Page One

Campus Computing News

Accessing SmartForce CBT

UNT General Access Labs: What We Did This Summer

EduTex: It's Closer Than You Think

Getting Started With at ColdFusion at UNT

Today's Cartoon

RSS Matters

SAS Corner

The Network Connection

List of the Month

WWW@UNT.EDU

Short Courses

IRC News

Staff Activities

Subscribe to Benchmarks *Online*

Research and Statistical Support
University of North Texas

RSS Matters

The Calculation of Statistical Power Using the Percentile Bootstrap and Robust Estimation

By Dr. Rich Herrington, Research and Statistical Support Consultant

Last month we examined the percentile bootstrap, this month we demonstrate the calculation of statistical power using the percentile bootstrap and robust estimation. The GNU S language, "R" is used to implement this procedure. R is a statistical programming environment that is a clone of the S and S-Plus language developed at Lucent Technologies. In the following document we illustrate the use of a GNU Web interface to the R engine on the "rss" server, http://rss.acs.unt.edu/cgi-bin/R/Rprog. This GNU Web interface is a derivative of the "Rcgi" Perl scripts available for download from the CRAN website, http://www.cran.r-project.org (the main "R" website). Scripts can be submitted interactively, edited, and be re-submitted with changed parameters by selecting the hypertext link buttons that appear below the figures. For example, clicking the "Run Program" button below samples 1000 random numbers from a normal distribution, then uses nonparametric density estimation to fit a density curve to the data. To view any text output, scroll to the bottom of the browser window. To view the density curve, select the "Display Graphic" link. The script can be edited and resubmitted by changing the script in the form window and then selecting "Run the R Program". Selecting the browser "back page" button will return the reader to this document.

```
x<-rnorm(1000)
density(x)
plot(density(x))
```

Run Program

The Importance of Power and Effect Size

The techniques of statistical power analysis, effect size estimation, and sample size estimation are important methods in statistics and research methodology (Cohen, 1988). Briefly, the power of a statistical test is the probability of rejecting the null hypothesis given that the alternate hypothesis is true; the effect size is the degree to which the null hypothesis is false in relation to the alternate hypothesis; type II error is the probability of failing to reject the null hypothesis when it needs to be rejected in favor of the alternate hypothesis; and type I error is the probability of incorrectly rejecting the null hypothesis. Proper sample size estimation allows one to achieve an acceptable level of power for a statistical test, thereby setting the type II error at a pre-specified level. Historically, for the social sciences, neglect of these topics have led to a long standing controversy surrounding the interpretation of statistical tests in the research community (Cohen, 1993). Following Jacob Cohen?s (1965, 1962) initial work on the power of statistical tests in behavioral research, many researchers and authors have pointed out the importance of statistical power analysis. Textbooks and articles have appeared that provide tables of power and sample sizes (Cohen, 1988). Additionally, several computer programs which perform exact power analysis assuming normal theory have appeared (Bradley, Helmstreet, & Zeigenhagen, 1992; Faul & Erdfelder, 1992). Despite these recommendations, and availability of resources for power calculation, Cohen has argued that researchers continue to ignore power analysis (Cohen, 1994). Sedlmeier and Gigerenzer, G. (1989) reported a power review of the 1984 volume of the Journal of Abnormal Psychology showing that there was not any marked improvement in the power of statistical tests appearing in the literature. As recent as 1997, a methodological study has found that the power of statistical tests are not taken into account by researchers and that they continue to run a high risk of type II error (Clark-Carter, 1997). Cohen (1988) has suggested that the neglect of power analysis exemplifies the slow movement of methodological advance. Cohen has also suggested a lack of consciousness regarding effect size, coupled with an overriding concern with the accompanying ?p? value (Cohen, 1992; 1994). Despite this unawareness on the part of editors and researchers, there has been a recent shift in the editorial practices of the American Psychological Association (APA, 1994). The manual notes that, ?Neither of the two types of probability values reflects the importance or magnitude of an effect because both depend on sample size?you are encouraged to provide effect-size information (APA, 1994, p.18).? Following these editorial changes, in 1996 APA established a task force that, among other goals, reexamined the role of statistical hypothesis testing in the methodological practices of Psychology (http://www.apa.org/science/tfsi.html). The Task Force on Statistical Inference (TFSI) met twice in two years after which a preliminary report was circulated that indicated its intention to examine issues beyond null hypothesis significance testing. After the second meeting, the task force recommended several possibilities for further action, one of which was to revise the statistical sections of the American Psychological Association Publication Manual (APA, 1994). A report was generated following those meetings (http://www.apa.org/journals/amp/amp548594.html). Neglect of power not only decreases the recognition of interesting effects (type II error), but it also has a negative effect on the ability of researchers to establish statistical consensus through replication. Ottenbacher (1996) points out that, ?The apparently paradoxical conclusion is that the more often we are well guided by theory and prior observation, but conduct a low power study, the more we decrease the probability of replication... The responsible investigator must be concerned with statistical power. A concern with power, however, cannot end with its calculation. Because the ability to detect treatments must be optimized, the responsible scientist must also be concerned with factors that determine effect size?. Most treatments of power analysis deal with the calculation of power for parametric statistics where normal theory assumptions are required (e.g. t-test, F-tests). The calculation of power for robust statistics or nonstandard nonparametric statistics are not addressed at a practical level. For example, Cohen?s book on power analysis (1988) concentrates mainly on ANOVA and regression models and some standard nonparametric tests such as the chi-square test. What is not addressed is how violations of normality assumptions affect power estimates. The bootstrap technique can be useful for exploring how statistical power is affected by non-normality.

