Page One

Campus Computing News

New Software Available

Important Summer Reading

Free Virus Protection for Home PCs

Today's Cartoon

RSS Matters

SAS Corner

The Network Connection

Link 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 previous article in this series can be found in a previous issue of Benchmarks Online: Controlling the False Discovery Rate in Multiple Hypothesis Testing - Ed.

Using Robust Mean and Robust Variance Estimates to Calculate Robust Effect Size

By Dr. Rich Herrington, Research and Statistical Support Consultant

This month we demonstrate the calculation of robust effect sizes. 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 creates a vector of 100 random normal deviates; displays the results; sorts and displays the results; then creates a histogram and a density plot of the random numbers. To view any text output, scroll to the bottom of the browser window. To view any graphical output, 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(100)
y<-sort(x)
y
hist(x)
density(x)
```

Run Program

Introduction - Calculating Power and Effect Size

Power analysis involves the relationships between four variables involved in statistical inference: sample size (N), a significance criterion (α), the population effect size (d_{cohen}), and statistical power. For any statistical inference, these relationships are a function of the other three (Cohen, 1988). For research planning, it is most useful to determine the N necessary to have a specified power for a given α and d_{cohen} . The statistical power of a test is the long term probability of rejecting H_0 (null hypothesis) given a specified α criterion and sample size N. When the effect size is not equal to zero, H_0 , is false, so a failure to reject H_0 is a decision error on the part of the researcher. This is called a type II error (β) and is related mathematically to power. The probability of rejecting the null if it needs to be rejected (power) is one minus the type II error ($1-\beta$). Figure 1. below is a graphical representation of the relationship between the null distribution, the alternate distribution, and the critical scores under the null distribution. The area underneath the H_1 distribution (the alternate distribution), past the critical score of the left tail of H_0 , and past the critical score of the right tail of H_0 , represents the power of the statistical test being performed (the shaded area).

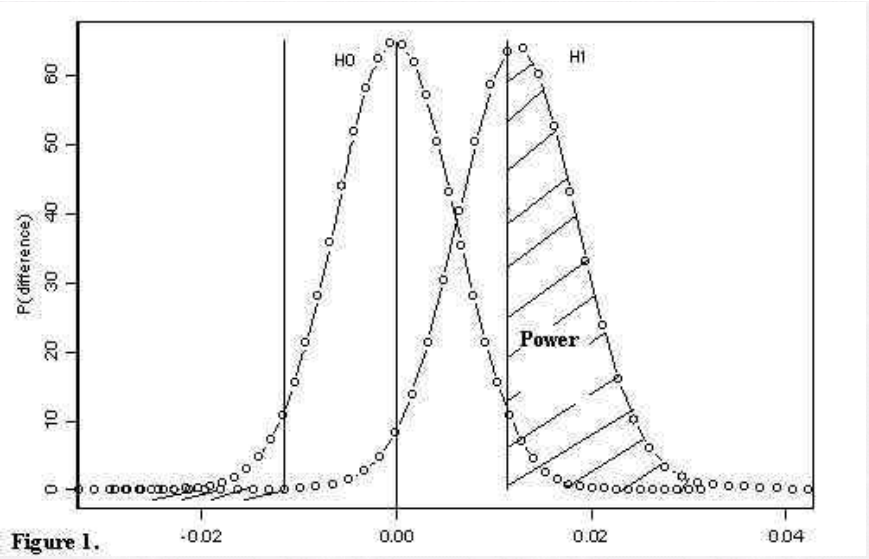


Figure 1.

Effect size is the degree to which H_0 (null hypothesis) is false and is indexed by the discrepancy between the null hypothesis and the alternate hypothesis. Power analysis specifies a non-centrality parameter to quantify this discrepancy. The noncentrality parameter for the difference between means is:

$$\delta = \left| \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_{diff}} \right|$$

where the difference between estimated population means is scaled in $\hat{\sigma}_{diff}$ units (known as the estimated standard error of the difference between means):

