A Gentle Introduction to Resampling Techniques

Dale Berger
Claremont Graduate University

Email: dale.berger@cgu.edu

Source: http://wise.cgu.edu Guides and Downloads; Special Topics

160727

A Gentle Introduction to Resampling Techniques

Overview:
Resampling techniques are rapidly entering mainstream data analysis; some statisticians believe that resampling procedures will soon supplant common nonparametric procedures and may displace most parametric procedures as well. This paper introduces the vocabulary, logic, and demonstrates basic applications of permutation and bootstrap resampling methods.

Resampling methods have become practical with the general availability of cheap rapid computing and new software. Compared to standard methods of statistical inference, these modern methods often are simpler and more accurate, require fewer assumptions, and have greater generalizability. Resampling provides especially clear advantages when assumptions of traditional parametric tests are not met, as with small samples from non-normal distributions. Additionally, resampling can address questions that cannot be answered with traditional parametric or nonparametric methods, such as comparisons of medians or ratios. The resampling methods for testing means, medians, ratios, or other parameters are the same, so we do not need new methods for these different applications. Thus, resampling also has advantages of conceptual simplicity.

Parametric tests can be criticized because they require restrictive assumptions, tests may be difficult to interpret, and no tests are readily available for some interesting statistics. More importantly, parametric tests can fail to detect important effects or give misleading results under some conditions. For example, adding a relatively extreme observation can reduce the sensitivity of a parametric test, even if the observation is in the direction of observed effects. An Excel spreadsheet demonstrating the failure of an independent samples *t*-test can be accessed at http://wise.cgu.edu ; Guides and Downloads; Demonstrations Using Excel; Failure of the *t*-test.

Three resampling methods are commonly used for different purposes:

Permutation methods use sampling without replacement to test hypotheses of 'no effect';
Bootstrap methods use sampling with replacement to establish confidence intervals;
Monte Carlo methods use repeated sampling from populations with known characteristics to determine how sensitive statistical procedures are to those characteristics.

Permutation Methods

With permutation methods (also called the randomization technique), we randomly redistribute all of the observed scores into two groups according to our observed N_1 and N_2, and calculate a statistic of interest, such as the difference in means or medians. In our example, we randomly assign the nine observed scores into two groups of four and five. If we do this many times, say 1000 times or 10000 times, we generate a distribution of observed values for the statistic of interest under the null hypothesis of no difference between the two populations. We compare our observed statistic to this empirical sampling distribution to determine how unlikely our observed statistic is if the two population distributions are the same. If the empirical sampling distribution includes 32 samples out of 1000 as extreme or more extreme than our observed sample we conclude that the probability of such an extreme outcome is only about .032, one-tailed. With conventional levels of statistical significance we reject the hypothesis that the two populations are the same.

Randomization allows us to generate the sampling distribution for **any statistic of interest** without making any assumptions about the shape or other parameters of the population distributions. The empirical sampling distribution (or reference distribution) emerges from the multiple randomizations of our observed data. We can determine the percentile location for our observed statistic on the empirical sampling distribution, and determine how unlikely the observed statistic is if the null hypothesis is true. With small samples, we could generate the sampling distribution by calculating the statistic of interest for each possible order. However, when samples are even of modest size, the number of possible orders is so large that it is more practical to use randomization to generate the sampling distribution. With 10 cases per group, there are ${}_{20}C_{10}$ = 184,756 possible randomizations.

Bootstrapping

With bootstrapping, we are able to estimate confidence intervals for a parameter of interest. We assume that our original sample is reasonably representative of the population from which it comes. We randomly sample **with replacement** from the observed scores to produce a new sample of the same size as our original sample. Now we can calculate the statistic of interest (e.g., median) from the new sample. With a large number of new samples, at least 1000, we generate an empirical sampling distribution for the statistic of interest and we can determine upper and lower confidence limits for this statistic. If we have two groups, we can generate a bootstrapped sample from each group separately and calculate the statistic of interest (e.g., difference between medians, a t-test value, or a difference in variance). With multiple replications, we generate a sampling distribution for the statistic of interest. Thus, bootstrapping produces confidence intervals around observed effects.

Monte Carlo

With Monte Carlo techniques, we can specify several populations with known characteristics, and sample randomly from these distributions. With many replications, we generate sampling distributions for the statistics of interest. For example, we may be interested in the sensitivity of the t-test to violations of the assumption of equal variance or normality. We can generate populations that have specific characteristics, and then with multiple resampling we can generate sampling distributions for the statistics of interest. Monte Carlo studies have demonstrated that when two samples are equal in size, the t-test for independent groups is remarkably unaffected by differences in population variance. However, when the samples are small, unequal in size, and the populations have substantially different variance, the t-test is either too liberal or too conservative. The ‘triple whammy’ of small sample size (e.g., $n < 20$), unequal n (e.g., ratio > 4:1), and unequal variance (e.g., ratio > 4:1) generates a test that is too liberal when the sample with the smaller n is from the population with the larger variance. Monte Carlo methods have been used to examine the effects of a wide range of population characteristics on various univariate and multivariate statistics.

Summary

In all three resampling procedures, a statistic of interest is calculated from multiple samples.
Permutation reshuffles the observed cases, sampling **without** replacement.
Bootstrapping selects from the populations of observed cases, sampling **with** replacement.
Monte Carlo typically samples with replacement from theoretical distributions with specific characteristics.

Example 1: Girl Scout Cookie Sales

To test the effect of motivators for Girl Scouts selling cookies, a troop of nine was randomly divided into two groups. A control group of four girls was given the standard instructions, while the remaining five girls were also told that profits from sales would be used for a trip to Disneyland. The data represent boxes of cookies sold in one week. Do we have statistically significant evidence that the Disneyland motivation was effective?

	Control	Treatment
	0	3
	1	6
	2	7
	5	8
		10
Count	4	5
mean	2.000	6.800
SD	2.16	2.59
Pooled variance =		5.83
Pooled SD =		2.41
df =		7
t =		2.964
two-tailed p =		0.021

	Control	Treatment
	0	3
	1	6
	2	7
	5	8
		16
Count	4	5
mean	2.000	8.000
SD	2.16	4.85
Pooled variance =		15.43
Pooled SD =		3.93
df =		7
t =		2.277
two-tailed p =		0.057

In the first set of data a t-test detects a significant difference between the two groups, p = .021.

In the second set of data, we have even stronger evidence of an effect. There is no obvious extreme score (maximum z=2.04). Yet the t-test is less sensitive to the difference between means because of the increased SD and the assumption that the underlying distributions are normal, $t(7) = 2.277$, $p = .057$.

Permutations test: The null hypothesis is that there is no difference between the two populations, such that our observed sample is just one of the many possible ways these nine scores could be distributed between the two groups. To test this hypothesis, we randomly distribute the nine observed scores into two groups of N_1=4 and N_2=5 and compute a statistic of interest, such as *t* or the difference between medians. We record that statistic. With a large number of such random redistributions we can generate an empirical distribution of our statistic under the assumption that there is no difference between the two groups in the population. Then we can determine where our observed statistic falls on this empirical null distribution.

Demonstration of Permutations

	Control	Treatment	Observed		Permutation 1		Permutation 2	
	0	3	0	3	8	5	1	0
	1	6	1	6	7	2	16	8
	2	7	2	7	3	0	6	5
	5	8	5	8	1	16	7	2
		16		16		6		3
Count	4	5	4	5	4	5	4	5
Median	1.5	7	1.5	7	5	5	6.5	3
Mean	2.000	8.000	2	8.000	4.75	5.800	7.5	3.6
SD	2.16	4.85	2.16	4.85	3.3	6.18	6.24	3.05
Pooled variance =		15.43		15.43		26.51		22
Pooled SD =		3.93		3.93		5.15		4.69
df =		7		7		7		7
t =		2.277		2.277		0.304		-1.24
two-tailed p =		0.057		0.057		0.770		0.26

Example from Howell: "Hurry up Fella!"

This example is adapted from Howell's website. It uses heuristic data generated by Howell to represent data collected by Ruback and Juieng (1997). This study was motivated by the perception that when you are waiting for someone's parking space, the driver you are waiting for takes longer than necessary to move out of the space. Our data represent the time it took 40 drivers to begin moving out of their parking space from the time they touched their car. For half of the observations someone was waiting for the space.

