RESAMPLING METHODS:
Not just for statisticians anymore.

William M. Duckworth and W. Robert Stephenson, Iowa State University
Dept. of Statistics, 327 Snedecor Hall, Iowa State University, Ames, IA 50011-1210

Key Words: Bootstrap, Jackknife, Teaching Statistics, Health Sciences

Abstract

Resampling methods in statistics have been around for a long time. Over forty years ago Tukey coined the term jackknife to describe a technique, attributed to Quenouille (1949), that could be used to estimate bias and to obtain approximate confidence intervals. About 20 years later, Efron (1979) introduced the "bootstrap" as a general method for estimating the sampling distribution of a statistic based on the observed data. Today the jackknife and the bootstrap, and other resampling methods, are common tools for the professional statistician. In spite of their usefulness, these methods have not gained acceptance in standard statistics courses except at the graduate level. Resampling methods can be made accessible to students at virtually every level. This paper will look at introducing resampling methods into statistics courses for health care professionals. We will present examples of course work that could be included in such courses. These examples will include motivation for resampling methods. Health care data will be used to illustrate the methods. We will discuss software options for those wishing to include resampling methods in statistics courses.

1 Background

Over forty years ago John Tukey coined the term jackknife for a rough and ready tool that could be used to come up with approximate confidence limits. Tukey's (1958) jackknife was based on an idea of Quenouille (1949, 1956) of using parts of a sample to estimate bias and thus come up with an estimator with reduced bias. The jackknife is based on the idea of computing estimates with the i^{th} sample observation omitted. Twenty years later Efron (1979) laid out the theoretical groundwork for the bootstrap as a generalization to the jackknife. Since that time, statisticians have embraced the jackknife, bootstrap and other resampling methods. Statisticians have used these methods extensively in their research. A quick literature search turns up over 175 articles using these methods in 2002 alone.

Although statisticians have embraced resampling methods for their own use they have not, in general, included them in their teaching. The exception to this rule are Julian Simon and Peter Bruce. Julian Simon hit upon using resampling in teaching statistics prior to Efron's seminal paper on the bootstrap. Since that time Simon and Bruce have crusaded for the use of resampling in the teaching of statistics at all levels. They have developed a computer program (Resampling Stats) to simplify the computing aspect of resampling.

Several articles in *Teaching Statistics* have dealt with the use of resampling and the bootstrap in teaching statistics courses. Ricketts and Berry (1994) discuss using resampling to teach hypothesis testing. They compare the means of two independent samples using the Resampling Stats program mentioned above. Taffe and Garnham (1996) discuss using resampling to accomplish estimation of a population mean based on a single sample. They also provide a Minitab macro for the comparison of means of two independent samples. Johnson (2001) presents bootstrap methods for estimating standard errors and constructing confidence intervals. All of these articles discuss topics that are core to introductory statistics courses. Hesterberg (1998) gives a very nice review of simulation and bootstrapping in teaching statistics. There is practical advice on software for simulation and the bootstrap. There is also an extensive set of references.

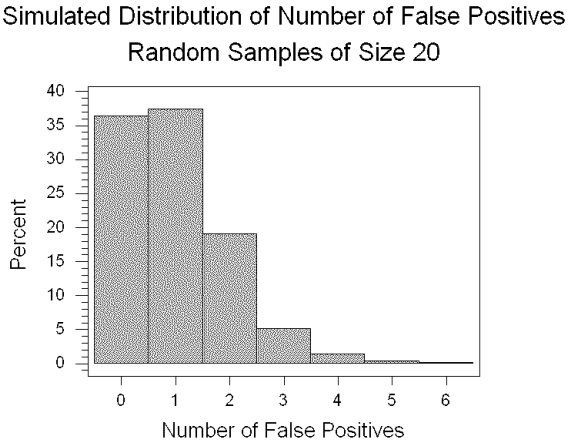
Resampling methods and the bootstrap have even found their way into published textbooks for a beginning course in statistical methods. *Statistics: Making Sense of Data*, 2^{nd} Ed. by Stout, Travers and Marden (1999) uses the bootstrap to estimate

the standard error of the sample median, test a hypothesis for a population mean and to test for differences among population means. *The Practice of Business Statistics* by Moore, *et. al.* (2003) has an optional companion chapter, primary author Tim Hesterberg, entitled "Bootstrap Methods and Permutation Tests".