Estimating Power with the Bootstrap

Beran (1986) provided mathematical and simulation results that show that a statistical test for a null hypothesis can be constructed by bootstrapping the null distribution for the test statistic. Beran also proved that the power of the test against an alternative can itself be estimated by simulation. The uniform consistency of these simulated power functions is the main result of Beran?s mathematical proof. Additionally, Beran performed a limited numerical study of the univariate bootstrap t-test and the associated power function. The null hypothesis value of the mean difference was zero; the nominal test level alpha was .05; and the sample size was 20. The critical value of the bootstrap test was obtained from the simulated null distribution using 200 bootstrap samples. The power of the bootstrap t-test was approximated by Monte Carlo simulation using 1000 standard normal samples. Thus, the simulation used 200 bootstrap samples for the critical value loop and 1000 bootstrap samples for the outer loop. Even at sample size 20, Beran found that the power function of the bootstrap test is almost indistinguishable from that of the classical t-test. Yuan (2001) applied Beran?s general theory of re-sampling to a covariance structure analysis framework. Yuan found that, with several data sets, robust estimators could be combined with the bootstrap to allow researchers to be in the position of finding an almost optimal procedure for evaluating covariance structure models (Yuan, 2001). Additionally, based on Beran?s results, Yuan provided an algorithm for determining sample size for a given level of power. A great advantage of calculating the critical value from the simulated null sampling distribution is that the empirical estimate of the critical value is asymptotically consistent with the true population value, and no assumptions are made regarding the shape of the null sampling distribution. Consequently, each statistical test (i.e. mean difference test) that is performed on a simulated bootstrap sample, is compared to this critical value, and since the critical value was constructed from the observed data (under the assumption of the null hypothesis), and according to Beran (1986), is a consistent estimate of the population critical value, we can expect proper coverage of the mean difference statistic with the bootstrap confidence intervals, based on this critical value. This is essential for calculating power estimates of test statistics whose sampling distributions are unknown (under the null or the alternate hypothesis), because of violations of assumptions (i.e. normality) or mathematical intractability. Re-sampling under the null hypothesis seems to be the preferred approach in calculating probability values for an observed test statistic (Hall and Wilson, 1991, p. 757). Hall and Wilson give the following guidelines for bootstrap testing in univariate situations, ?The first guideline says that care should be taken to ensure that even if the data might be drawn from a population that fails to satisfy Ho, re-sampling should be done in a way that reflects Ho? (Hall and Wilson, 1991). Bootstrapping under the null hypothesis, for a two group difference test of means, would involve mean centering each group around their respective group means, and sampling with replacement from the whole collection of mean centered scores to produce two new groups of scores (two bootstrap samples) which reflect group differences when the null hypothesis is true (Westfall and Young, 1993, p. 35-36). Furthermore, if one is bootstrapping measures of location other than the mean, one must be sure to create a bootstrap population where the observations are centered around that alternative measure of location (Westfall and Young, 1993, p. 143-144). For example, if one is using a median, or an M-estimate as a measure of location, then one would center around that measure to insure that the null hypotheses are true in the bootstrap population.