No one waiting
36.30 42.07 39.97 39.33 33.76 33.91 39.65 84.92 40.70 39.65
39.48 35.38 75.07 36.46 38.73 33.88 34.39 60.52 53.63 50.62

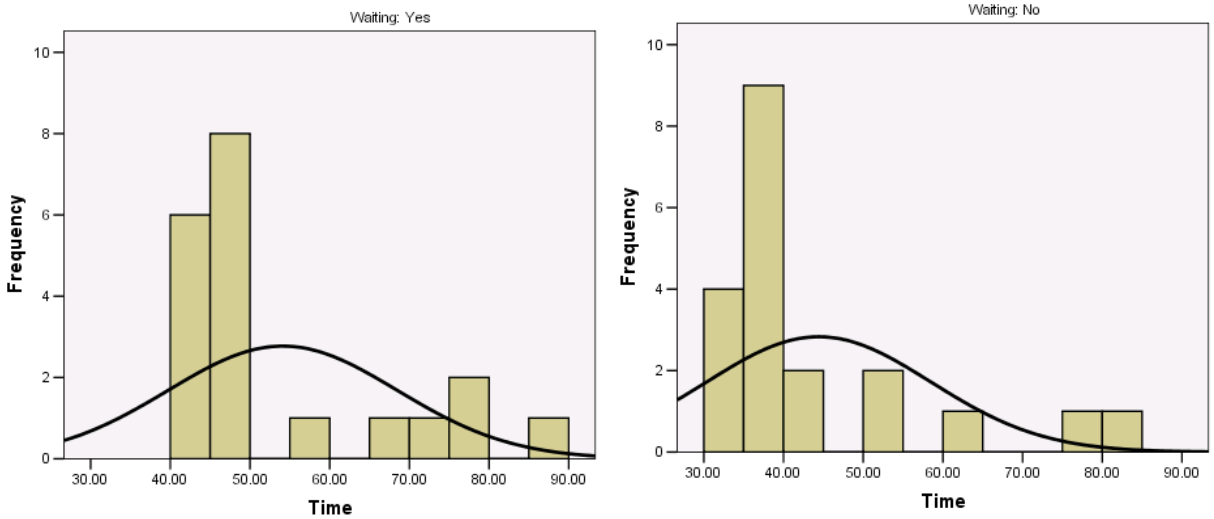
Someone waiting
49.48 43.30 85.97 46.92 49.18 79.30 47.35 46.52 59.68 42.89
49.29 68.69 41.61 46.81 43.75 46.55 42.33 71.48 78.95 42.06

Parametric *t*-test:
How do we test whether people take longer to vacate a parking space when someone is waiting? The standard *t*-test for independent samples shows $t(38) = \mathbf{-2.150}$, p=.038 (two-tailed).

Group Statistics

	Waiting	N	Mean	Std. Deviation	Std. Error Mean
Time	1 No	20	44.4210	14.09798	3.15240
	2 Yes	20	54.1055	14.39418	3.21864

Is this an appropriate test? Have we satisfied the assumptions of the *t*-test? Let's check the distributions.



These distributions should make us skeptical about the accuracy of the *p* value from the *t*-test because that *p* value is computed from a theoretical normal sampling distribution. With relatively small samples that are so skewed, the assumption of normality is questionable.

Nonparametric Mann-Whitney-Wilcoxon test: The null hypothesis for the MWW test is that the population distributions of scores are identical. If we randomly choose one score from population A and one score from population B, $p(A > B) = ½$. If the null hypothesis is true, when we rank the pooled scores from the two populations, the expected value of the average rank is the same for scores from population A or population B.

Test Statistics[b]

	Time
Mann-Whitney U	81.000
Wilcoxon W	291.000
Z	-3.219
Asymp. Sig. (2-tailed)	.001
Exact Sig. [2*(1-tailed Sig.)]	.001[a]

a. Not corrected for ties.

b. Grouping Variable: Waiting

Ranks

	Waiting	N	Mean Rank	Sum of Ranks
Time	1 No	20	14.55	291.00
	2 Yes	20	26.45	529.00
	Total	40		

In our data, the average rank for scores from the 'No one waiting' group is 14.55 while the average rank from the 'Someone waiting' group is 26.45. The two-tailed p value is 0.000933417. (I got this from SPSS by double-clicking the tabled p value in the output window.)

A limitation of the MWW test is that it may not provide a test of a statistic of interest. For example, we may be more interested in comparing the median time or a standardized difference between means. As noted earlier, the MWW test is a test of the equality of medians only if the two populations are assumed to have the same shape and dispersion. It is a test of means only if the distributions are also symmetrical. Those conditions are not satisfied here.

Permutation tests (also called Randomization tests):
The permutation method of resampling provides an alternative approach that does not require any assumptions about the shapes of the distributions, but it can provide a test using a measure of your choice.

Suppose that there is absolutely no effect of someone waiting. Then any one of these scores is equally likely to be observed in either group. Any random shuffling of these 40 scores is equally likely. Theoretically, we could generate all possible combinations and determine whether the observed combination is an extreme outcome.

But how should we measure the difference between groups? The *t*-test is a measure of the standardized difference between group means. The nonparametric MWW test uses average rank the scores in each group. In practice, we may be more interested in a difference in medians, ranges, variances, or something else.

***t* as a measure of an effect:**
Suppose that we really are interested in the standardized difference between means, but we are reluctant to use the parametric *t*-test because it assumes that the sampling distribution for *t* values is based on an underlying normal distribution. With resampling we can use a calculated *t*-value as a measure of the group difference, but we can test it against an empirical sampling distribution for the *t*-value. In our example, the *t*-value for our data is -2.150. We can randomly reshuffle the data into two groups of N=20 each and recompute the *t*-value. If we do this for many reshuffles

of the data (e.g., 10,000) we can generate an empirical distribution of the *t*-value. This distribution is NOT necessarily distributed according to the parametric t distribution. However, we can determine how extreme our observed value of -2.150 is in this distribution. If only 214 of the 10,000 shuffles produce a *t*-value as small as -2.150, we can conclude that the probability is only about .0214 of observing a *t*-value as small as or smaller than our observed value if the null hypothesis is true.

Demonstration of Resampling Procedures using Howell's program

David Howell provides a free program that does resampling for some selected statistics. The program, instructions for using the program, references on resampling, and discussions of resampling can be accessed through his web page https://www.uvm.edu/~dhowell/StatPages/ . Howell provides a treasure trove of thoughtful explanations of statistical concepts.

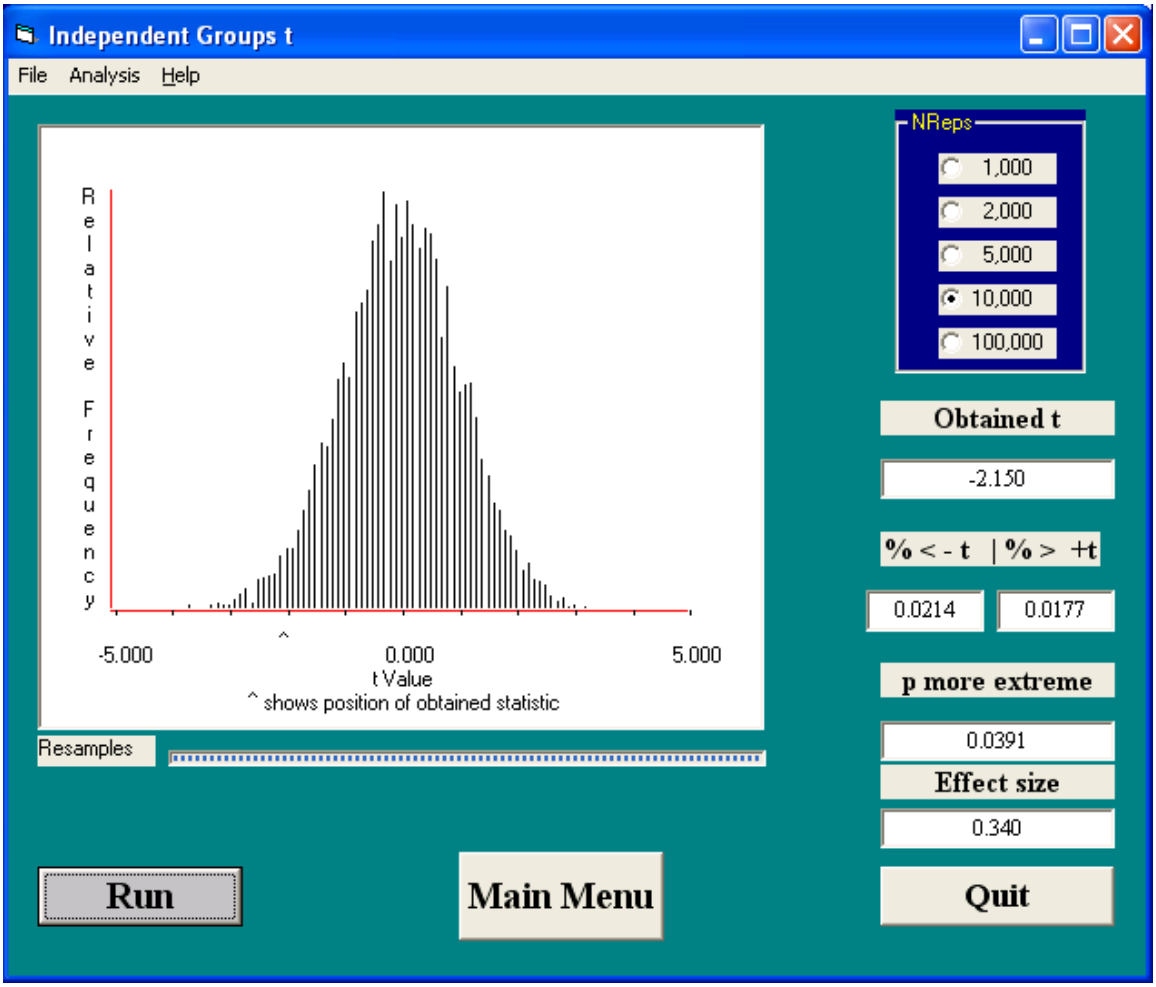
To replicate the analyses shown here, you need to install Howell's program on your computer (follow his instructions for installation).

