goff and Simms (1993) were interested in the number of alternative personalities among people who had been diagnosed with multiple personality disorder (now dissociative identity disorder). They compared cases from 1800–1965 with those from 1980–1988. They found the mean number of alternative personalities in the early cases was 3.0 and the mean of the more recent cases was 12.3. They had 54 cases in their recent sample, but for illustrative ease suppose they identified 10 cases and the number of alternative personalities each had were: 1, 1, 2, 2, 4, 5, 7, 15, 30, and 56. These data are very skewed as is often the case in clinical measures (see earlier). The mean is high because of a couple of really large values. The data are not normally distributed and therefore the traditional methods for calculating the 95% confidence interval for the mean are inappropriate; this situation is one in which the median may be more appropriate than the mean. The median is the middle data point (or observation) and is often preferred to the mean because

extreme values or high levels of skew affect it less. In Goff and Simms the median for early cases was 2.0 alternates and the median for 1980 cases was 4.5 alternates. These values are much more similar than the aforementioned means. Unfortunately, calculating SEs and CIs for the median using traditional approaches is difficult (Wilcox, 2005) and not available in the main statistics packages used by psychologists. Bootstrapping allows these to be calculated in a simple and automatic fashion.

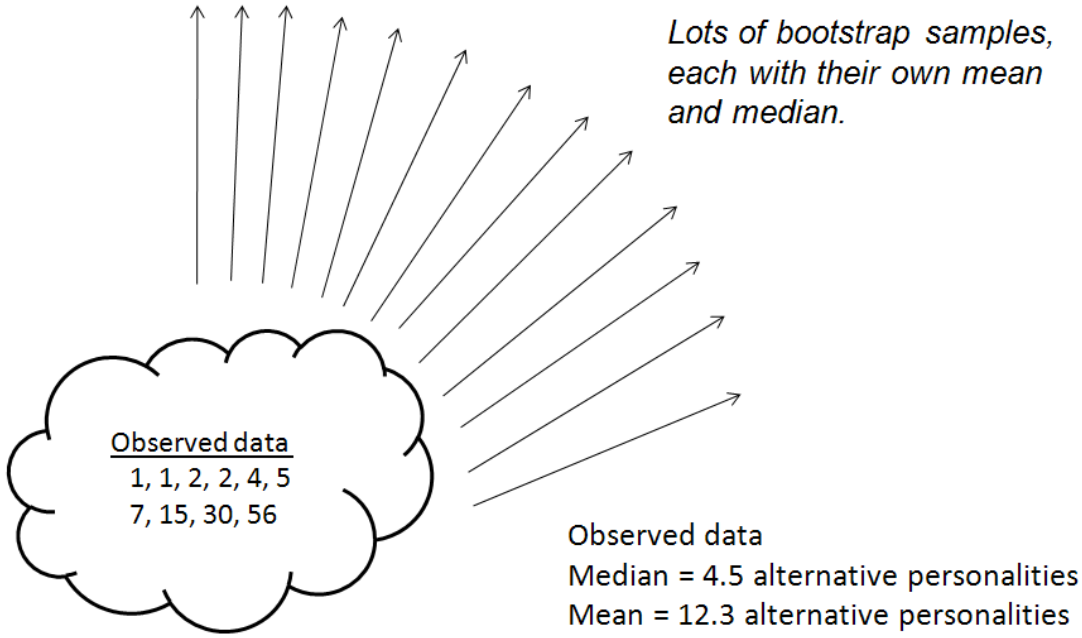

Figure 2: Bootstrapping the mean and median for data based on Goff and Simms (1993). For illustration purposes, n = 10 people.

Bootstrap sampling can be explained using the analogy of picking ping-pong balls from a box. In our example of multiple personalities, imagine that each person's number of alternate personalities is written on a ping-pong ball and placed in a box. The researcher randomly picks one ball, writes down the score (the number of alternative personalities), and returns the ball to the box. Returning the ball to the box is important. This process is called *sampling with replacement,* and it means that one person's number could be in the bootstrap sample multiple times and another person's value may not be in the bootstrap sample. This process continues until 10 balls have been picked (and the 10 values on them have been noted). These 10 values are the first bootstrap sample. The norm is to pick the same number of cases in each bootstrap sample as was in the original sample. The researcher then calculates the statistic of interest, here the mean and median, for this first bootstrap sample and writes the result down. The researcher then choses another 10 balls (replacing the ball each time) and notes down the 10 values. This is the second bootstrap sample and the mean and median are calculated for this sample and written down. This process is repeated several times (each time is known as a *replication*). Figure 3 shows this process for 5 bootstrap samples.

If we were drawing bootstrap samples and calculating statistics by hand each time then the process would soon become time consuming and tedious. Fortunately, computers work faster than humans and do not get bored; therefore, in practice a computer is 'drawing the balls from the box', and computing statistics which means that the number of replications can be large (most researchers recommend at least 2000).