The General Bootstrap Power Simulation Algorithm

Beran?s (1986) simulation algorithm is presented as a sequence of steps (for a two-sided difference of location):

Step 1 - Generate the bootstrap null distribution using bootstrap re-sampling: A) Re-center the data vector x and the data vector y around their respective measures of location. B) Stack the data vectors x and y into a single vector, z. Vector z is now considered the in-hand, proxy population. C) Re-sample with replacement from vector z to produce a bootstrap sample for group x1 with length of the original group x. Repeat this re-sampling to produce a group y1. D) Calculate a measure of location for both groups (e.g. mean, M-estimate, trimmed mean, or Windsorized mean). E) Subtract the two location measures. This difference is one bootstrap sample which comprises the empirical null sampling distribution. F) Repeat steps C-E a large number of times to generate the empirical null distribution (suggestions vary widely, 1000 seems to be a sufficient number of bootstrap samples; one might resample 10,000 bootstrap samples to be conservative). The empirical null distribution will be centered on zero.

Step 2. ? Calculate the critical scores that correspond to the 2.5th and 97.5th critical alpha regions under the empirical null distribution: The critical scores are the scores that correspond to the 2.5th and 97.5th percentiles of the empirical null distribution. We can calculate the percentiles using the following approach: round((.05/2)x(#bootstrap samples)) for lower percentile; and round((1-(.05/2))x(#bootstrap samples)). Next, locate the scores that correspond to those percentiles.

Step 3. ? Generate the bootstrap alternative distribution: A) Re-sample with replacement from vector x with replacement to generate a bootstrap sample, x1, with length of original vector x. B) Re-sample with replacement from vector y with replacement to generate a bootstrap sample, y1, with length of original vector y. C) Calculate measures of location for both bootstrap samples x1 and y1. D) Subtract the two measures of location. This is one bootstap difference, and represents the difference between measures of location under the empirical alternate distribution. This empirical distribution is centered on the population difference under the alternate hypothesis.

Step 4. ? Calculate the empirical power of the statistical test: A) Using the upper and lower critical scores for the empirical null hypothesis calculated in step 2., Calculate the number of difference scores in the empirical alternative sampling distribution that are as or more extreme than the critical scores under the null distribution. B) Take the count tallied in step A) and divide by the total number of bootstrap samples. This value is the empirical power for the statistical test that tests differences between groups using whatever location measure is under consideration.

The Data Set

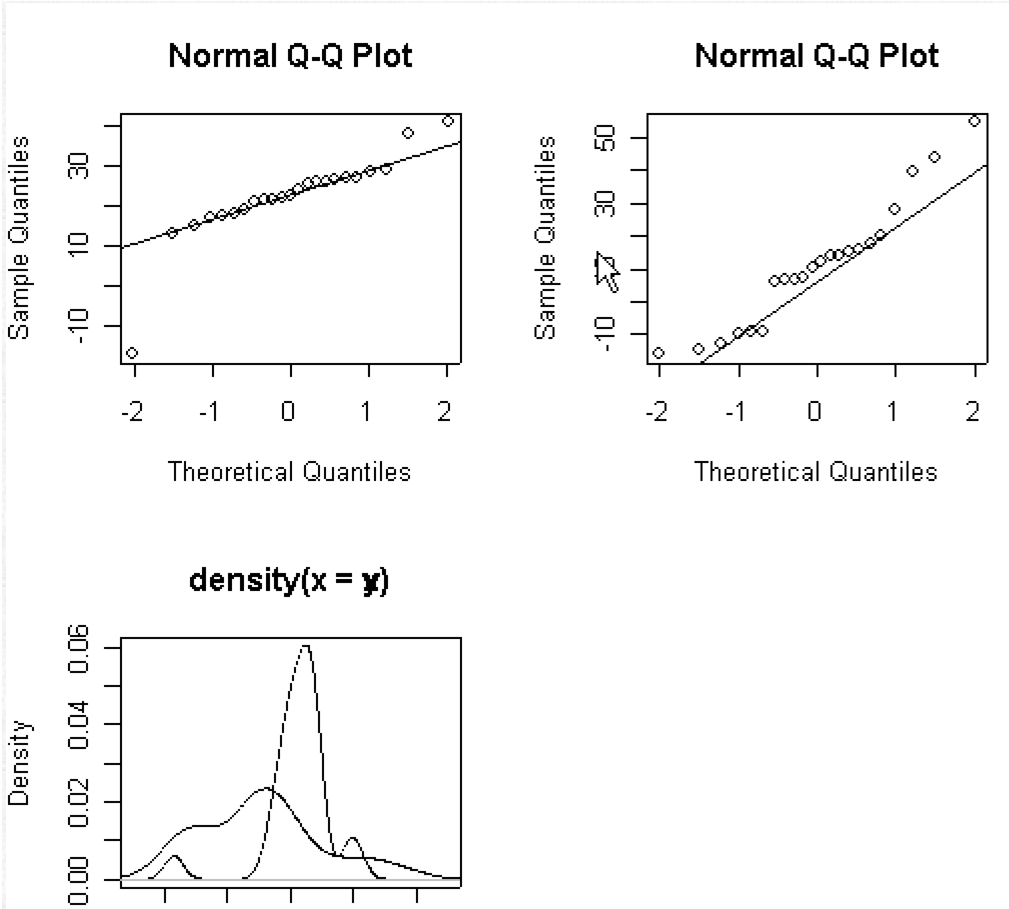
Doksum & Sievers (1976) report data on a study designed to assess the effects of ozone on weight gain in rats. The experimental group consisted of 22 seventy-day old rats kept in an ozone environment for 7 days (group y). The control group consisted of 23 rats of the same age (group x), and were kept in an ozone-free environment. Weight gain is measured in grams. The following R code produces quantile-quantile plots and non-parametric density plots of the two groups of data:

```
###Control Group

x<-c(41.0,38.4,24.4,25.9,21.9,18.3,13.1,27.3,
28.5,-16.9,26.0,17.4,21.8,15.4,27.4,19.2, 22.4,17.7,26.0,29.4,21.4,26.6,22.7)

###Experimental Group (Ozone environment)

y<-c(10.1,6.1,20.4,7.3,14.3,15.5,-9.9,6.8, 28.2,17.9,-9.0,-12.9,14.0,6.6,12.1,15.7,
39.9,-15.9,54.6,-14.7,44.1,-9.0)

summary(x)
summary(y)

#### Plot all graphs on same page in a
# 2x2 grid

par(mfrow=c(2,2))
qqnorm(x)
qqline(x)
qqnorm(y)
qqline(y)

#### Plot both graphs on same graph
plot(density(x), ylim=c(0,.06), xlim=c(-30,70))
par(new=T)
plot(density(y), ylim=c(0,.06), xlim=c(-30,70))
```