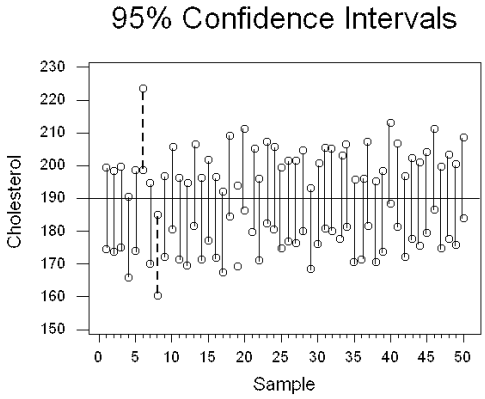
Given that much has already been said about resampling and the bootstrap, what do we have to offer? There is no doubt in our mind that simulation and elements of resampling and the bootstrap have a place in beginning statistics courses. What then, as a first step, should be included in a first statistics course for health care professionals?

2 Simulation

Simulation should be a part of every introductory statistics course. Many of the core concepts in probability and inference rely on an understanding of long run relative frequency arguments. Simulation can be used to make students more comfortable with these arguments. For example, everyone knows that diagnostic tests are not 100% accurate. Health care professionals should be aware of the sensitivity and specificity of diagnostic tests. They should be aware of the idea, and associated probabilities, of a false positive and a false negative result. How can one introduce the idea of the chance of 2 false positives in a random sample of 20 diagnostic tests? One could try to introduce the binomial probability function or one could simulate. If the probability of a false positive is 0.05, the simulated probability of 2 false positives in a random sample of 20 is approximately 0.20.



In addition to the simulation of probabilities, simulation of random sampling from a population is helpful when discussing the sampling distribution of a sample statistic or the interpretation of what 95% confidence means. There are several interactive applets available on the web. Two we have found useful look at the sampling distribution of the sample mean **http://www.ruf.rice.edu/∼lane/stat_sim/sampling_dist/index.html** and the interpretation of confidence **http://www.stat.sc.edu/∼west/javahtml/ConfidenceInterval.html**. The latter simulation looks at how many times a confidence interval covers the true population mean. From this we try to get students to see that confidence refers to the method of constructing confidence intervals. That is, when one repeatedly constructs 95% confidence intervals based on random samples, about 95% of the intervals will cover the population mean.



3 The Bootstrap

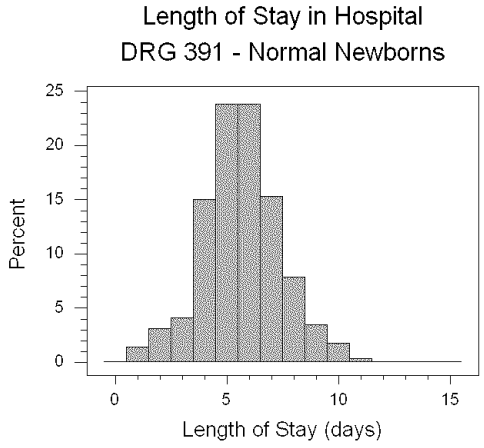
Once students are familiar with simulation, especially simulated sampling from a known population, it is fairly easy to introduce the idea of resampling. It is important to differentiate between simulated sampling from a population and resampling from the sample. These are different operations. However, what one gets out is a distribution, either a simulated sampling distribution or a resampling distribution, of possible values for a sample statistic.

3.1 Confidence Intervals

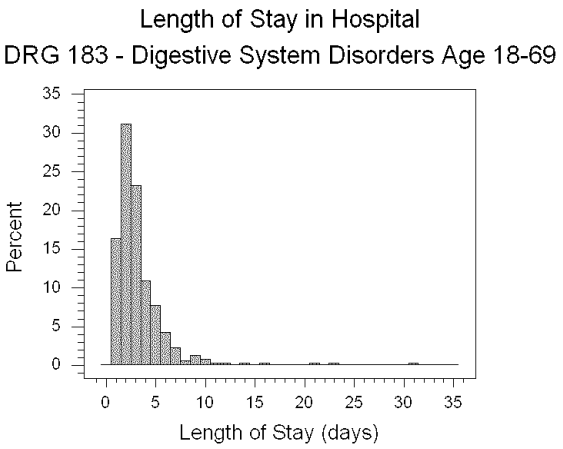
The usual treatment of confidence intervals in most introductory courses begins with the sampling distribution of a statistic, like the sample mean. From the sampling distribution, one argues that 95% of the time the sample mean, $\overline{X}$, will fall between

$\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$. Through some simple algebra, the random interval $\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ will cover the population mean, μ, 95% of the time. If the population standard deviation is not known, them the random interval becomes $\overline{X} \pm t^*\frac{s}{\sqrt{n}}$. The value of t^* is obtained from a table of the t-distribution with $n-1$ degrees of freedom and the appropriate confidence level. The necessary condition for the latter interval is for the original measurements to come from a normal, or at least approximately, normal distribution.