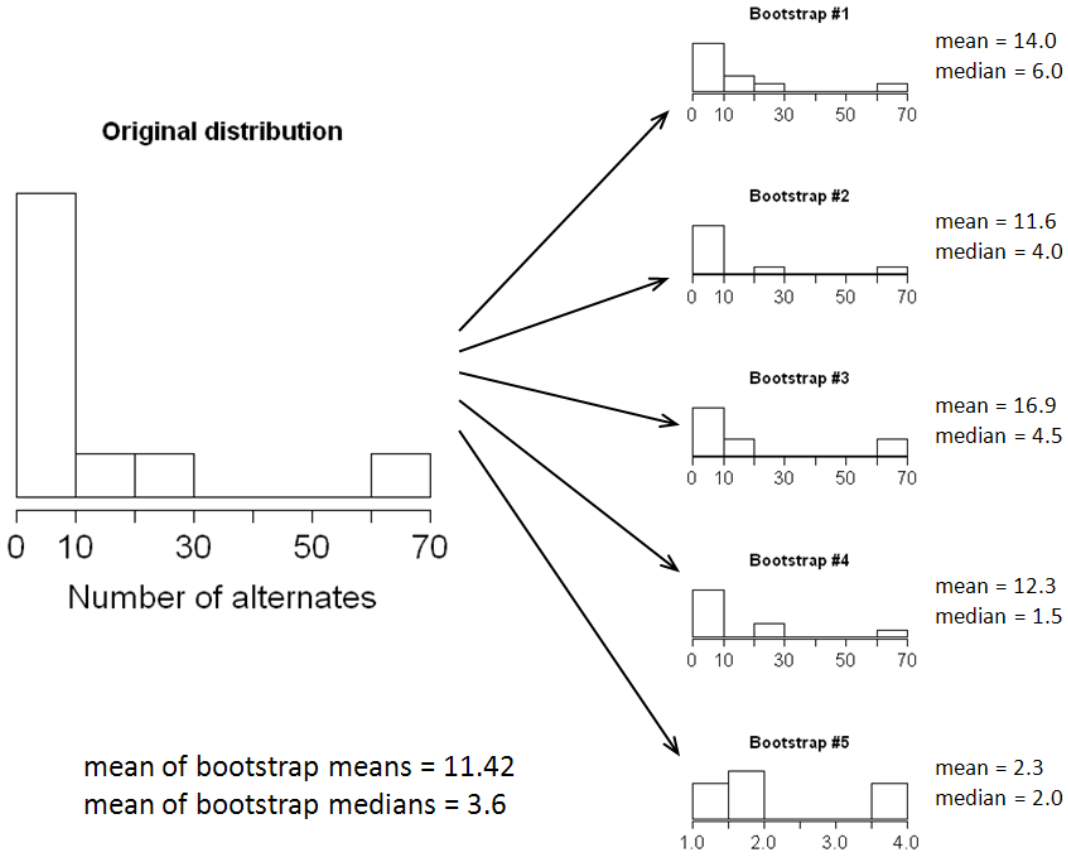

Figure 3: The distribution and statistics for 5 bootstrap samples for the data shown in Figure 2.

Having drawn 2000 bootstrap samples and computed, for example, the mean for each, we can compute the standard deviation of these means and use this value as an estimate of the standard error of the mean. Similarly, the 95% confidence interval for the mean can be estimated from the values that enclose the middle 95% of the bootstrap sample means (although a better method is described below). For our multiple personality data, the middle 95% of the mean values go from 3.70 to 24.80 alternates, and the middle 95% of the median values go from 2.0 to 16.0 alternates. These are called the 95% *percentile* bootstrap confidence intervals.

Bootstrapping Categorical Data (Kappa, association for a 2×2 table)

Although bootstrap estimation is simple, it is worth stressing that it is just one part of the analytic process and the other stages remain difficult. In particular, deciding what statistic should be estimated to evaluate your model is difficult. The assumption that the mean is the only statistic of interest can blinker psychologists from considering other statistics that may be more appropriate. Once you have decided which statistic to use, figuring out how to calculate the precision of the estimate can be difficult with the traditional approach, but it is not with the bootstrap.

Much data is categorical and often researchers are interested in the association between two categorical variables. The standard statistical test for this is Pearson's χ^2 test. While there is much agreement about using the χ^2 test for hypothesis testing, there is less agreement about what effect size to use. We will consider the simplest situation for estimating the association between two categorical variables when both variables are binary (i.e., can take only 1 of 2 values).

Consider the following example: Reynolds and colleagues conducted a project in which young people diagnosed with obsessive-compulsive disorder (OCD), anxiety disorders, and no known mental health problems were compared on measures of cognitive appraisals such as inflated responsibility, thought-action fusion and perfectionism (Libby, Reynolds, Derisley, & Clark, 2004) as well as measures of their parents' mental health, coping and family-functioning (Derisley, Libby, Clark, & Reynolds, 2005). Within these studies the researchers recorded much background demographic information about the children. We will use this dataset to look at how the different anxiety groups differed on demographic variables.

For the first analysis suppose that the interest is in whether the proportion of skilled versus professional father's occupations is the same for the OCD group as the anxious group. Table 1 shows a breakdown of this information. There is a huge literature on analyzing categorical data (Agresti, 2002, is an excellent graduate textbook). There has been over 100 hundred years of debate about how to analyze the difference between two proportions. Kraemer has argued in several sources (e.g., Kraemer & Gibbons, 2009; see also Gilchrist, 2009) that the κ (the Greek letter kappa) statistic should be used. There are a few forms for κ; we will use Cohen's (1960) version (called Cohen's κ or unweighted κ). κ ranges from −1 to +1, where 0 corresponds to no difference in proportions. Table 2 shows the formula for this. For this example, the formula for κ is:

$$\kappa = \frac{2((2)(5)-(14)(12))}{(2+14)\cdot(14+5)+(12+5)\cdot(2+12)} = -.58$$

We find a negative value because the frequencies on the off-diagonal (12 and 14) are larger than those on the main diagonal (2 and 5).

Table 1: Breakdown of the category of the child and father's occupational category from Libby, et al. (2004) and Derisley et al. (2005).

		Father's Occupational Group		
		Unskilled	**Skilled**	**Professional**
Group	OCD	9	2	14
	Anxious	7	12	5
	Control	13	15	21

Note: Unemployed and student occupational categories not included because of low numbers.

Table 2: Calculating Cohen's κ for a 2×2 contingency table

	Columns	
Rows	A=2	B=14
	C=12	D=5
$Cohen's\kappa = \frac{2(AD-BC)}{(A+B)(B+D)+(C+D)(A+C)}$		

Although formulae exist for estimating the confidence interval for κ, they do not perform well in all circumstances (Roldán Nofuentes, Luna del Castillo, & Montero Alonso, 2009), so bootstrapping is recommended. We created 2000 bootstrap samples and calculated κ in each; we, therefore, ended up with 2000 κ values. Figure 4 shows a histogram of these values. The middle 95% of these values, the percentile bootstrap confidence interval, ranged from −.813 to −.269. By using bootstrapping, a task

(computing a CI for κ) that would have been extremely complicated and involved complex formulae that yield sub-optimal results has become simple (and more accurate).