Run Program

Results

The following output is produced. Group x (control group) is in the upper left panel, and group y (experimental group) is in the upper right panel. Both groups show substantial deviations away from normality. Deviations away from the straight line indicate deviations away from normality. In the lower panel, non-parametric density estimates of both groups are plotted on the same graph. The more peaked, narrower density function is the control group, and the less peaked, more dispersed density function is experimental group.



Using GNU S ("R") to Calculate Statistical Power Using the Bootstrap and Robust Estimation

In this section, we use M-estimation as measures of location for the control and experimental group. Bootstrap p-values, confidence intervals and power for the difference between the M-estimates are calculated. Additionally, a classical t-test is calculated for comparison:

```
# Function for extracting M-Estimator from the MASS Library

library(MASS)
mest.MASS<-function(x, k=1.28){
location<-hubers(x, k)$mu
}

k<-1.28

#### Calculate the size of the two groups
N1<-length(x)
N2<-length(y)

##### Number of bootstrap samples
nboot<-399

### Alpha criterion

alpha<-.05

#### Non-directional test=2, directional test=1

empTAIL<-2

#### Assign type of estimator (mean, median, mest.MASS)
est<-mest.MASS

######## Re-center data so that H0 holds in original sample

z.x<-x
z.y<-y

# Center x vector according to estimation method

c.x<-(z.x-est(z.x, k))

# Center y vector according to estimation method

c.y<-(z.y-est(z.y, k))

# Stack x and y vector

cx.cy<-c(c.x,c.y)

# Sampling from H0: Population is stacked x and y vector;
# Sample with replacement
# Stack x and y vector

h0datax<-matrix(sample(cx.cy, size=N1 * nboot, replace=T), nrow=nboot)
h0datay<-matrix(sample(cx.cy, size=N2 * nboot, replace=T), nrow=nboot)
h0bvecx<-apply(h0datax,1,est,k)
h0bvecy<-apply(h0datay,1,est,k)
h0bvec<-sort(h0bvecx-h0bvecy)
```

```
# Function for extracting M-Estimator from the MASS Library

library(MASS)
mest.MASS<-function(x, k=1.28){
location<-hubers(x, k)$mu
}

k<-1.28

#### Calculate the size of the two groups
N1<-length(x)
N2<-length(y)

##### Number of bootstrap samples
nboot<-399

### Alpha criterion

alpha<-.05

#### Non-directional test=2, directional test=1

empTAIL<-2

#### Assign type of estimator (mean, median, mest.MASS)
est<-mest.MASS

######## Re-center data so that H0 holds in original sample

z.x<-x
z.y<-y

# Center x vector according to estimation method

c.x<-(z.x-est(z.x, k))

# Center y vector according to estimation method

c.y<-(z.y-est(z.y, k))

# Stack x and y vector

cx.cy<-c(c.x,c.y)

# Sampling from H0: Population is stacked x and y vector;
# Sample with replacement
# Stack x and y vector

h0datax<-matrix(sample(cx.cy, size=N1 * nboot, replace=T), nrow=nboot)
h0datay<-matrix(sample(cx.cy, size=N2 * nboot, replace=T), nrow=nboot)
h0bvecx<-apply(h0datax,1,est,k)
h0bvecy<-apply(h0datay,1,est,k)
h0bvec<-sort(h0bvecx-h0bvecy)

###### Calculate critcal cutoffs

critup<-quantile(h0bvec,.975)

critlow<-quantile(h0bvec,.025)

#### Calculate two-sided mean difference test probability for
#    observed difference

diff.empirical<-est(z.x, k)-est(z.y, k)
count<-length(h0bvec[abs(h0bvec)>=abs(diff.empirical)])
pvalue.empirical<-count/nboot

# Sampling from H1: Sample wih replacement from # x and y vectors seperately

h1datax<-matrix(sample(x, size=N1 * nboot, replace=T), nrow=nboot)