Howell's Resampling Procedures program includes the following subprograms:

Bootstrapping (Generation of confidence intervals)
- Mean of a single sample
- Correlation
- Median of a single sample
- Difference between two medians
- Mediation (not yet operational)
- Oneway ANOVA

Randomization Tests (permutations of an index of an effect assuming no effect)
- Two independent samples (independent t as the index of an effect)
- Two paired samples (dependent t as the index of an effect)
- Compare medians of two samples (difference between medians)
- Correlation (permutation program to test Ho: rho=0)
- Oneway ANOVA via randomization
- Paired correlations (not yet operational)
- Repeated measures one-way

Howell's Randomization Tests (using permutations)
Howell's randomization program *Independent Groups t* produces a graph of the empirical sampling distribution of a computed *t*-value under the assumption that there is no difference between the two populations. The output window shows the obtained *t*-value of **-2.150** along with the empirical probability of observing a *t*-value more extreme than the observed value. In this sample of 10,000 random permutations only 214 samples produced a *t*-value more negative than -2.150 and 177 samples larger than +2.150. Thus, the two-tailed probability of a *t*-value farther from zero than 2.150 is estimated to be .0214+.0177 = .0391.



The observed effect size in the data can be calculated as

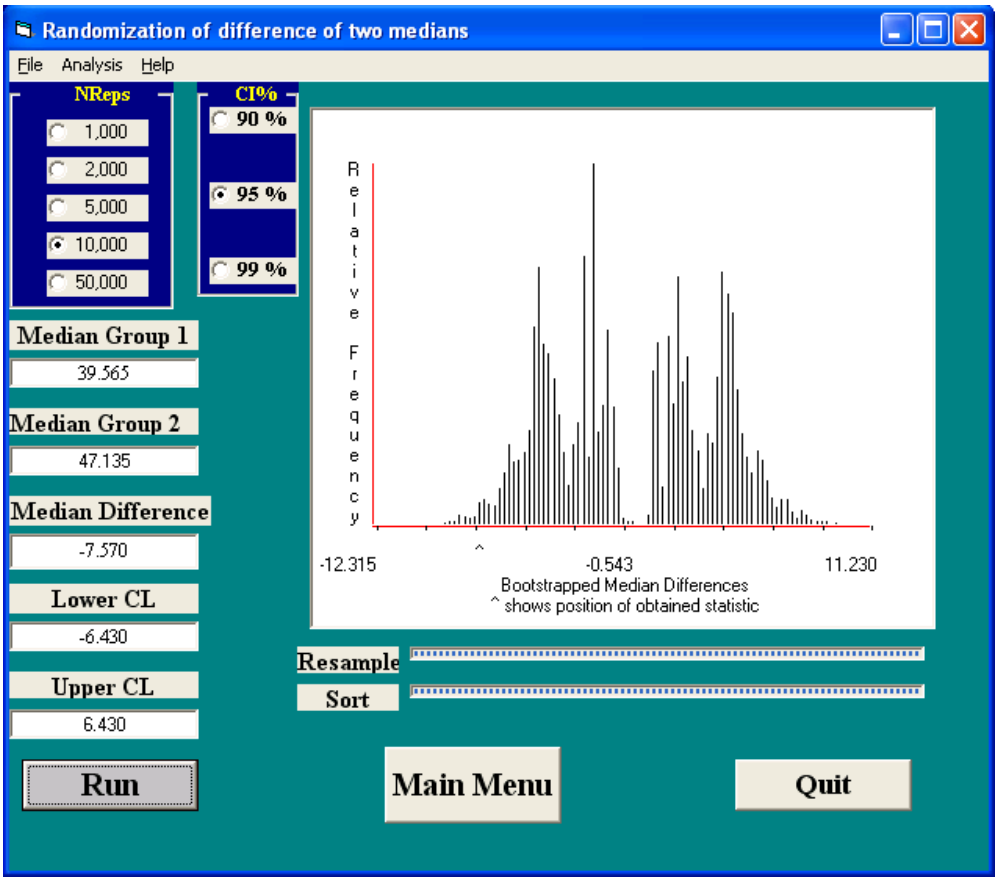
$$d = \frac{(M_1 - M_2)}{SD_{pooled}} = \frac{(44.4210 - 54.1055)}{14.246} = .680$$

(Howell's reported effect size of .340 appears to be an error. Maybe it is based on the sum of SD rather than pooled SD. Unconventional at least.)

A crucial point to emphasize is that we are using the *t*-value as a measure of the difference between the groups, not as a statistic to be compared to the parametric *t* distribution. Instead, we compared our observed value to an empirical sampling distribution that we generated for this statistic. Our estimated two-tailed *p*-value is .0391, which is virtually the same as the *p*-value of .038 computed for the parametric *t*-test. (I re-ran this with NReps = 100,000 and got .03697.)

Difference between medians as a measure of an effect: Because of skew in the data and the possibility that we may observe an exceptionally long wait time on occasion (e.g., someone makes a telephone call before driving), we may be more interested in the difference between medians than the difference between means. Howell provides a resampling module for this purpose.

In our sample, the median wait time is 39.565 seconds for the first group and 47.135 for the second group. The difference between medians is -7.570 seconds. Is this an extreme outcome if the scores are randomly scrambled between groups?



Howell's program incorrectly calls this a bootstrap analysis; the label at the top is correct – this is a randomization analysis

The output from Howell's module with NReps=10,000 shows the empirical sampling distribution for the difference between means under the null hypothesis that the scores are scrambled randomly between the two groups. The Lower CL = -6.430 is the lower limit for a 95% confidence interval. Because our observed value of -7.570 is less than this value, we conclude that our observed value is not likely if the null hypothesis is true.

Note that this test makes no assumptions at all about the shape of the population distributions. The empirical distribution shows the distribution of the difference between medians when the null hypothesis is true. Using alpha=.05 two-tailed, we can conclude that the typical person (median person) does take longer to exit a parking space when someone is waiting for it.

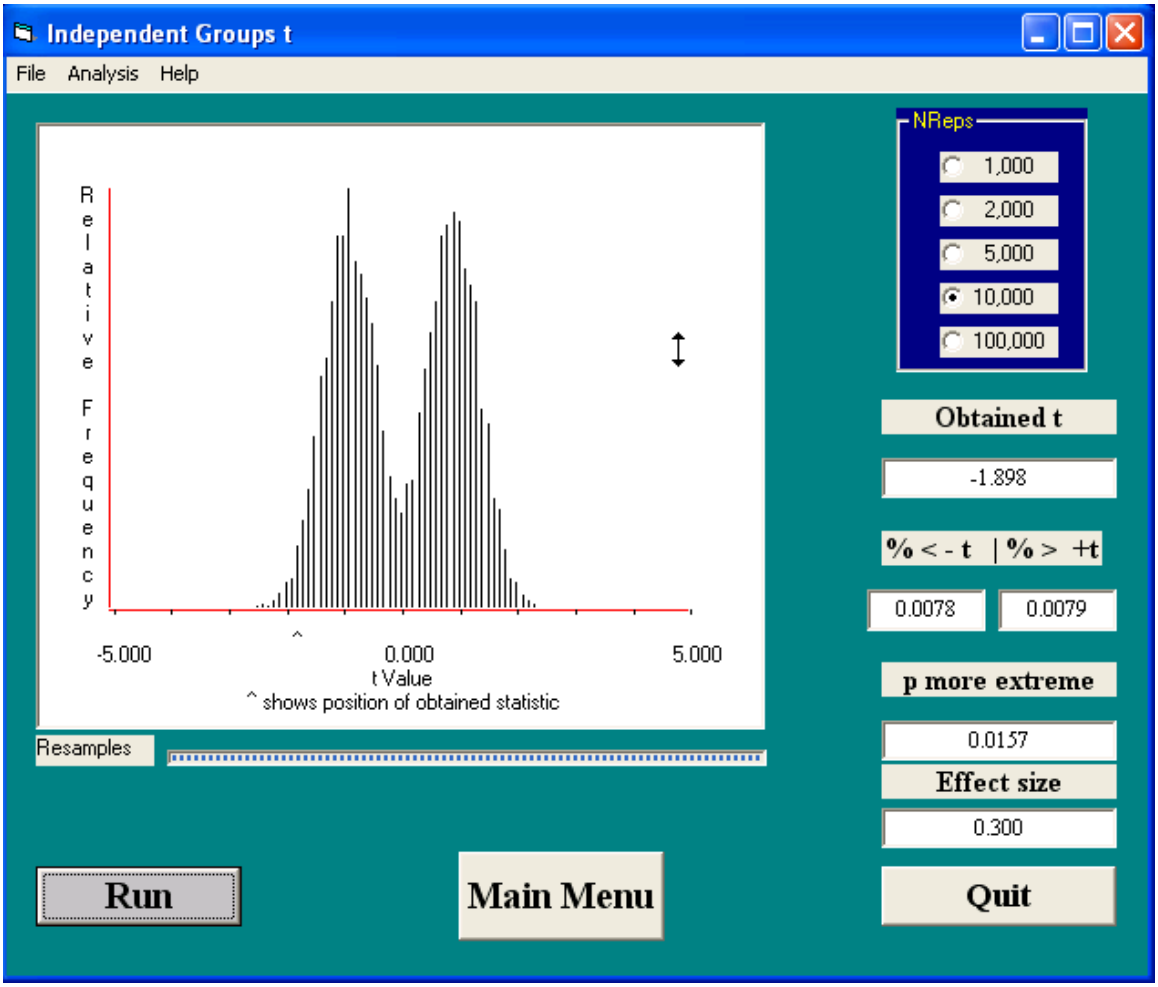
Impact of outliers

Suppose one person in the Waiting group decided to place a telephone call before leaving the parking space. To simulate this observation, we add 200 seconds to the last observation in the Waiting group, replacing 42.06 with 242.06.