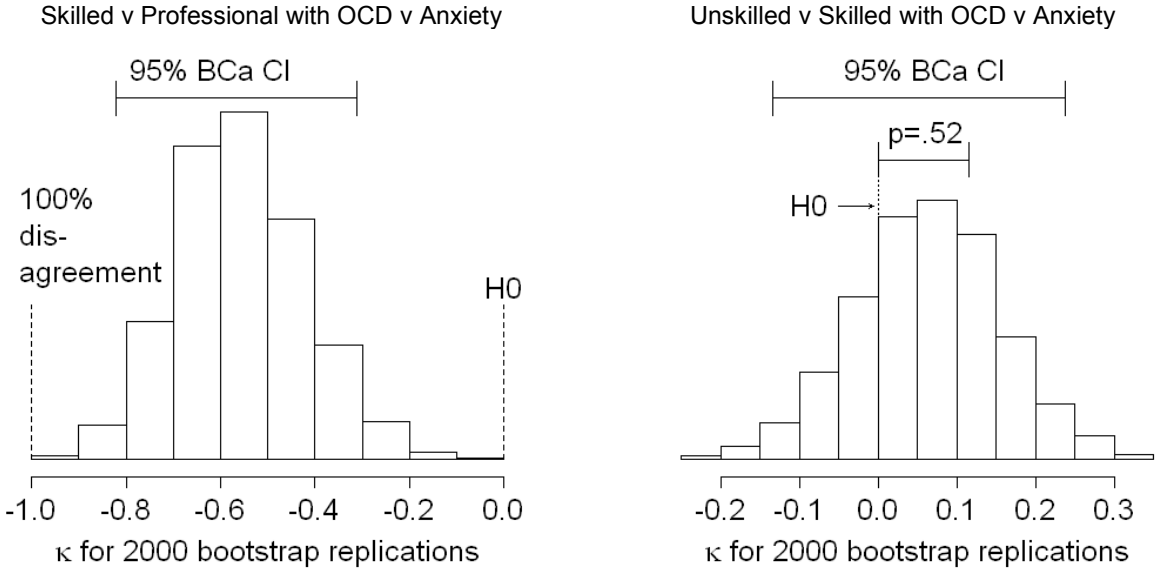

Figure 4: Distribution of 2000 bootstrap sample κ from the data in Table 1 comparing the fathers' occupation with child's diagnosis. The left panel compares skilled versus professional with OCD versus anxiety. The right panel compares unskilled versus combined skilled and professional with the control group versus the combined OCD and anxiety conditions.

The percentile bootstrap confidence interval that we have described above is not the only method for computing bootstrap CIs. Efron and colleagues have developed improved methods and recommend the *bias-corrected and accelerated* or BCa method. Using this method for the OCD data yields a 95% CI of −.823 to −.311. If the mean of the bootstrap statistics is biased (i.e., it is not the sample value) then the BCa method helps to correct the bias. The acceleration refers to the limits of the confidence interval converging more quickly. Several papers have shown that this alternative tends to produce more accurate intervals than the percentile method (Efron & Tibshirani, 1993, Chapter 22).

κ can be calculated for the association of other 2×2 comparisons that can be made from the data in Table 1. For example, suppose you wanted to compare unskilled versus the combined skilled and professional occupation groups for the two clinical groups (OCD and anxious) compared with controls (in the terms of Table 2, A=16, B=33, C=13, D=36). The κ is:

$$\kappa = \frac{2((16)(36)-(13)(33))}{(16+33)\cdot(33+36)+(13+36)\cdot(16+13)} = .061$$

The 95% BCa confidence interval for this is from −.133 to .238; because this interval overlaps with 0 most researchers would fail to reject the null hypothesis at α=.05.

Bootstrapping Correlations

Much psychopathology research (and psychology research more generally) involves comparing variables like rating scales or reaction times. Researchers are able to conceptualize these as being derived from some continuous psychological scale. Sometimes they pretend the data are normally distributed, but as we have argued, this assumption is seldom valid (Micceri, 1989). As Wilcox (1998) shows, falsely assuming that data are normally distributed is problematic. One of the most often used statistics is Pearson's correlation. Consider some recent data from Cartwright-Hatton et al. (in press).

This project investigated a parent-based intervention for anxiety disorders in very young children (under 10 years). The main outcome data are reported in Cartwright-Hatton et al. (in press), so here we focus on the relationship between age and anxiety as measured by the child-behaviour checklist (DSM subscale), CBCL-DSM. Figure 5 shows a scatterplot between age of 58 children and their anxiety score (upon entering the trial) on the TCBCL-DSM.

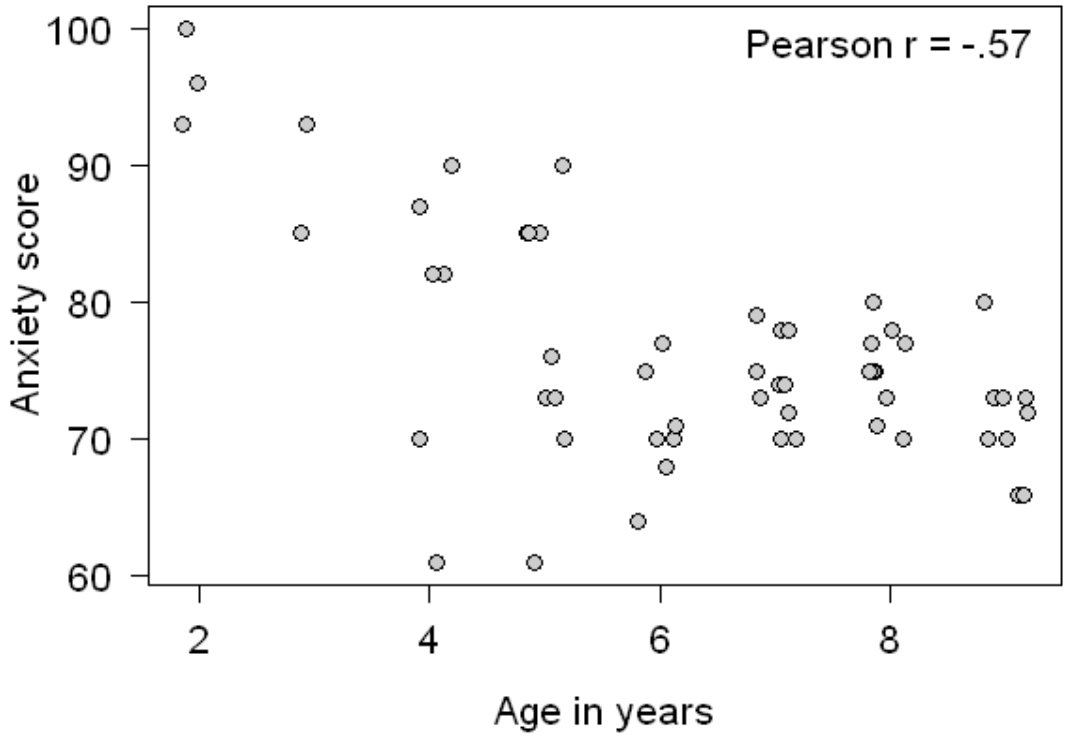

Figure 5: A scatter plot of anxiety with child's age. A small random jitter has been added to the age values so that all the data points can be seen. Data from Cartwright-Hatton et al. (in press).