$$\hat{\sigma}_{diff} = \sqrt{\hat{\sigma}^2_{pooled} \cdot \left(\frac{n_1 \cdot n_2}{n_1 + n_2} \right)}$$

where,

$$\hat{\sigma}^2_{pooled} = \frac{(n_1 - 1) \cdot \hat{\sigma}^2_1 + (n_2 - 1) \cdot \hat{\sigma}^2_2}{n_1 + n_2 - 2}$$

and $\hat{\sigma}$ is the sample estimate of the population standard deviation. The denominator of the non-centrality parameter represents the estimated standard deviation of the sampling distribution for the null hypothesis for differences between means. Usually, $\hat{\sigma}_{diff}$ is calculated on the basis of a formula that assumes normality in the population since the standard deviation of the null sampling distribution cannot be calculated directly on the basis of the observed data without normality assumptions. For robust measures of location (i.e. M-estimate), the numerator would be the difference between two M-estimates, and the denominator would represent the standard deviation of the null hypothesis re-sampling distribution for the difference between M-estimates. For robust estimates (as well as the sample mean), the standard error can be estimated directly by calculating the standard deviation of the bootstrap estimates of the differences between the robust estimates of location (see September 2001 issue of Benchmarks). An alternative effect size for group differences has been advocated by Cohen (1988). Cohen s d_{cohen} measure is based on the pooled estimated population standard deviation:

$$d_{cohen} = \left| \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_{pooled}} \right|$$

Cohen provides guidelines for interpreting the practical importance of an effect size based on d_{cohen} when no prior research is available to anchor d_{cohen} meaningfully. Cohen s rule of thumb for a small, medium and large effect size are based on a wide examination of the typical difference found in psychological data. A small effect size for d_{cohen} is .20; a medium effect size for d_{cohen} is .50; and a large effect size for d_{cohen} is .80 (Cohen, 1992). Equating δ and d_{cohen} using algebra, the expression for δ is:

$$\delta = d_{cohen} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

It is noted that d_{cohen} is not a robust measure of effect size. The pooled sample standard deviation, which is used to estimate the population standard deviation ($\hat{\sigma}$) will be inflated in the presence of outliers thereby biasing the effect size measure. Furthermore, d_{cohen} assumes a normal distribution in the calculation of power estimates.

Measures of Robust Effect Size

Several problems exist with the d_{cohen} measure of effect size. The assumption of equal variances in the population is often dealt with by substituting a pooled variance estimate for σ . With data that appear to have unequal variances, questions arise about how to interpret H_0 . Another criticism of d_{cohen} is that both the location and scale (mean difference and sample standard deviation) of the sample are non-resistant measures. One strategy would be to replace the means and standard deviation with more resistant measures of location and scale. For example, one variation might be a difference of medians divided by MAD (median absolute deviation):