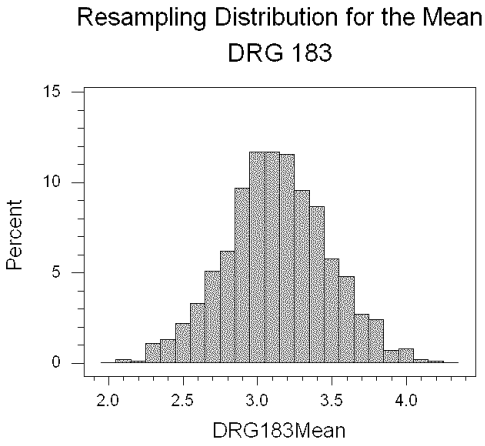
To illustrate the usual construction of a confidence interval we consider the length of stay in hospital for normal newborns (Diagnostic Related Group (DRG) 391). A normal newborn is actually a baby that experiences some minor problems at birth and is considered an admission separate from that of the mother. A random sample of 20 newborns, DRG 391, is selected and the length of stay noted. The sample statistics are $\overline{X} = 5.6$ and $s = 2.16$. A 95% confidence interval for the population mean length of stay is $5.6 \pm 2.09\frac{2.16}{\sqrt{20}}$ or 5.6 ± 1.0. It turns out the the distribution of length of stay for DRG 391 is symmetric and mounded in the middle as seen in the histogram below.



Continuing with another DRG, a random sample of 20 patients, DRG 183 - Digestive system disorders age 18-69, is taken and sample statistics calculated. The sample mean is $\overline{X} = 3.1$ and sample standard deviation is $s = 1.65$. This presents a problem as the relatively large sample standard deviation, and the fact that the shortest length of stay is 1 day, indicates that the distribution of length of stay for this DRG is most likely skewed. The distribution of length of stay for DRG 183 is highly skewed as indicated by the histogram below.



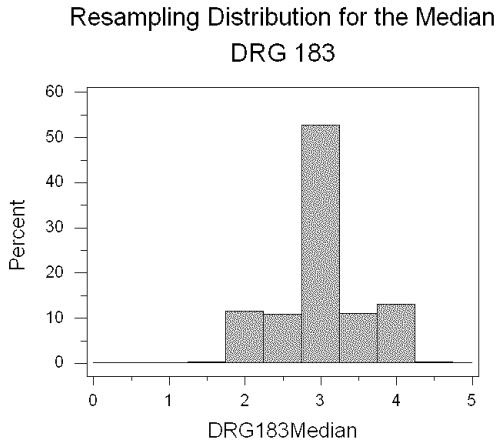
At this point we are confronted with a dilemma. Should we proceed with the normal theory method even though we have concerns that this is inappropriate? Should we stop and say we do not know what to do or that methods for proceeding are beyond the scope of the introductory course? The bootstrap provides a way out of this dilemma. The bootstrap can produce a resampling distribution that can be used to set confidence limits. Resampling, with replacement, 1000 times from the original sample of 20 patients' length of stays gives a resampling distribution like the one shown in the following figure.



Counting in 2.5% of the values from either end of the resampling distribution gives a 95% bootstrap confidence interval for the population mean. In our example, this turns out to be 2.4 days to 3.8 days.

The idea of resampling can be extended to other statistics. For the length of stay for DRG 183, the skewed nature of the population distribution might lead us to consider the median as a more robust measure of center. Again we can resample, with

replacement, from the original sample of 20 values and create a resampling distribution like the one in the following figure.



Again, counting in 2.5% of the values from either end of the resampling distribution gives a 95% bootstrap confidence interval for the population median. This turns out to be from 2 days to 4 days.