Parametric *t*: Outliers can wreak havoc with the traditional t test. They often inflate a mean and the associated group variance. The result often is less sensitivity for *t*, and concern that violations of assumptions make the test inaccurate anyway.

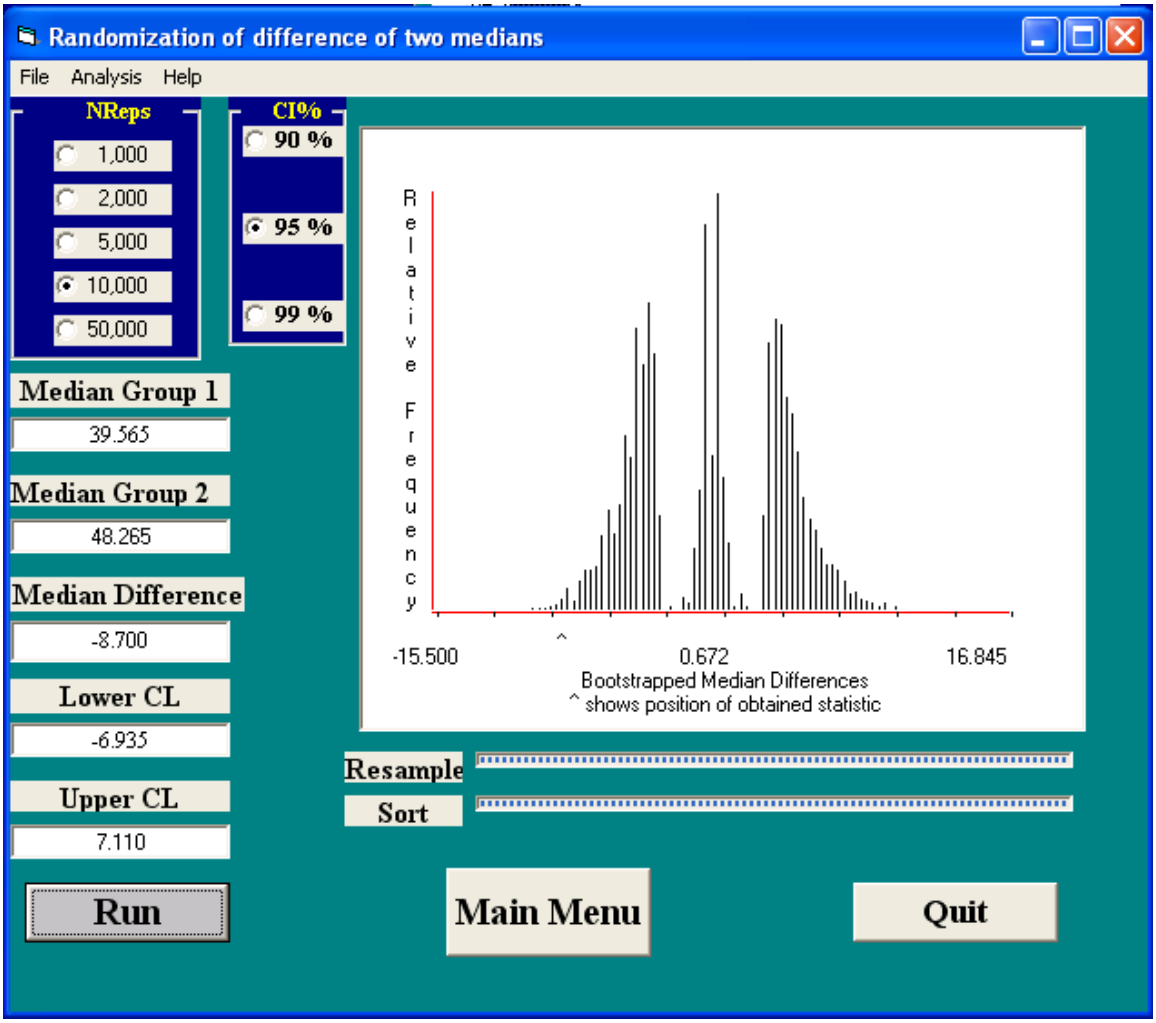
With this added delay time in the Waiting group, one might expect that the parametric *t*-test show an even smaller p value. The actual result not assuming equal variance is $t(22.8) = -1.898$, $p=.070$, two-tailed. Why is the *t*-test less sensitive when an extreme case is added in the direction of the observed effect? The within groups variance estimate is greatly increased. If we had a normal distribution with such a large variance, a sample mean with N=20 would be quite unstable. We can't trust this *p* value anyway because of the severe violation of the assumption of normality.

Permutation test using *t*: With resampling, the outlier is equally likely to fall into either group, resulting in an observed *t* that will jump between positive and negative values. The observed *t* of -1.898 certainly does not come from a nice normal sampling distribution.



With 10,000 random shuffles of the data, we find that our observed $t = -1.898$ is quite extreme. Only 157 out of 10,000 shuffles produced a *t* value farther from zero. Thus, we conclude that we have evidence of a difference between the groups ($p = .0157$, two-tailed). Note that in contrast to the parametric *t* test, when we use the resampling application to find the empirical sampling distribution of the *t*-value the evidence for a group difference is stronger in the presence of the outlier than when there was no outlier.

Permutation test using the difference between medians: Because the extreme score moved an observation from below the median to above the median for the Wait group, the observed median for the Wait group is 48.265 instead of 47.135, and the difference between medians is -8.700. The empirical sampling distribution for the difference between medians has some gaps because of gaps in the observed data – some median values are impossible.



The empirical lower limit for the 95% confidence interval is -6.935. Because our observed value of -8.700 is below this limit, we can conclude that our observed outcome is unlikely if there was no difference between the two groups.

Cookie Data: Comparison of parametric and nonparametric tests

Let's explore our Girl Scouts data from Example 1. Recall that our control group of four scouts sold 0, 1, 2, and 5 boxes of cookies while our Disneyland-inspired group of five scouts sold 3, 6, 7, 8, and 16 boxes of cookies.

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Sales	1 Control	4	2.00	2.160	1.080
	2 Treatment	5	8.00	4.848	2.168

We noted earlier that the traditional **independent *t*-test** was more sensitive to a difference when a treatment-group girl sold 10 boxes of cookies rather than 16, so we are skeptical about the value of a *t*-test. If we computed *t* anyway, we would find $t(7) = -2.277$, p=.057, two-tailed. The *t*-test provided by SPSS not assuming equal variances gives $t(5.76) = -2.48$, p=.050, two-tailed.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Sales	Equal variances assumed	.940	.364	-2.277	7	.057	-6.000	2.635	-12.231	.231
	Equal variances not assumed			-2.477	5.759	.050	-6.000	2.422	-11.987	-.013

Alternatively, we could use a **nonparametric** approach to test the null hypothesis that the distributions of scores for the two populations (C and T) are identical. How many distinct ways are there to order 4 Cs and 5 Ts? If we had 9 distinct objects, we could order them in 9! = 9x8x7x6x5x4x3x2x1 = 362,880 ways. However, if four of the objects are indistinguishable, then for each placement of these four objects in the sequence of nine, there are 4! = 4x3x2x1 = 24 ways these four objects could be reordered that would look the same. Similarly, each placement of five indistinguishable objects represents 5! = 5x4x3x2x1 = 60 orders for these five objects that look the same. Thus, the total number of distinguishable arrangements of 4 Cs and 5Ts is 9!/(4!5!) = 126. This is a combinations problem, ${}_9C_4 = 9!/(4!5!) = 126$.

If the population distributions of C and T are identical, then the probability of any specific order is 1/126 = .00794. When we order our observed scores from low to high, we observe the sequence CCCTCTTTT. The probability of observing this exact sequence is 1/126 = .00794. There is only one arrangement that would show a greater advantage for the treatment group, CCCCTTTTT. Thus, the probability of observing an arrangement of 4 Cs and 5 Ts that shows an advantage for T as great or greater than we observed is 2/126 = .01587. Using the conventional

.05 level of statistical significance we would interpret our observed data as statistically significant evidence of an advantage for the treatment group. The *p* value for a two-tailed test would be double the *p* value for a one-tailed test, or two-tailed *p*=.03175.

Score	Group	Rank 1	R1C	R1T	Rank 2	R2C	R2T
0	C	1	1		9	9	
1	C	2	2		8	8	
2	C	3	3		7	7	
3	T	4		4	6		6
5	C	5	5		5	5	
6	T	6		6	4		4
7	T	7		7	3		3
8	T	8		8	2		2
16	T	9		9	1		1
	Sum	**45**	**11**	**34**	**45**	**29**	**16**

This test result is exactly what we would observe with a Wilcoxon W test which is based on ranks of the observed scores. W is the sum of ranks for the smaller group, using the smaller sum of ranks when scores are ranked from larger to smaller or smaller to larger. In our example, C is the smaller group (N_1=4) and the sum of ranks is 1+2+3+5 = 11. Mann-Whitney U is the sum of the number of Cs greater than each T or vice versa, whichever is smaller. In our observed sequence, only one C is greater than one T, so U=1.

Ranks

	Group	N	Mean Rank	Sum of Ranks
Sales	1 Control	4	2.75	11.00
	2 Treatment	5	6.80	34.00
	Total	9		

Test Statistics[b]

	Sales
Mann-Whitney U	1.000
Wilcoxon W	11.000
Z	-2.205
Asymp. Sig. (2-tailed)	.027
Exact Sig. [2*(1-tailed Sig.)]	.032[a]

a. Not corrected for ties.

b. Grouping Variable: Group

The Mann-Whitney U and Wilcoxon W are based on different statistics, but they give identical p values. In the SPSS table of test statistics, Exact Sig. [2*(1-tailed Sig.)] = .032 matches our calculation of *p*=.03175. For small samples, tables are available for W and U (e.g., Siegel & Castellan, 1988). For large sample (e.g., N_1>25), the sampling distribution of W approaches normal and the asymptotic Z test provides an approximation. In our example, N_1=4 is much less than 25, and the Z approximation is not very accurate (*p*=.027 two-tailed instead of *p*=.032).