The data in Figure 5 appear to show an odd pattern. With the small number of cases we will not speculate on possible grouping from this scatter plot, but it appears not to have bivariate normality. This is confirmed with a multivariate version of the Shapiro-Wilk statistic, MVW = .96, p = .03 (Villasenor-Alva & Gonzalez-Estrada, 2009). This result makes the traditional confidence interval for the correlation suspect. Instead, a bootstrap method can be used.

As before, cases are resampled to create bootstrap samples. Pearson's correlation coefficient, -.58 for the original sample, is calculated for each of these bootstrap samples. Figure 6 shows the histogram of the correlations from 2000 bootstrap samples. The BCa 95% confidence interval goes from −.76 to −.28. The data in Figure 5 suggest the relationship may not be linear. Spearman's ρ (rho) has the value -.42 for the original sample. The BCa 95% bootstrap CI is -.65 to -.11.

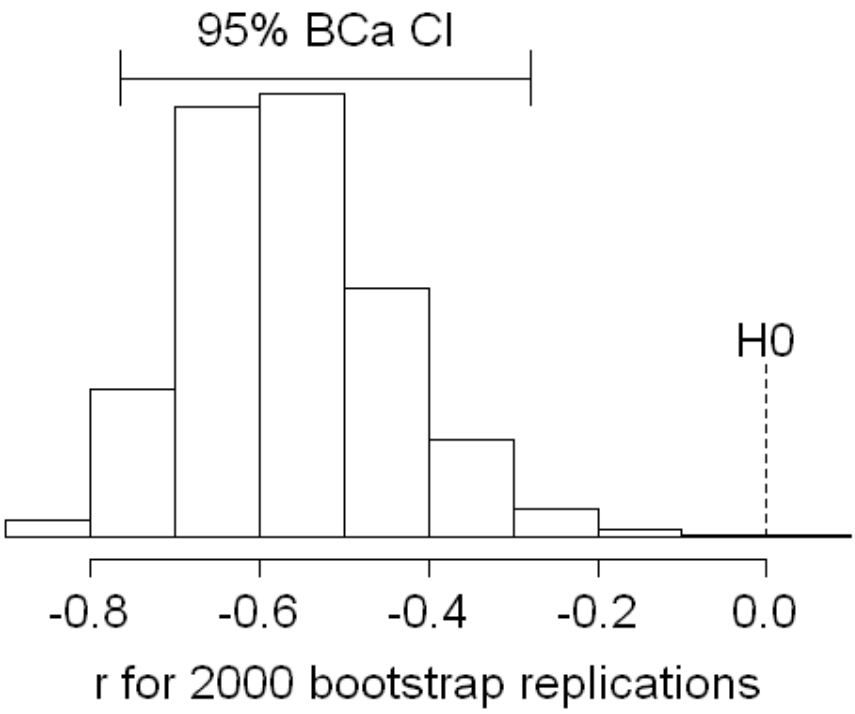

Figure 6: A histogram of the correlations from 2000 bootstrap samples. Data from Cartwright-Hatton et al. (in press).

Bootstrapping Regression Coefficients

Just as we can bootstrap correlation coefficients, we can also bootstrap regression parameters (i.e., the intercept and any predictors in the model). For example, using the correlation example from above, we could produce the following regression model:

$$Anxiety\ (TCBCL)_i = b_0 + b_1 Initial\ Age_i + \varepsilon_i$$

In which anxiety (as measured by the child's initial TCBCL-DSM score) is predicted from their age at entering the trial. Bootstrapping provides estimates of the confidence intervals around b_0 and b_1 without making as many assumptions about the shape distribution of residuals. As for the correlation coefficient, cases are resampled to create 2000 bootstrap samples. The regression coefficients, *b*s, are calculated for each of these samples. Figure 7 shows the histogram of the regression coefficients for initial age (as a predictor of anxiety) from 2000 bootstrap samples. The BCa 95% confidence interval goes from −3.20 to −1.18.