h1datay<-matrix(sample(y, size=N2 * nboot, replace=T), nrow=nboot)

#### Apply M-estimator to each data vector and accumulate in
#    new data vectors

h1bvecx<-apply(h1datax, 1, est, k)
h1bvecy<-apply(h1datay, 1, est, k)

#### Sort and subtract Bootstrap M-estimate data vectors
#    and accumulate into a M-estimator difference vector

h1bvec<-sort(h1bvecx-h1bvecy)

### Calculate the upper and lower cuttoff percentiles for
#   the lower and upper alpha criterion

effectlow<-round((alpha/empTAIL)* nboot)
effectup<-round((1-alpha/empTAIL)* nboot)

#### Calculate Empirical Power

countup<-length(h1bvec[h1bvec>=critup])
countlow<-length(h1bvec[h1bvec<=critlow])
power.twotail<-(countup+countlow)/nboot

### Calculate M-estimate differences that correspond to the
#   upper and lower alpha criterion cuttoffs

h1.ci<-list(ci=c(h1bvec[effectlow], h1bvec[effectup]))

###Display Results

t.test(x,y)
pvalue.empirical
power.twotail
h1.ci
```

```
# Function for extracting M-Estimator from the MASS Library

library(MASS)
mest.MASS<-function(x, k=1.28){
location<-hubers(x, k)$mu
}

k<-1.28

#### Calculate the size of the two groups
N1<-length(x)
N2<-length(y)

##### Number of bootstrap samples
nboot<-399

### Alpha criterion

alpha<-.05

#### Non-directional test=2, directional test=1

empTAIL<-2

#### Assign type of estimator (mean, median, mest.MASS)
est<-mest.MASS

######## Re-center data so that H0 holds in original sample

z.x<-x
z.y<-y

# Center x vector according to estimation method

c.x<-(z.x-est(z.x, k))

# Center y vector according to estimation method

c.y<-(z.y-est(z.y, k))

# Stack x and y vector

cx.cy<-c(c.x,c.y)

# Sampling from H0: Population is stacked x and y vector;
# Sample with replacement
# Stack x and y vector

h0datax<-matrix(sample(cx.cy, size=N1 * nboot, replace=T), nrow=nboot)
h0datay<-matrix(sample(cx.cy, size=N2 * nboot, replace=T), nrow=nboot)
h0bvecx<-apply(h0datax,1,est,k)
h0bvecy<-apply(h0datay,1,est,k)
h0bvec<-sort(h0bvecx-h0bvecy)

###### Calculate critcal cutoffs

critup<-quantile(h0bvec,.975)

critlow<-quantile(h0bvec,.025)

#### Calculate two-sided mean difference test probability for
#    observed difference

diff.empirical<-est(z.x, k)-est(z.y, k)
count<-length(h0bvec[abs(h0bvec)>=abs(diff.empirical)])
pvalue.empirical<-count/nboot

# Sampling from H1: Sample wih replacement from # x and y vectors seperately

h1datax<-matrix(sample(x, size=N1 * nboot, replace=T), nrow=nboot)
h1datay<-matrix(sample(y, size=N2 * nboot, replace=T), nrow=nboot)

#### Apply M-estimator to each data vector and accumulate in
#    new data vectors

h1bvecx<-apply(h1datax, 1, est, k)
h1bvecy<-apply(h1datay, 1, est, k)

#### Sort and subtract Bootstrap M-estimate data vectors
#    and accumulate into a M-estimator difference vector

h1bvec<-sort(h1bvecx-h1bvecy)

### Calculate the upper and lower cuttoff percentiles for
#   the lower and upper alpha criterion

effectlow<-round((alpha/empTAIL)* nboot)
effectup<-round((1-alpha/empTAIL)* nboot)

#### Calculate Empirical Power

countup<-length(h1bvec[h1bvec>=critup])
countlow<-length(h1bvec[h1bvec<=critlow])
power.twotail<-(countup+countlow)/nboot

### Calculate M-estimate differences that correspond to the
#   upper and lower alpha criterion cuttoffs

h1.ci<-list(ci=c(h1bvec[effectlow], h1bvec[effectup]))

###Display Results

t.test(x,y)
pvalue.empirical
power.twotail
h1.ci
```