The null hypothesis for the Mann-Whitney-Wilcoxon test is that the two samples are taken from identical distributions. If the two populations are assumed to have the same shape and dispersion, then the test is a test of equality of the medians. If the distributions are also symmetrical, the test is a test of means.

Entering our own data into Howell's program

Create your data file in Word. To prepare data for analysis, we need to create a text file (also called an ASCII file or a .DOS file). You can create the file in Word, but you must save it using the "Plain text" format and you must give it a name ending in .DAT. The data can be entered on one or more lines, separated by spaces or new lines. In our example the information could be on two lines, as follows:

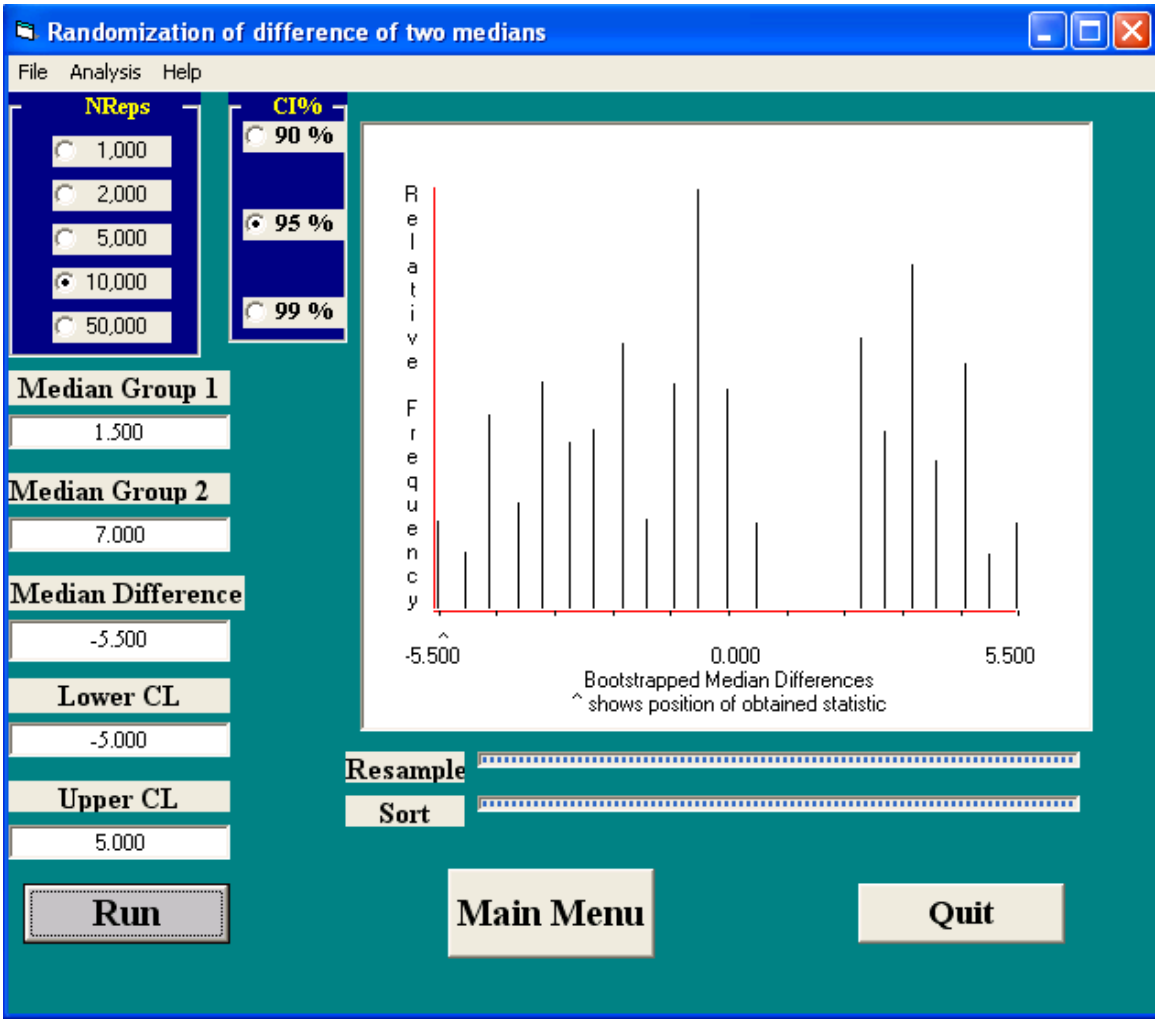
4 5
0 1 2 5 3 6 7 8 16

The data should begin at the top of the page and there should be no leading spaces. In Word, click File, Save As…, click Save as type…, select Plain text, and name the file with an extension of .DAT (e.g., **Cookies.DAT**). Word wants to give the file the extension .TXT, but you must insist on using .DAT. A File Conversion window opens, click OK. The Windows default is OK here.

Test difference between medians with randomization procedure: Is the observed difference between the medians unusual for this set of nine observations split into two groups of 4 and 5 cases? The observed median of the first group is 1.5 and the median of the second group is 7.0, so the observed difference between medians is 5.500 (or -5.500, depending which way we compare groups). The randomization procedure randomly assigns the nine observations into two groups of four and five cases, and calculates the difference in the medians. Repeating this 10,000 times gives a probability distribution of the possible outcomes. We can compare our observed value of 5.500 to this empirical sampling distribution to determine how unlikely our observed result is for a random permutation of this set of nine observations.

In the Resampling Procedures Main Menu, select Analysis, select Randomization Tests, select Compare Medians of 2 Samples. Select File, Open, and locate your data file, select it, and click Open. You can select the number of replications. Select 10,000. Now click Run. The output is on the next page (Randomization of difference of two medians).

The program shows the median for Group 1 (1.500), the median for Group 2 (7.000), and the Median Difference (-5.500). Based on the empirical sampling distribution, the 95% Lower CL is -5.000 and the Upper CL is 5.000. Thus, our observed value of -5.500 is below the lower limit of the confidence interval.



The null hypothesis that we are testing is that the two groups have the same distribution in the population. If the null hypothesis is true, then our nine scores came from the equivalent distributions. If we randomly divide the nine observed scores into a group of four and a group of five, we would expect the medians for those two groups to be approximately equal. The 95% confidence limits of -5.0 to +5.0 indicate that if the null hypothesis is true we would observe differences in medians in this range 95% of the time when we subtract the median for the smaller sample from the median of the larger sample.

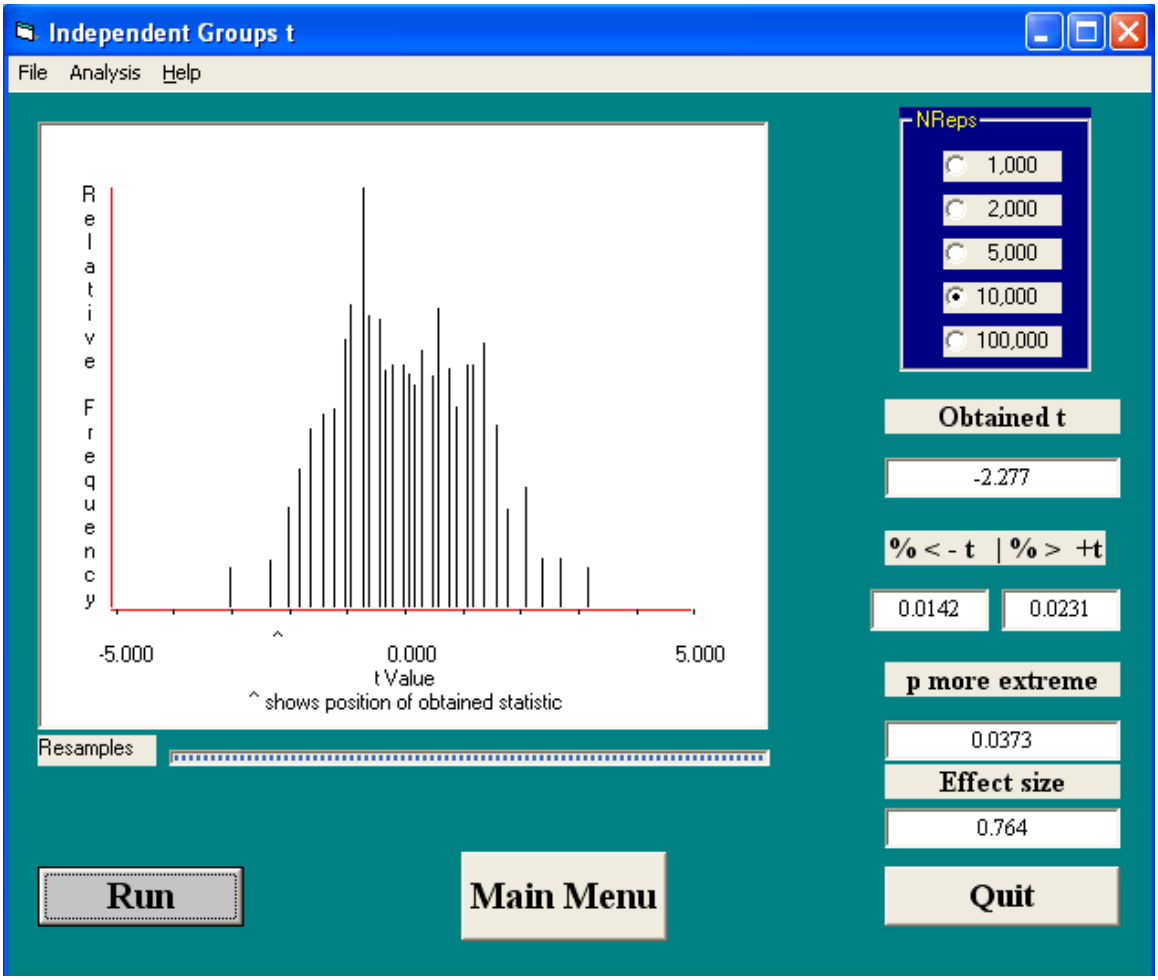
Because our observed difference in medians is below the lower limit of the confidence interval, we reject the null hypothesis that the populations represented by our samples have the same distribution.