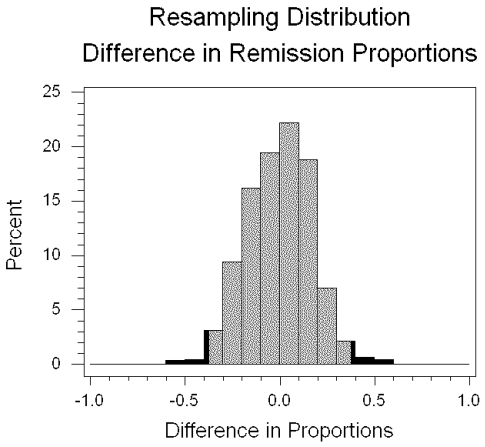
3.2 Testing Hypotheses

Testing statistical hypotheses is an important, but often difficult, topic in any introductory statistics course. For health care professionals, determining whether two treatment regimens are statistically different is very important. Here the use of the bootstrap and resampling can be a great help.

Consider the data presented in the article by Regan, Hellmann and Stone (2001). There are two treatment groups, one with 17 patients and the other with 19 patients. Patients are being treated for Wegener's granulomatosis. The data consists of the number of patients in remission, and not in remission, for each treatment. The data are reproduced in the table below.

	Trmt 1	Trmt 2	
Yes	6	14	20
No	11	5	16
	17	19	36

The proportion in remission for Treatment 1 is 0.353 or 35.3% while the proportion in remission for Treatment 2 is 0.737 or 73.7%. This looks like a large difference in sample proportions, but is this difference statistically significant? How likely is it to get a difference in proportions as large, or larger than, the observed difference of 0.384 if there is actually no difference in the population remission proportions? This question is asking for a P-value. How might this P-value be approximated? If the population remission proportions for each treatment are not different, then we would expect to see $\frac{20}{36} = 0.556$, or 55.6% of the patients in each treatment group in remission. For the first group simulate the number of patients in remission out of 17 in Treatment 1 using a probability of remission of 0.556. Do the same for the number of patients out of 19 in Treatment 2. Compute the difference in simulated remission proportions and construct a histogram like the one in the figure below.



The darker shaded bars represent those differences that are as large as, or larger than, the observed difference of 0.384. There are 26 out of 1000 such cases. This gives a bootstrap P-value of 0.026. It is not likely that as extreme a difference in observed remission proportions could have occurred if the two treatments had been the same. There is evidence that the two treatments have different remission proportions.

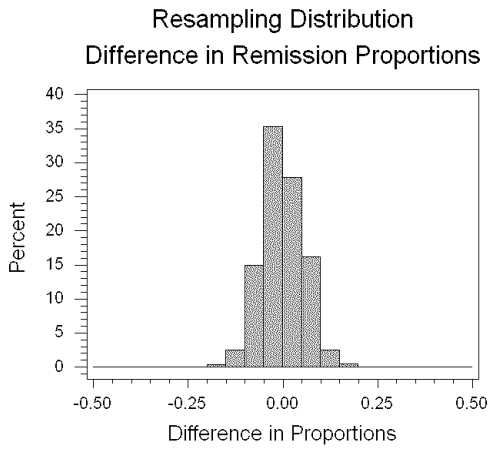
Although we simulate the number of patients in remission in each treatment group, we do this using the observed sample remission proportion of 0.556. Simulating with a value derived from a sample is referred to as a parametric bootstrap.

In the same article by Regan, Hellmann, and Stone (2001) there is another study involving many more patients. The data on remission in each of two treatment groups is given below.

	Trmt 1	Trmt 2	
Yes	84	119	203
No	71	39	110
	155	158	313

The proportion in remission for treatment 1 is 0.542 or 54.2% while the proportion in remission

for treatment 2 is 0.753 or 75.3%. The difference in observed proportions is 0.211. Performing the parametric bootstrap with observed overall remission proportion of 0.649 produces the resampling distribution of the difference in proportions as illustrated in the figure below.



None of the 1000 resampled differences are more extreme than the observed difference of 0.21. The bootstrap P-value is 0.000. This is a good way to show the influence of sample size on the test of hypothesis. Even though the observed difference in proportions is smaller than before (0.211 compared to 0.384), the bootstrap P-values indicate that 0.211 is a more extreme difference.