Run Program

Results

The following results are produced:

Welch Independent Two Sample t-test:

data: x and y
t = 2.4585, df = 32.909, p-value = 0.01938

alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
(1.964178, 20.826336)
sample estimates:
mean of x mean of y
22.40435 11.00909

Bootstrap statistics based on difference between M-estimates:

Bootstrap Empirical P-value
>

pvalue.empirical
[1] 0

Bootstrap Empirical Power
> power.twotail
[1] 0.9331104

Bootstrap Confidence Intervals
> h1.ci
$ci
[1] 4.117494 21.818252

The M-estimate confidence intervals are much narrower that the classical confidence intervals. With 399 bootstrap samples, not one bootstrap sample exceeded the observed difference, giving a p value less than 1/399=.0025. The non-parametric bootstrap power for the difference in M-estimates was .933.

Conclusions

The bootstrap and robust estimation provide a method for improving statistical power whenever the data can be described as having heavy-tailed distributions. Furthermore, an estimate of power based on the percentile bootstrap is non-parametric, and does not depend on normal theory assumptions. Bootstrap power estimation is a general methodology that can be used to calculate power for many different kinds of statistical estimators (e.g. mean, median, or M-estimates).

Announcements

GNU S ("R") on SOL

The Research and Statistical Support Office (RSS) in conjunction with the UNIX support group in the Academic Computing Center have made the decision to place GNU S or "R" on the main UNIX research computer, SOL. We are hoping to get R and it's supporting libraries installed in the next month. This will provide an alternative to the S-PLUS language that already exists on SOL. SOL accounts are available to both students and faculty for research purposes.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association (4th ed.).* Washington, DC: Author.

Beran, R (1986). Simulated Power Functions. *The Annals of Statistics, 14(1),* 151-173.

Bradley, D. R., R. L. Helmstreet, and S. T. Zeigenhagen. 1992. A simulation laboratory for statistics. *Behaviour Research Methods, Instruments, and Computers 24:* 190-204.

Clark-Carter, D. (1997). The account taken of statistical power in research published in the British Journal of Psychology. *British Journal of Psychology, 88,* 71-83.

Cohen, J. (1995). The earth is round (p < .05): Rejoinder. *American Psychologist, 49(12),* 1103.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49(12),* 997-1003.

Cohen J. (1992) A power primer. *Psychological Bulletin, 112,* 155-159.

Cohen, J. (1990). Things I have learned (So Far). *American Psychologist, 45,* 1304-1312.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition.* Lawrence Erlbaum
Associates, Inc., Hillsdale, New Jersey.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology (pp. 95-121).* New York: McGraw-Hill.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology,* 65, 145-153.

Doksum, K.A. & Sievers, G.L. (1976). Plotting with confidence: graphical comparisons of two populations. Biometrika 63, 421-434.

Ottenbacher, K.J. (1996). The Power of Replications and Replications of Power. *The American Statistician, 50(3),* 271-275.

Hall P, Wilson SR (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics, 47,* 757-762.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Yuan, K. (2001). Bootstrap Approach to inference and power analysis based on three test statistics for covariance structure models. Under review.

Westfall, P.H. (1993). *Re-sampling based multiple testing: examples & methods for p-Value adjustment.* Wiley, New York.