Suggested APA format: "A permutation test with 10,000 replications was computed using Howell's re-sampling software (Howell, 2002). The observed difference in medians of 5.5 fell outside of the 95% confidence interval based on the null hypothesis of no difference, 95% CI[-5.0, 5.0], $p < .05$."

Test for group difference using *t*-value as a measure of group difference:

When we computed a *t*-test to compare the two sample means, we found $t(7) = 2.277$, $p=.057$. However, it may not be appropriate to compare our calculated *t* to the parametric *t* distribution because we do not have normal distributions. Instead, we can compare our observed *t*-value to the empirical distribution of *t*-values under the assumption of no difference between the two populations.

In the Resampling Procedures Main Menu, select Analysis, select Randomization Tests, select Two Independent Samples. Select File, Open, and locate your data file, select it, and click Open. You can select the number of replications. Select 10,000. Now click Run. The output is below.



Here we see that our observed *t*-value of -2.277 is quite unlikely if there really is no difference between the two groups. Only 142 samples produced a *t*-value smaller than -2.277 and 231 produced a *t*-value greater than +2.277. Thus, we conclude that there is a statistically significant difference between the two groups, $p=.0373$.

It is important to note that we did not need to assume anything about the shapes of the population distributions. We are using the *t*-value as a measure of group difference, not as a statistic to be compared to a tabled parametric distribution.

Bootstrapping Demonstrations

With bootstrapping, our goal is to estimate the confidence interval for a statistic of interest, such as a calculated *t*-value or the difference between the medians for two groups.

We use our observed sample distribution as our best representation of the population from which we are sampling. Thus, bootstrapping works better when we have larger samples, especially if there are outliers in the population. If we have a sample with N_1 cases, we draw a new sample of N_1 cases with replacement from our observed sample.

When we have samples from two groups, as with the Girl Scout Cookie Study, we draw separate bootstrapped samples from each group, as illustrated in the Excel example below.

Demonstration of Bootstrapping

	Control	Treatment	Observed		Bootstrap 1		Bootstrap 2	
	0	3	0	3	2	7	5	6
	1	6	1	6	2	16	0	3
	2	7	2	7	5	16	5	8
	5	8	5	8	1	6	1	3
		16		16		16		8
Count	4	5	4	5	4	5	4	5
Median	1.5	7	1.5	7	2	16	3	6
Mean	2.000	8.000	2	8	2.5	12.2	2.75	5.6
SD	2.16	4.85	2.16	4.85	1.73	5.22	2.63	2.51
Pooled variance =		15.43		15.4		16.8		6.56
Pooled SD =		3.93		3.93		4.1		2.56
df =		7		7		7		7
t =		2.277		2.28		3.52		1.66
two-tailed p =		0.057		0.06		0.01		0.14

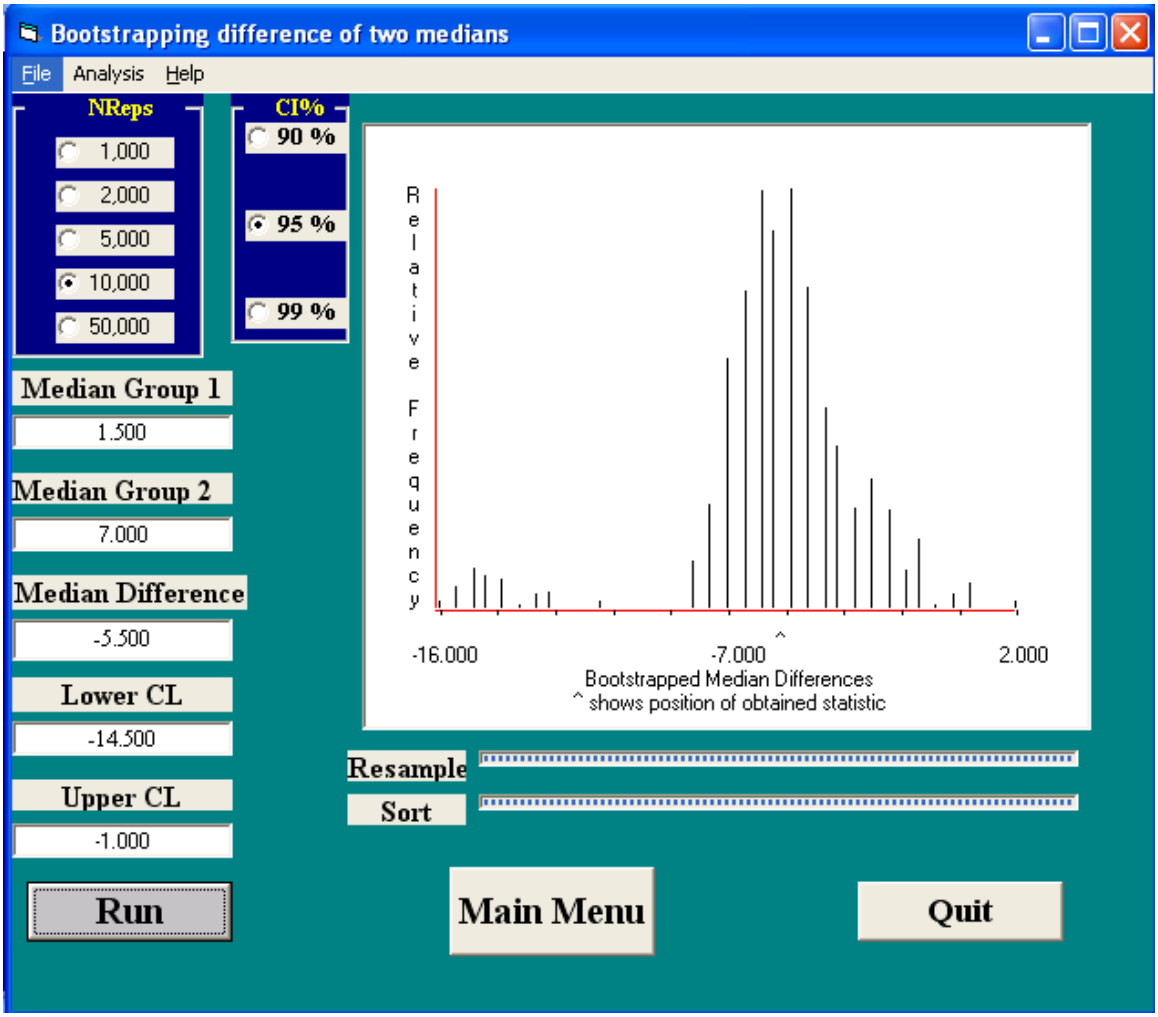
Our samples in this example are much too small for serious bootstrapping analyses, but they work well for illustration. Our observed samples (N_1=4 and N_2=5) generated a *t*-value of 2.277 and a difference between medians of 1.5 – 7.0 = -5.5.

In any given bootstrap draw, each observed value in a sample is equally likely to be drawn. Bootstrap 1 shows a possible result if we randomly sample with replacement from each of our observed samples. Here we happened to draw the largest score in the Treatment Group three times out of five. In Bootstrap 2 we didn't draw that value at all.

For each bootstrap sample we record the statistic of interest, such as the difference between medians (here we have 2 -16 = -14 and 3-6 = -3). The observed t-values were 3.52 and 1.66. We can generate a sampling distribution for our statistic representing possible values if cases were drawn from a population that looks like our observed samples.

The most extreme medians that would be possible from populations that looked like our observed sample would be 0 or 5 for the Control group and 3 or 16 for the Treatment group. Thus the bootstrapped samples could find values as extreme as 0-16 = -16 on one end or 5-3=2 on the other. The bootstrapped distribution will be centered on the observed sample, however, which showed a difference in medians of 1.5-7.0 = 5.5.

Below is a bootstrapped sampling distribution based on 10,000 resamples. The 95% confidence interval extends from -14.5 to -1.0. Thus, a median difference of zero is unlikely given populations that look like our samples. Certain median differences such as -9 or -10 are impossible given our data. A difference of -11 is possible but not -12 or -13.