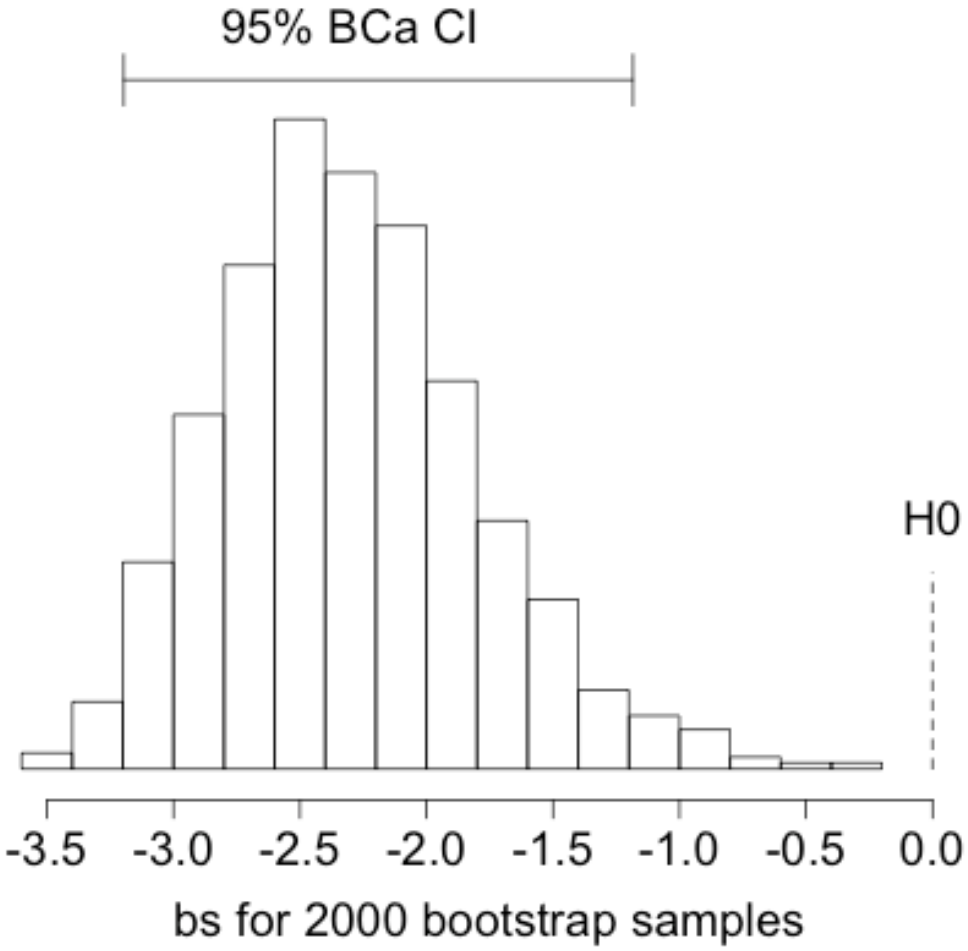

Figure 7: A histogram of the regression coefficient, b_1, of child's age as a predictor of initial anxiety (CBCL) from 2000 bootstrap samples. Data from Cartwright-Hatton et al. (in press).

Other Statistics

Bootstrapping can be used in a vast number of situations. In this paper only the tip of the iceberg has been touched, and to ground the reader in the basic principles without getting bogged down in complexity we have focussed on simple situations familiar even to those with only a basic level of statistics training. For example, having shown you that it is straightforward to bootstrap a regression coefficient, *b*, it is not difficult to see how we could, therefore, use bootstrapping in more advanced forms of regression such as structural equation modelling, hierarchical linear models, logistic regression, mediation analysis, and any other situation in which *b* is estimated or a general linear model is applied to the data (i.e. ANOVA, ANCOVA, MANOVA etc.). Examples in this issue on bootstrapping statistics for mediation analysis are given in Sadler and colleagues (2011) and Woody (2011). By extension of our regression example, it should be clear that the possibilities for applying the basic bootstrapping technique are nearly endless.

This flexibility empowers us to use bootstrapping in situations for which traditional nonparametric tests are not available. Most clinicians and researchers will be familiar with rank-based tests such as the Wilcoxon, Mann-Whitney, Kruskal-Wallis and Friedman tests as replacements for *t*-tests and one-way ANOVA when normality assumptions have not been met. However, research designs are seldom as

simple as comparing two or more groups on a single independent variable. More often than not, there are two independent variables, mixed designs, or perhaps a covariate to be added. When the assumptions of tests based on the normal distribution are not met and the researcher has one of these designs they will find no traditional nonparametric test to come to their aid. The typical response might, therefore, be to pretend there is not a problem and resort to the usual ANOVA (often with some mis-informed statement that the *F*-statistic is robust). In fact, the *F*-statistic performs very strangely in some situations (see Wilcox, 2005, or Field, 2009, for a less technical summary of the issues). However, there are versions of ANOVA (factorial, independent, repeated measures and mixed) and ANCOVA based on bootstrapping that do provide robust tests of group differences. Wilcox (2005) has implemented many of these tests in the statistical package R (see below).

Originally many people were suspicious of bootstrapping: it seemed too good to be true. When Efron and others laid down the mathematical justification for these methods, their use in statistics quickly spread. There are still some bootstrapping problems that statisticians are working on, for example when the parameter of interest can arise from many different patterns of data (like a multiple R^2 or the loadings of a principal component analysis) or when one parameter is highly dependent on others as in some time series analyses. Details on these more complex cases are covered in the textbooks listed in the references (e.g., Chernick, 2008; Efron, & Tibshirani, 1993; Good, 2006; Lunneborg, 2000).

Doing bootstrapping

Most of the main statistics packages have some facilities for bootstrapping. For example, SPSS, a particularly popular piece of software for psychologists, recently implemented a bootstrap add-on to their core package. One of us (APF) has created a webcast for doing bootstrapping with SPSS, which is available at http://www.statisticshell.com/woodofsuicides.html. This is a good way to get started in a package that is probably familiar to you. It is also possible to bootstrap regression parameters in AMOS and a basic guide to this process in the context of mediation analysis can be found at http://amosdevelopment.com/video/indirect/flash/indirect.html.

However, although SPSS offers some scope for bootstrapping, it is quite limited in what it can do.

We believe that the most versatile software package for applying bootstrapping is R (e.g., R Development Core Team, 2010). R is an environment/language for statistical analysis and is the fastest growing statistics package. It is freely available from cran.r-project.org. If you have never used R it will be valuable to read some introduction to R, for example Chapter 3 of Field and Miles (in press) or Chapter 1 of Wright and London (2009). There are thousands of R packages that can be freely accessed from within R. The R code for all of our examples is linked from this article as an online resource. R is a command language and so we type in commands that R then executes; when referring to commands typed into R we will use blue text. R is a powerful programming language, but to the uninitiated it appears bewildering. Many people find learning to use R difficult, but if you plan to do much data analysis and graphing we believe the initial difficulty will be worth it. These brief tutorials should get you started.

The main R URL is: http://www.r-project.org/. Much information is available there including instructions on downloading the package and online tutorials.

Using R to Bootstrap Estimates for the Median and Mean

In this section we briefly describe how to use the R package *boot* (Canty & Ripley, 2010; Davison & Hinkley, 1997). Although there are several packages for bootstrapping, the package *boot* is the most recommended. To have access to the functions within the boot package, it has to be installed and

loaded. When you download R the package boot should be installed automatically (if not, type *install.packages("boot")*), but the functions will not be part of your active R session unless you load the package using the command library(boot).