4 Computing

In order to do simulation and/or bootstrapping effectively, one needs access to computing. There are extremes to the range of computing possibilities. At one end is a package specifically designed for resampling like Resampling Stats by Simon and Bruce. Ricketts and Berry (1994) use Resampling Stats to perform resampling. For reviews of Resampling Stats see Albert and Berliner (1994) and Reeves (1995). Be aware that the reviews are approaching 10 years old and new versions of Resampling Stats have been released. At the other end of the spectrum are computing packages, or languages, that have some of the capabilities needed for simulation and resampling. Willemain (1994) suggests using a spreadsheet program like Excel to perform resampling. Johnson (2001) as well as Taffe and Garnham (1996) use the command language in Minitab. Hesterberg (1998) favors the use of S-plus which has a built in bootstrap function.

The choice one makes may depend on what computing is already available to you. Some may not want to purchase an additional package like Resampling Stats or S-plus just to do resampling. On the other hand, without programming skills in Minitab macros or JMP scripting or Visual Basic for Excel, it may be difficult to make your favorite statistical computing package into a tool for resampling methods.

We have taken two approaches to computing in putting this paper together. The first is to use the built in functions available in R, a free statistical programming language. The second is to develop Minitab Macros. We are making these available to anyone interested via a website. The URL is **http://www.public.iastate.edu/~wrstephe/stateduc.html**.
Available on the website will be the population and sample data for length of stay for Diagnosis Related Groups 183 and 391 and the two tables on remission counts. There will be instructions and a link to download R and instructions on how to use the bootstrap function for the examples in this talk. Also, there will be a short introduction to simulation and macros using Minitab as well as the Minitab Macros for the examples. If there is interest additional illustrations using R and associated Minitab Macros will be posted.

5 References

Albert, J. and Berliner, M. (1994) "Review of 'Resampling Stats (Version 3.14)' ". *The American Statistician*, **48**, 129-131.

Baglivo, J.A. (2001), "Teaching permutation and bootstrap methods", *ASA Proceedings of the Joint Statistical Meetings*.

Boomsma, A. and Molenaar, I.W. (1991), "Resampling with more care (Disc: p30-31+)", *Chance, New Directions for Statistics and Computers*, **4 (1)**, 25-29.

Bruce, P.C. (1992), "Resampling as a complement to 'Against All Odds' ", *ASA Proceedings of the Section on Statistical Education*, 85-93.

Efron, B. (1979), "Bootstrap methods: Another look at the jackknife", *Annals of Statistics*, **7**, 1-26.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM [Society for Industrial and Applied Mathematics] (Philadelphia).

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman & Hall Ltd.

(London; New York)

Hesterberg, T.C. (1998), "Simulation and bootstrapping for teaching statistics", *ASA Proceedings of the Section on Statistical Education*, 44-52.

Johnson, R.W. (2001), "An introduction to the bootstrap", *Teaching Statistics*, **23 (2)**, 49-54.

Moore, D.S., McCabe, G.P., Duckworth, W.M. and Sclove, S.L. (2003), *The Practice of Statistics: Using data for decisions*, W.H. Freeman and Co. (New York)

Quenouille, M. (1949), "Approximate tests of correlation in time series", *Journal of the Royal Statistical Society, Series B*, **11**, 18-84.

Quenouille, M. (1956), "Notes on bias in estimation", *Biometrika*, **43**, 353-360.

Reeves, J. (1995), "Resampling stats", *Teaching Statistics*, **17 (3)**, 101-103.

Regan, M., Hellmann, D. and Stone, J. (2001), "Treatment of Wegener's Granulomatosis", *Rheumatic Diseases Clinics of North America*, **27 (4)**, 863-886.

Ricketts, C. and Berry, J. (1994), "Teaching statistics through resampling", *Teaching Statistics*,**16**, 41-44.

Simon, J.L. (1992), "Resampling and the ability to do statistics", *ASA Proceedings of the Section on Statistical Education*, 78-84.

Simon, J.L. and Bruce, P. (1991), "Resampling: A tool for everyday statistical work", *Chance, New Directions for Statistics and Computers*, **4 (1)**, 22-32.

Stout, W.R., Travers, K.J. and Marden, J. (1999), *Statistics: Making Sense of Data*, 2^{nd} Ed., Mobius Communications Ltd., Rantoul, IL

Taffe, J. and Garnham, N. (1996), "Resampling, the bootstrap and Minitab", *Teaching Statistics*, **18**, 24-25.

Tukey, J.W. (1958), "Bias and confidence in not quite large samples (abstract)", *The Annals of Mathematical Statistics*, **29**, 614.

Willemain, T.R. (1994), "Bootstrap on a shoestring: Resampling using spreadsheets", *The American Statistician*, **48**, 40-42.