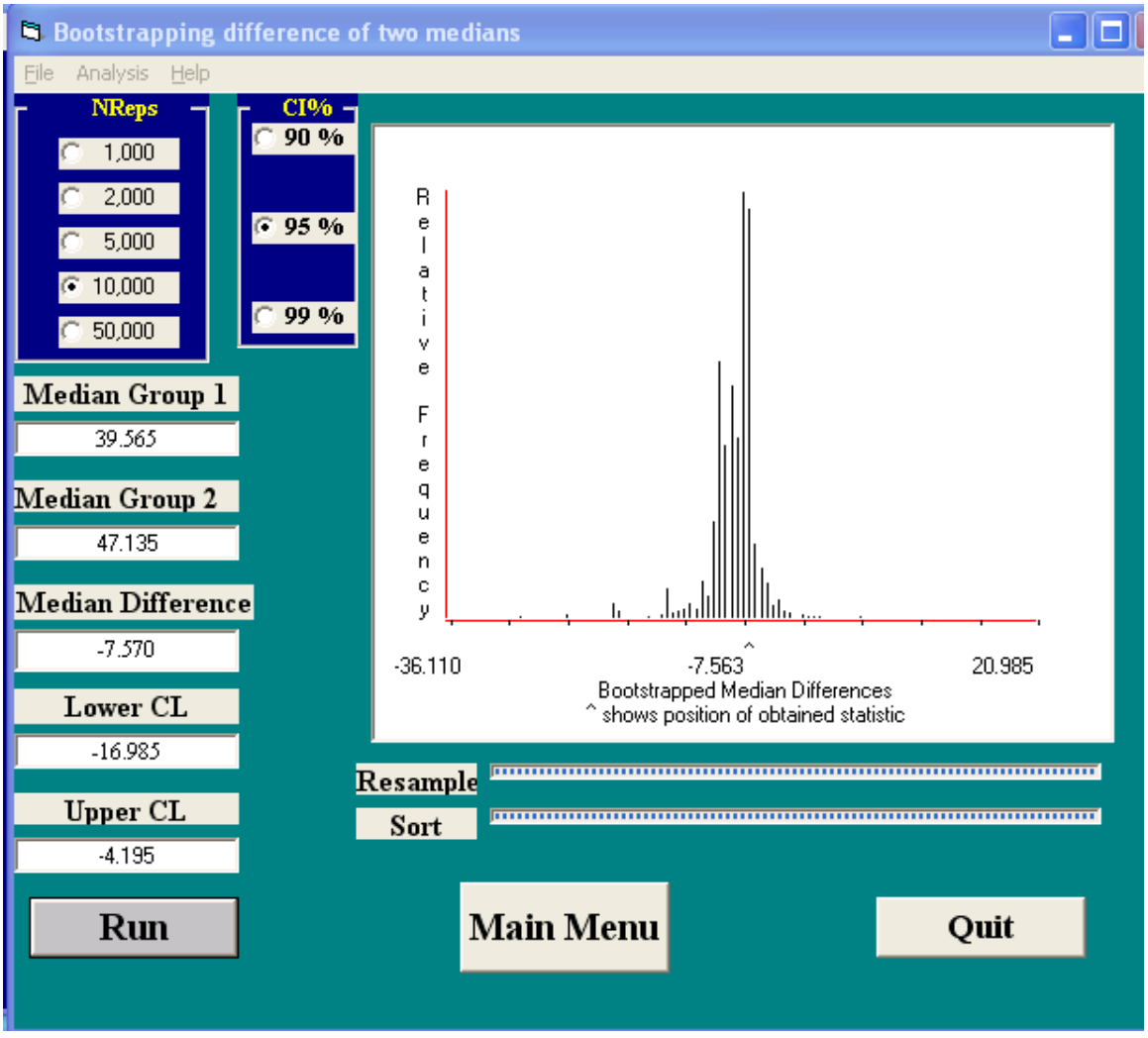
Suggested format for presenting bootstrapped results, APA style:
"The median number of cookies sold by the experimental group was 7.0 compared to only 1.5 for the control group. A bootstrapped 95% confidence interval using Howell's resampling software (Howell, 2002). Those in the experimental group sold significantly more cookies than those in the control group, $p < .05$. The median difference = 5.5, 95% CI with 10,000 replications [1.0, 14.5] does not include zero." Note – We could re-enter the data with the order of groups reversed so that the differences in medians would all be positive. I simply reversed the signs for all of the differences between medians to accomplish the same goal.

Bootstrapping the difference between medians for the Parking Lot data.

The Parking Lot data from Howell is large and stable enough that bootstrapping can be applied with confidence. In this example we use the difference between the medians as our statistic of interest. We sample WITH REPLACEMENT from each observed set of scores to create new samples the same sizes as the observed samples. We generate an empirical sampling distribution for the statistic (difference between medians) by repeating this process many times.

From Howell's main menu, click Analysis, select Bootstrapping Procedures, select Compare 2 Medians via Bootstrapping. Select File, Open, and locate your data file, select it, and click Open. Under the number of replications, select 10,000. Now click Run.

Here is an example using 10,000 replications.



The sampling distribution is centered on the observed value of -7.570 and the upper and lower limits indicate the likely range for the statistic. In our example, the value of zero falls well outside of the range. The sampling distribution for medians is often lumpy and asymmetrical.

Resampling with correlations

The difference between randomization and bootstrapping is especially clear with correlations. When we use randomization, the null hypothesis is that the X,Y pairs are randomly scrambled. Thus, the multiple sample correlations are centered on zero, and we are interested in how unusual our observed correlation is on that sampling distribution of correlations.

With bootstrapping, we keep each X,Y pair together, but we randomly sample these pairs with replacement. Thus, the sampling distribution of correlations is centered on the observed correlation, and the empirical sampling distribution may be quite skewed. Our interest is in setting confidence intervals for the population correlation.

As an example, we will use data from Howell representing scores on an SAT-type test where students were asked to answer questions about a passage that they didn't read. Performance on this test reflects test-taking skills. How are these skills related to SAT test performance?

Score	58	48	48	41	34	43	38	53	41	60	55	44	43	49
SAT	590	590	580	490	550	580	550	580	550	700	800	600	650	580
Score	47	33	47	40	46	53	40	45	39	47	50	53	46	53
SAT	660	590	600	540	610	580	620	600	560	560	570	630	510	620

Below are examples of a Permutation (randomization) and a Bootstrap for the first ten pairs of scores. With the Permutation test, Score is unchanged but the ten SAT scores are in scrambled order. The ID for Permutation refers to the ID score for the SAT value only from the Observed. With the Bootstrap, pairs are kept together but are resampled with replacement. The ID for Bootstrap refers to the ID for the pair in the Observed sample. Note that Pair 10 was drawn first, then Pair 4, etc. Pair 5 was drawn twice in a row but Pair 3 was never drawn.

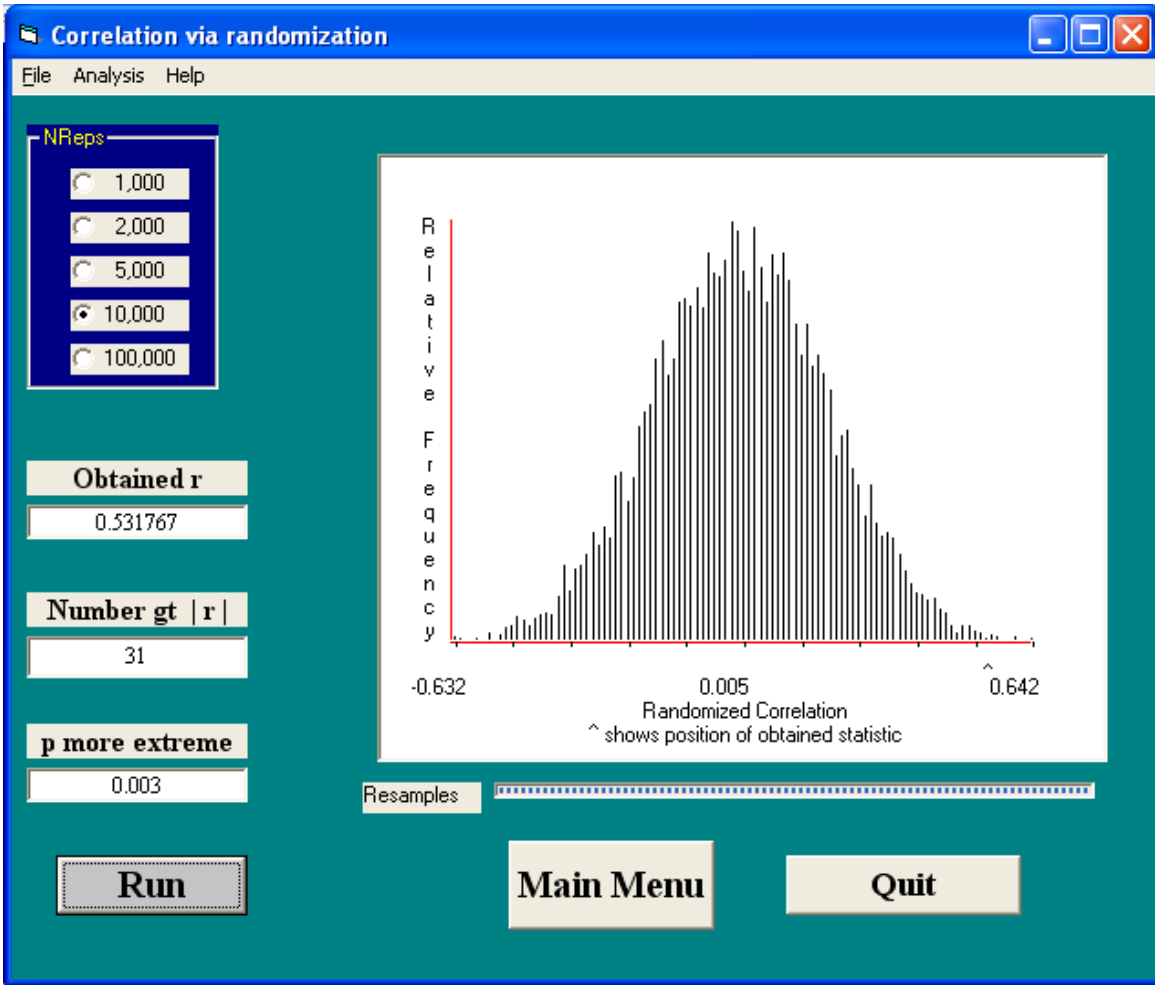
If we use this sample of ten cases, the observed r=.743 will be the middle of the Bootstrapped distribution, but the middle of the Permutation distribution will be zero.