To illustrate the package *boot*, consider the 10 cases from the multiple personality data illustrated in Figure 2. R works by creating objects from functions. The object-oriented aspect of R makes it particularly powerful (Chambers, 2008). An object for the data can be created with:

```
MultiplePersonalityData <- c(1,1,2,2,4,5,7,15,30,56)
```

This function creates an object called 'MultiplePersonalityData' that is a variable (sometimes referred to as a vector) containing the numbers listed in parentheses. In short, it is like a column representing a variable in Excel or SPSS.

Suppose we wanted to find the 95% BCa bootstrap estimate for the median as we did when we discussed this example above. Within R, the functions *mean* and *median* take a variable as its input, and then output the mean and median respectively:

```
mean(MultiplePersonalityData)
median(MultiplePersonalityData)
```

In this case R produces 10.7 for the mean and 4.5 for the median. The boot function has the following generic form:

```
boot(data, function, replications)
```

in which *data* names the data object to be bootstrapped and *function* represents the thing that you want calculated (e.g., the mean). The value that you input for *replications* is the number of bootstrap replications. There are several other options, but these will suffice for our purposes. To calculate the mean and the median for 2000 bootstrap samples from the original sample of *MultiplePersonalityData*, type:

```
boot_mean<-boot(MultiplePersonalityData, function(x,i) mean(x[i]), 2000)
boot_median<-boot(MultiplePersonalityData, function(x,i) median(x[i]), 2000)
```

The only confusing part of these commands is why, for example, *function(x,i) median(x[i])* has to be written rather than just 'median'. The short answer is: "It just has to". The longer answer is that the original variable is x, and x[i] refers to one of the 2000 bootstrap samples. This extra notation is necessary if bootstrapping more complex functions. These two commands create objects called *boot_median* and *boot_mean* respectively (in less than a second unless you are using a computer relic).

boot_median and *boot_mean* are bootstrap objects. Objects in R vary in complexity. Both of these objects are a list of many different sets of numbers. It includes the value for the original sample. It calls this t0. To access just this but you need to use the $ symbol. For most complex objects in R the $ is used to access different parts of it. Thus, if you type *boot_median$t0*, then the median of original sample will be printed on the screen. The object also includes the values for all the bootstrap samples, called t. If you type *boot_median$t* the values for the median of all 2000 bootstrap samples will be printed on the screen. The object *boot_median$t* is a variable with 2000 cases. This variable can be used in other functions. As such, if you wanted to plot a histogram showing the distributions of the medians and means for the 2000 bootstrap samples, you could use these commands:

```
hist(boot_median$t)
hist(boot_mean$t)
```

There is another function, *boot.ci*, which creates confidence intervals for bootstrap objects. To get the 95% confidence intervals for the mean and median, we can use these commands:

```
boot.ci(boot_median, type = c("perc","bca"), conf = .95)
boot.ci(boot_mean, type = c("perc","bca"), conf = .95)
```

The *boot.ci* command takes the bootstrap object (e.g., *boot_median*, which we created earlier) and makes bootstrap confidence intervals. We can specify the type of confidence interval that we want: “perc” uses the percentile method, “bca” uses the bias corrected accelerated method, and in this example we have asked for both. Finally, the command “conf=.95” tells the computer to compute the 95% confidence interval. If you prefer a 90% confidence interval change this command to "conf = .90" (95% confidence intervals are the default so you do not need to type "conf = .95").

Using R to Bootstrap Estimates for the Correlation Coefficient

For this example, we will look at the correlation between the age of a child entering a treatment trial (*AgeInitial*) and their initial levels of anxiety (*TCBCL_DSM*). These data come from Cartwright-Hatton et al.’s (in press) data set and are stored in the file **cartwright-hatton_bootstrap.sav**. By storing the data in an SPSS file we can show you how to access these files in R. To enable R to read SPSS files we use two commands; the first is to load a package called *foreign* (in case you do not have this installed), and the second initiates this package in the current session:

```
install.packages("foreign")
library(foreign)
```

Next we need to choose the SPSS data file and import it into R. The following two commands do this; *file.choose* opens a dialog box that enables us to select the file in the usual way in Windows or MacOS, it then sets the object, which we have called ‘filename’ to be this file. The *read.spss* command takes the object ‘filename’ (which we defined before as the file that you have just chosen) and reads it into a data frame that we have called *cartwrightData*. (You can call the data frame whatever you like but it makes life easier, especially when looking back at code that you wrote months before, if you assign sensible, informative names to objects that you create in R.)

```
filename <- file.choose()
cartwrightData <- read.spss(filename, use.value.labels=FALSE, to.data.frame=TRUE)
```

We now have a data frame called *cartwrightData* that contains all of the variables that were in the SPSS files. The variables will have the same name as in the SPSS file, and we can refer to them by appending $ and the name of the variable to the name of the data frame. For example, to access the variable *AgeInitial* in the data frame *cartwrightData* we would write *cartwrightData$AgeInitial*.

We can use these variables now to compute correlation coefficients in each of several bootstrap samples. Remember for the mean and median we had to define the function that we wanted to bootstrap. We have the same process for the correlation coefficient except that the function is now more complex; because of this we have chosen to define the function as a separate object (which we have called *pearsonR*). As with the mean and median, we use a command called function, name the data to be used (*cartwrightData*) and use *i* to denote each bootstrap sample. However, in this case we use *cor* to compute the correlation between the variables *AgeInitial* (remember we have to define this in full as *cartwrightData$AgeInitial*) and *TCBCL_DSM_Pre* for each sample, *i*. The ‘use = "complete.obs"’ option is there because we had missing data in the original file; this tells R to use complete observations only (i.e., cases for which there is a score for both initial age and anxiety on the CBCL). One of the most

common errors in using the *boot* command is getting the square brackets in the right places so be careful to define your function properly.

```
pearsonR <- function(cartwrightData,i) cor(cartwrightData$AgeInitial[i],
    cartwrightData$TCBCL_DSM_Pre[i], use = "complete.obs")
```

Having created the object, the function *pearsonR*, which computes the correlation coefficient between the two variables of interest, we can insert this object into the *boot* command as we did for the mean and median. The code below creates a new object called *boot_correlation* that is made from bootstrapping the *pearsonR* function (created above) in the *cartwrightData* data frame, and replicating 2000 times. The confidence intervals can then be obtained as for the mean and median by using the *boot.ci* command. (The code below shows how to ask for 95% percentile and BCa confidence intervals in separate commands; note that we have not used 'conf=' because 95% is the default.)

```
boot_correlation <- boot(cartwrightData, pearsonR, 2000)
boot.ci(boot_correlation, type="perc")
boot.ci(boot_correlation, type="bca")
```

Using R to Bootstrap Regression Parameters

Most of the traditional statistics, like the standard deviation and ANOVA, are greatly affected by outliers. When a statistical technique is greatly affected by a few points statisticians say it is not robust. When it is not greatly affected by a few outliers they say it is robust. There are a large set of robust techniques available. One set of robust functions is from Rand Wilcox and described in his 2005 book. These functions are stored on his web page, and to access these functions in R we need to type:

```
source("http://www-rcf.usc.edu/~rwilcox/Rallfun-v13")
```

For this to work you need to be connected to the Internet. Wilcox updates these functions quite regularly so if you get an error when you run the command it might be because this web address no longer exists; in which case try replacing 'v13', which is the version number, with a higher number such as 'v14'. The functions are currently being put together as a package, which will make them easier to access, but at present it is better to access them from his web page.

Let's consider two of his functions. The first is *tsreg*, which estimates the regression coefficients (the *b*s) using something called the Theil-Sen estimator (see Wang, 2005; Wilcox, 1998, 2005, for details). The second function is *regci*, which calculates bootstrap confidence intervals around the Theil-Sen estimated *b*s. Both of these functions take the general form *function(predictor, outcome*), as such, all we need to do is specify the predictor variable (*cartwrightData$TCBCL_DSM_Pre*) and the outcome variable (*cartwrightData$AgeInitial*). For *regci* we should also specify the number of bootstrap samples using *nboot* because the default is 599 and here we use 2000. Having sourced Wilcox's functions above, we can type the following to get the robust regression coefficients and their bootstrapped confidence intervals:

```
tsreg(cartwrightData$AgeInitial, cartwrightData$TCBCL_DSM_Pre)
regci(cartwrightData$AgeInitial, cartwrightData$TCBCL_DSM_Pre, nboot = 2000)
```

The benefits of bootstrapping

When Efron (1979) first introduced the bootstrap most statisticians and scientists were skeptical. It seemed too easy, particularly compared with the difficult mathematical tasks it was replacing (Chernick, 2008). Bootstrapping began to catch on only when a large number of statisticians showed that in

situations where traditional statistics were known to work (i.e., with normal distributions which Micceri, 1989, said are as rare in psychology as unicorns are in nature) bootstrap methods performed as well and in other situations they performed better. There are situations in which the simple bootstrap has difficulties. For example, although Chernick (2008, p. 174) argues that samples as small as n = 20 work with some problems, he says a good rule of thumb is have at least n = 50. For some procedures like longitudinal methods more complex bootstrapping procedures are necessary. Also, because of the way bootstrapping is conducted, the importance of having a sample that is truly representative of the population is essential (but this is true with traditional methods also).

The bootstrap is a paradigmatic shift in the mindset of doing statistics from problems that are either mathematically hard or yet-to-be-solved, to problems that can be solved with modern computing and minimal mathematics. In the past, psychologists often tried to fit their hypotheses to a small but well-known set of statistics with mathematical formulae for calculating the standard error, and then acted as if the distributional assumptions of these methods were met. Bootstrapping allows psychologists to avoid this restrictive and often delusional approach. The advantages of bootstrap estimation over traditional estimation include: (1) Statisticians have shown it works better; (2) It is much more flexible; and (3) It is easy and automatic— once you decide on the statistic, then the computer does the hard work.

Bootstrapping also offers a pedagogical advantage. For decades people have taught statistics with resampling to avoid the formal mathematics of the traditional methods. For example, before personal computers, Simon (1969a, b) showed how playing cards could be used to teach probability and statistics by shuffling the deck and then dealing the cards repeatedly. Most of his examples have sampling without replacement so his approach differs from Efron's bootstrap, but he gives similar intuitive arguments. He showed this approach worked with high school students and argued it could be used with pre-high school students. More recently Good (2006) and Lunnerborg (2000) have texts that take this approach and are appropriate for psychology undergraduates.

Conclusions

Bootstrapping is a highly flexible method that allows the estimation of different statistics. Bootstrapping is less hampered by standard distributional assumptions of some tests than traditional methods for constructing confidence intervals. Further, it is flexible enough to be applied in situations where no formulae exist for calculating standard errors and confidence intervals. In this paper, we have argued that bootstrapping is an appropriate technique to use in a variety of situations. We echo Howell (2007) in arguing that bootstrapping procedures are bound to become more commonplace in standard psychology training and empirical articles. The age of high powered personal computers is here and with it new and more robust statistical methods are emerging.

Acknowledgements

We are grateful to Shirley Reynolds and Sam Cartwright-Hatton for allowing us to use their data sets for our examples.

References

Agresti, A. (2002). *Categorical data analysis (2nd ed.)*. Hoboken, NJ: John Wiley & Sons. doi:10.1002/0471249688

Arostegui, I., Nunez-Anton, V., & Quintana, J. M. (2007). Analysis of the short form-36 (SF-36): The beta-binomial distribution approach. *Statistics in Medicine, 26*(6), 1318–1342. doi:10.1002/sim.2612

Barber, J. A., & Thompson, S. G. (2000). Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine, 19*(23), 3219–3236.

Binks, C. A., Fenton, M., McCarthy, L., Lee, T., Adams, C. E., & Duggan, C. (2006). Psychological therapies for people with borderline personality disorder. *Cochrane Database of Systematic Reviews*(1). doi:10.1002/14651858.cd005652

Canty, A. & Ripley, B. (2010). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.2-43. cran.r-project.org..

Cartwright-Hatton, S., McNally, D., Field, A. P., Rust, S., Laskey, B., Dixon, C … & Woodham, A. (In Press). A new parenting-based group intervention for young anxious children: Results of a randomized controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry.*

Casella, G. (2003). Introduction to the Silver Anniversary of the Bootstrap. *Statistical Science*, *18,* 133–134. doi:10.1214/ss/1063994967

Chambers, J.M. (2008). *Software for data analysis: Programming with R.* New York: Springer. doi:10.1007/978-0-387-75936-4

Chernick, M.R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed). Hoboken, NJ: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. doi:10.1177/001316446002000104

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge UK: Cambridge University Press.