	Permutation				Observed				Bootstrap		
	Score	SAT	ID		ID	Score	SAT		ID	Score	SAT
	58	580	6		1	58	590		10	60	700
	48	550	7		2	48	590		4	41	490
	48	590	1		3	48	580		1	58	590
	41	580	3		4	41	490		5	34	550
	34	550	9		5	34	550		5	34	550
	43	490	4		6	43	580		2	48	590
	38	550	5		7	38	550		9	41	550
	53	590	2		8	53	580		5	34	550
	41	700	10		9	41	550		8	53	580
	60	580	8		10	60	700		9	41	550
Mean	46.4	576				46.4	576			44.4	570
r	0.077					0.743				0.729	

Randomization (Permutation) Test for Correlation

The data are in a text-only file with the extension of .DAT. The first row has the number of cases, 28. Each row thereafter has a pair of observations separated by a space.

With resampling, the X values are fixed and the Y values are randomly redistributed to be paired with the X values on each resample.



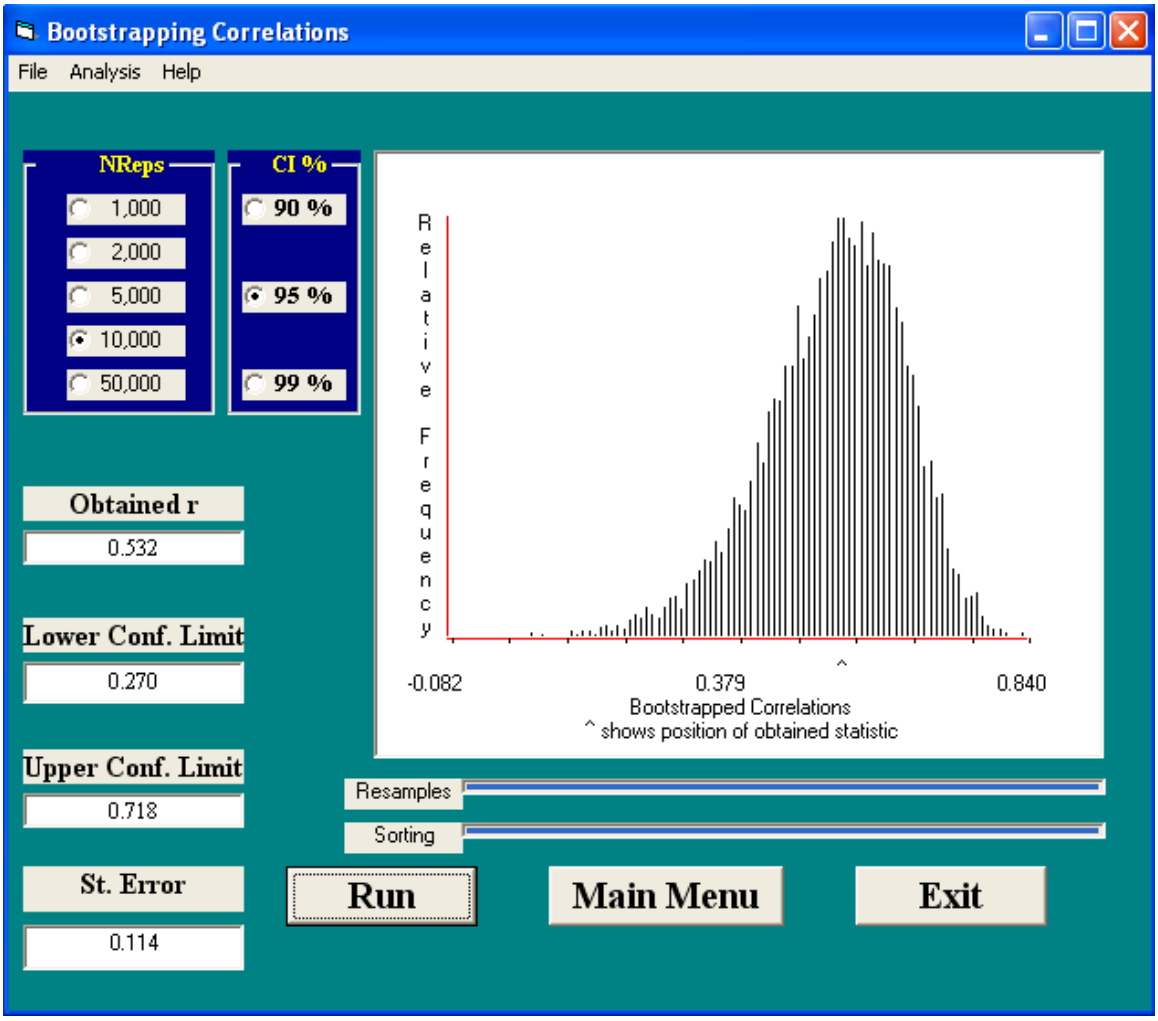
From Howell's main menu, click Analysis, select Randomization Tests, select Correlation via Randomization. Select File, Open, and locate your data file, select it, and click Open. Under the number of replications, select 10,000. Now click Run.

The sampling distribution is centered on zero, and it shows that the observed correlation of .532 is quite unlikely if the observed pairing of X,Y values were randomly scrambled.

Bootstrapping Correlations

With bootstrapping we are interested in determining a confidence interval for the population correlation based on our sample of N=28 pairs. With bootstrapping, the 28 X,Y pairs remain linked, but on each resample we draw 28 of those pairs with replacement.

The empirical sampling distribution is centered on the observed correlation of .532, and the sampling distribution is skewed. No assumption is made about the distribution of the population.



Based on this empirical sampling distribution we can conclude that the population correlation is unlikely to be smaller than .27. The interval is quite wide because our sample is so very small.

Howell shows that a conventional confidence interval constructed by using Fisher's transformation runs from .200 to .756. As is generally found, the bootstrapped interval is a bit narrower.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-12121	3083.0		-3.932	.000	-18179	-6063
	salbegin Beginning Salary	1.914	.046	.882	41.271	.000	1.823	2.005
	jobtime Months since Hire	172.297	36.276	.102	4.750	.000	101.014	243.581

a. Dependent Variable: salary Current Salary

An SPSS macro to estimate bootstrapped confidence limits for B and beta regression weights was applied to the SPSS data set of salaries of bank employees. Above is the regression analysis and below is output from two bootstrap runs, one with N=100 replications and the other with N=1000 replications.

In this example, the bootstrap limits agree quite closely with the results from regression. The sample is large and the distributions do not differ substantially from normal.

Statistics

		salbegin_B	salbegin_Beta	jobtime_B	jobtime_Beta
N	Valid	100	100	100	100
	Missing	0	0	0	0
Percentiles	2.5	1.75618	.85160	87.21404	.05243
	50	1.92066	.88115	175.29314	.10296
	97.5	2.18686	.91620	261.56327	.14651

Statistics

		salbegin_B	salbegin_Beta	jobtime_B	jobtime_Beta
N	Valid	1000	1000	1000	1000
	Missing	0	0	0	0
Percentiles	2.5	1.76992	.85045	101.20384	.05973
	50	1.92205	.88450	174.83254	.10289
	97.5	2.13976	.91242	246.29098	.14380

References for Resampling

Efron, Bradley (1982). *The jackknife, the bootstrap, and other resampling plans*. CBMS 38, SIAM-NSF. This is the classic reference.

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Hesterberg, T., Moore, D.S., Monaghan, S., Clipson, A., & Epstein, R. (2003). Bootstrap methods and permutation tests. Chapter 14 in Moore, D.S. & McCabe, G.P. *Introduction to the practice of statistics, 5th edition*.

Mooney, C.Z. & Duval, R.D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-095. Newbury Park, CA: Sage.

Mooney, C.Z. (1997). *Monte Carlo simulation*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-116. Thousand Oaks, CA: Sage.

Cirincione, C. and G. A. Gurrieri (1997). *Research methodology: Computer intensive methods in the social sciences. Social Science Computer Review, Vol. 15, No. 1*: 83-97.

Parking Lot study: Ruback, R.B. and D. Juieng (1997). "Territorial Defense in Parking Lots: Retaliation Against Waiting Drivers." *Journal of Applied Social Psychology* 27(9):821-834.

Software:

R is a free software environment for statistical computing and graphics. www.**r-project**.org/ S-PLUS has special libraries for resampling. Hesterberg provides additional code.

Good, Phillip (2005). *Introduction to Statistics via Resampling Methods and R/S-PLUS*. Wiley.

SAS macro at http://www2.sas.com/proceedings/sugi22/STATS/PAPER295.PDF for bootstrapping a confidence interval for Cohen's kappa.

Howell, D.C. http://www.uvm.edu/~dhowell/StatPages