Delucchi, K. L., & Bostrom, A. (2004). Methods for analysis of skewed data distributions in psychiatric clinical studies: Working with many zero values. *American Journal of Psychiatry, 161*(7), 1159–1168. doi:10.1176/appi.ajp.161.7.1159

Derisley, J., Libby, S., Clark, S., & Reynolds, S. (2005). Mental health, coping and family-functioning in parents of young people with obsessive-compulsive disorder and with anxiety disorders. *British Journal of Clinical Psychology, 44*, 439-444. doi: 10.1348/014466505x29152 doi:10.1348/014466505X29152

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7,* 1–26. doi:10.1214/aos/1176344552

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Field, A. P. (2009). *Discovering statistics using SPSS: and sex and drugs and rock 'n' roll* (3rd edition). London: Sage.

Field, A. P., & Miles, J. N. V. (in press). *Discovering statistics using R: and sex and drugs and rock 'n' roll*. London: Sage.

Gilchrist, J.M. (2009). Weighted 2×2 kappa coefficients: Recommended indices of diagnostic accuracy for evidence-based practice. *Journal of Clinical Epidemiology*, *62*, 1045-1053. doi:10.1016/j.jclinepi.2008.11.012

Goff, D.C. & Simms, C.A. (1993). Has multiple personality disorder remained consistent over time? *Journal of Nervous and Mental Disease*, *181*, 595–600. doi:10.1097/00005053-199310000-00003

Good, P.I. (2006). *Resampling methods: A practical guide to data analysis* (3rd ed). Boston: Birkhäuser.

Hlatky, M. A., Boothroyd, D. B., & Johnstone, I. M. (2002). Economic evaluation in long-term clinical trials. *Statistics in Medicine, 21*(19), 2879–2888. doi:10.1002/sim.1292

Howell, D.C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.

Kraemer, H.C., & Gibbons, R.D. (2009). Where do we go wrong in assessing risk factors, diagnostic and prognostic tests? The problems of two-by-two association. *Annals of Psychiatry*, *39,* 711–718. doi:10.3928/00485713-20090625-05

Libby, S., Reynolds, S., Derisley, J., & Clark, S. (2004). Cognitive appraisals in young people with obsessive-compulsive disorder. *Journal of Child Psychology and Psychiatry, 45*(6), 1076-1084. doi:10.1111/j.1469-7610.2004.t01-1-00300.x

Lunneborg, C.E. (2000). *Data analysis by resampling: Concepts and applications.* Pacific Grove, CA: Duxbury Press.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105,* 156–166. doi:10.1037/0033-2909.105.1.156

Picardi, A., Battisti, F., De Girolamo, G., Morosini, P., Norcio, B., Bracco, R., & Biondi, M. (2008). Symptom structure of acute mania: A factor study of the 24-item Brief Psychiatric Rating Scale in a national sample of patients hospitalized for a manic episode. *Journal of Affective Disorders, 108*(1–2), 183–189. doi:10.1016/j.jad.2007.09.010

R Development Core Team. (2010). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, Retrieved from http://www.R-project.org

Roldán Nofuentes, J.A., Luna del Castillo, J.D., & Montero Alonso, M.A. (2009). Confidence intervals of weighted kappa coefficient of a binary diagnostic test. *Communications in Statistics: Simulation and Computation*, *38,* 1562–1578. doi:10.1080/03610910903039473

Rutter, C. M., & Miglioretti, D. L. (2003). Estimating the accuracy of psychological scales using longitudinal data. *Biostatistics, 4*(1), 97–107. doi:10.1093/biostatistics/4.1.97

Sadler, P., Ethier, N. & Woody, E. (2011). Tracing the interpersonal web of psychopathology: Dyadic data analysis methods for clinical researchers. *Journal of Experimental Psychopathology, 2*(2) 95-138. doi:10.5127/jep.010310

Simon, J. L. (1969a). *Basic research methods in social science: The art of empirical investigation*. New York: Random House.

Simon, J. L. (with Holmes, A.) (1969b). A new way to teach probability statistics. *The Mathematics Teacher*, *62,* 283–288.

Tang, L. Q., Song, J. W., Belin, T. R., & Unutzer, J. (2005). A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine, 24*(14), 2111–2128. doi:10.1002/sim.2099

Tyrer, P., Nur, U., Crawford, M., Karlsen, S., McLean, C., Rao, B., & Johnson T. (2005). The social functioning questionnaire: A rapid and robust measure of perceived functioning. *International Journal of Social Psychiatry, 51*(3), 265–275. doi:10.1177/0020764005057391

Villasenor-Alva, J. A. & Gonzalez-Estrada, E. (2009). A generalization of Shapiro-Wilk's test for multivariate normality. *Communications in Statistics: Theory and Methods*, *38,*1870–1883. doi:10.1080/03610920802474465

Wang, X. (2005). Asymptotics of the Theil-Sen estimator in the simple linear regression model with a random covariate. *Journal of Nonparametric Statistics, 17(1)*, 107-120. doi:10.1080/1048525042000267743

Wilcox, R. R. (1998). A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. *Biometrical Journal, 40(3),* 261-268.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). Burlington, MA: Elsevier.

Wilkinson, L., and the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54,* 594–604. doi:10.1037/0003-066X.54.8.594

Woody, E. (2011). An SEM perspective on evaluating mediation: What every clinical researcher needs to know. *Journal of Experimental Psychopathology*, *2*(2), 210-251. doi:10.5127/jep.010410

Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology*, *73,* 123–136. doi:10.1348/000709903762869950

Wright D.B. & London, K. (2009). *Modern regression techniques using R: A practical guide for students and researchers*. London: Sage.

Zimmerman, M., Chelminski, I., & Posternak, M. (2004). A review of studies of the Hamilton depression rating scale in healthy controls - Implications for the definition of remission in treatment studies of depression. *Journal of Nervous and Mental Disease, 192*(9), 595–601. doi:10.1097/01.nmd.0000138226.22761.39