$$d_{median} = \frac{\mu_{M1} - \mu_{M2}}{MAD}$$

where $MAD = MED\left[\left|X_1 - M\right|, \ldots\ldots \left|X_n - M\right|\right]$ and M is the median of the scores in the control group. This effect size estimator does not seem like a good candidate since both the median and MAD are both known to be inefficient for Normal distributions compared to the mean and standard deviation.

Robust Effect Size based on M-estimators

Lax (1985) examined the performance of 17 different estimators of scale with heavy tailed distributions. Lax examined the performance of these scale estimators with the Normal distribution; a distribution with Cauchy tails (large kurtosis relative to the Normal The Slash dist.); and a mixture distribution of N(0,1) and N(0,100) for samples of size 20. The mixture distribution had 19 points sampled from N(0,1) and 1 point sampled from N(0,100) (One-Wild dist.). Lax combined the efficiencies (see July 2001 issue of Benchmarks) of the estimators for the three distributions into what was defined as triefficiency . The biweight midvariance (with c=9) estimator performed best, with favorable efficiencies across all scenarios: Normal (86.7%), One-Wild (85.8), and Slash (86.1). Following Wilcox (1997) the biweight midvariance can be calculated as follows. Setting (with c=9, M=sample median):

$$Y_i = \frac{X_i - M}{c \cdot MAD} \qquad \text{and} \qquad a_i = \begin{cases} 1, \text{ if } |Y_i| < 1 \\ 0, \text{ if } |Y_i| \geq 1 \end{cases}$$

the following is calculated:

$$\hat{\xi}_{bi} = \frac{\sqrt{n} \cdot \sqrt{\sum a_i (X_i - M)^2 (1 - Y_i^2)^4}}{\left|\sum a_i (1 - Y_i^2)(1 - 5Y_i^2)\right|}$$

The square of $\hat{\xi}_{bi}$ is called the biweight midvariance. It appears to have a breakdown point of approximately .5 (Hoaglin, Moesteller, & Tukey, 1983). Based on this robust variance, the following robust effect size can be calculated:

$$d_{robust} = \frac{\hat{\mu}_{Mest1} - \hat{\mu}_{Mest2}}{\hat{\xi}_{bi1}}$$

where, $\hat{\mu}_{Mest1}$ is the robust M-estimator for group 1 (using Huber objective function, with k=1.28 for both groups), $\hat{\mu}_{Mest2}$ is the robust M-estimator for group 2, and $\hat{\xi}_{bi1}$ is the square root of the biweight midvariance for group 1 (control group). The robust effects size d_{robust} does not assume equal variances among groups since only the robust variance for the control group is used (alternatively, a pooled estimated of both the control and experimental group biweight midvariances could be used, assuming equal variances).

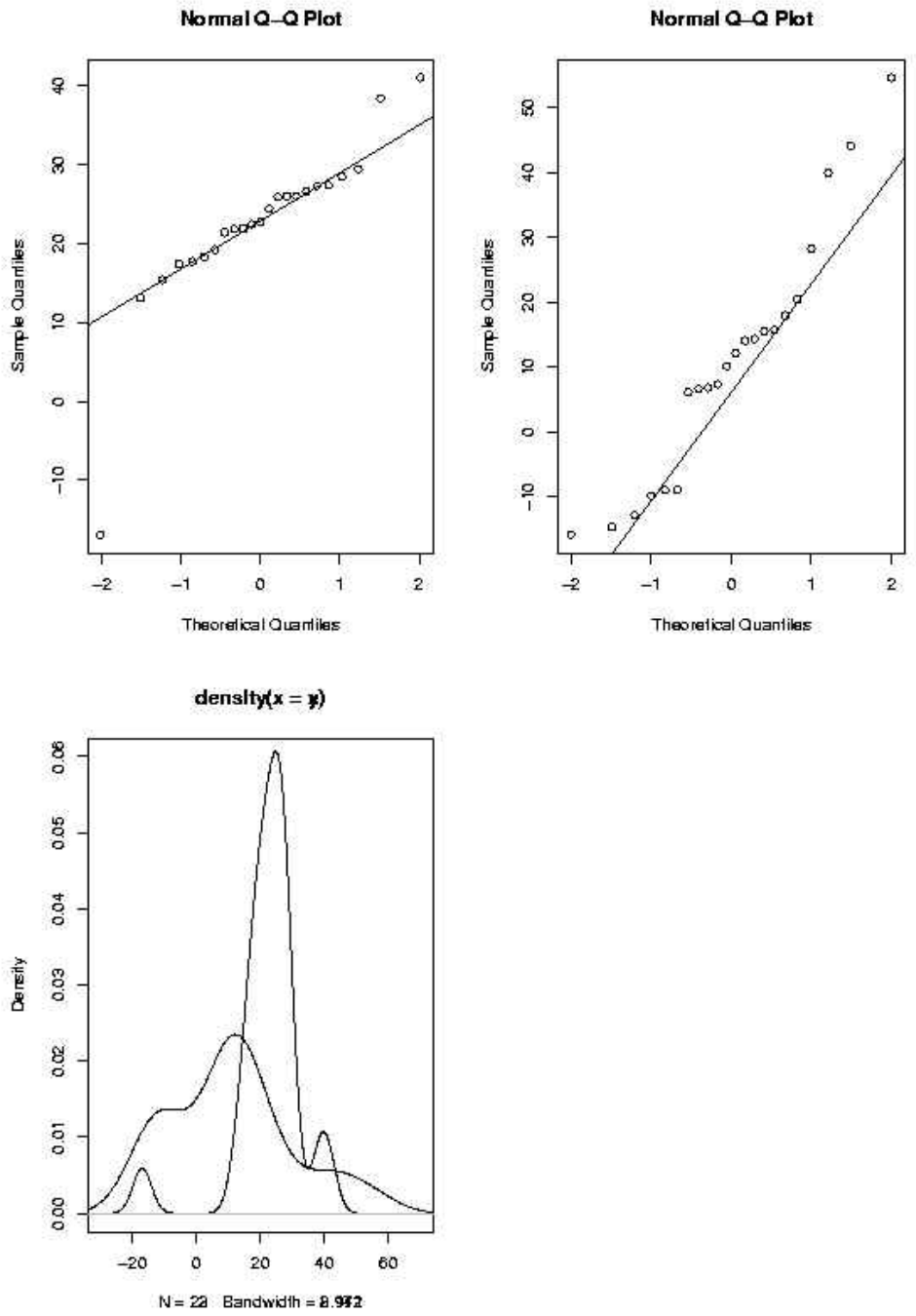
An Example Using GNU-S ("R")

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:

```
#### Assign data to vector x and y for the
#    two independent groups

library(eda)

#### Control Group (Ozone free environment)

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)

par(mfrow=c(2,2)) #### Plot all graphs on same page in a
              #    2x2 grid

qqnorm(x)
qqline(x)

qqnorm(y)
qqline(y)

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

Run Program

Resulting qqnorm plots and density plots from R code above:



Both groups appear to have right and left tails which are "heavy". It appears as if the classical mean difference between the groups will be underestimated (smaller). The R code below estimates both the classical means, and robust means; classical estimated pooled standard deviation, and estimated pooled robust root biweight midvariance.

```
## Function to calculate pooled variance
std.dev.pooled<-function(std.dev1, std.dev2, n1, n2){
    pooled.std.dev<-(((n1-1)*std.dev1^2+(n2-1)*std.dev2^2)/
                    (n1+n2-2))^.5
}

## M-estimator function
mest<-function(x,bend=1.28){
#
#  Compute M-estimator of location using Huber's Psi.
#  The default bending constant is 1.28
#
if(mad(x)==0)stop("MAD=0. The M-estimator cannot be computed.")
y<-(x-median(x))/mad(x)  #mad in splus is madn in the book.
A<-sum(hpsi(y,bend))
B<-length(x[abs(y)<=bend])
mest<-median(x)+mad(x)*A/B
repeat{
y<-(x-mest)/mad(x)
A<-sum(hpsi(y,bend))
B<-length(x[abs(y)<=bend])
newmest<-mest+mad(x)*A/B
if(abs(newmest-mest) <.0001)break
mest<-newmest
```

```
## Function to calculate pooled variance
std.dev.pooled<-function(std.dev1, std.dev2, n1, n2){
   pooled.std.dev<-(((n1-1)*std.dev1^2+(n2-1)*std.dev2^2)/
                  (n1+n2-2))^.5
}

## M-estimator function
mest<-function(x,bend=1.28){
#
#  Compute M-estimator of location using Huber's Psi.
#  The default bending constant is 1.28
#
if(mad(x)==0)stop("MAD=0. The M-estimator cannot be computed.")
y<-(x-median(x))/mad(x)  #mad in splus is madn in the book.
A<-sum(hpsi(y,bend))
B<-length(x[abs(y)<=bend])
mest<-median(x)+mad(x)*A/B
repeat{
y<-(x-mest)/mad(x)
A<-sum(hpsi(y,bend))
B<-length(x[abs(y)<=bend])
newmest<-mest+mad(x)*A/B
if(abs(newmest-mest) <.0001)break
mest<-newmest
}
mest
}

hpsi<-function(x,bend=1.28){
#
#   Evaluate Huber`s Psi function for each value in the vector x
#   The bending constant defaults to 1.28.
#
hpsi<-ifelse(abs(x)<=bend,x,bend*sign(x))
hpsi
}


## Biweight midvariance function
bivar<-function(x){
# compute biweight midvariance of x
m<-median(x)
u<-abs((x-m)/(9*qnorm(.75)*mad(x)))
top<-length(x)*sum((x[u<=1]-m)^2*(1-u[u<=1]^2)^4)
bot<-sum((1-u[u<=1]^2)*(1-5*u[u<=1]^2))
bi<-top/bot^2
bi
}


#### Control Group (Ozone free environment)

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)

## Classical Mean
mean(x)
mean(y)

## Robust Mean (M-estimator)
mest(x)
mest(y)

## Classical Pooled Variance
std.dev.pooled.x.y<-std.dev.pooled(var(x)^.5, var(y)^.5, length(x), length(y))
std.dev.pooled.x.y

## Pooled Robust variance
robust.std.dev.pooled.x.y<-std.dev.pooled(bivar(x)^.5, bivar(y)^.5, length(x), length(y))
robust.std.dev.pooled.x.y

## Cohens d (non-robust)
cohen.d<-(mean(x)-mean(y))/std.dev.pooled.x.y
cohen.d

## Robust Cohen's d
robust.d<-(mest(x)-mest(y))/robust.std.dev.pooled.x.y
robust.d
```

```
## Function to calculate pooled variance
std.dev.pooled<-function(std.dev1, std.dev2, n1, n2){
    pooled.std.dev<-(((n1-1)*std.dev1^2+(n2-1)*std.dev2^2)/
                (n1+n2-2))^.5
}

## M-estimator function
mest<-function(x,bend=1.28){
#
#  Compute M-estimator of location using Huber's Psi.
#  The default bending constant is 1.28
#
if(mad(x)==0)stop("MAD=0. The M-estimator cannot be computed.")
y<-(x-median(x))/mad(x)  #mad in splus is madn in the book.
A<-sum(hpsi(y,bend))
B<-length(x[abs(y)<=bend])
mest<-median(x)+mad(x)*A/B
repeat{
y<-(x-mest)/mad(x)
A<-sum(hpsi(y,bend))
B<-length(x[abs(y)<=bend])
newmest<-mest+mad(x)*A/B
if(abs(newmest-mest) <.0001)break
mest<-newmest
}
mest
}

hpsi<-function(x,bend=1.28){
#
#   Evaluate Huber`s Psi function for each value in the vector x
#   The bending constant defaults to 1.28.
#
hpsi<-ifelse(abs(x)<=bend,x,bend*sign(x))
hpsi
}


## Biweight midvariance function
bivar<-function(x){
# compute biweight midvariance of x
m<-median(x)
u<-abs((x-m)/(9*qnorm(.75)*mad(x)))
top<-length(x)*sum((x[u<=1]-m)^2*(1-u[u<=1]^2)^4)
bot<-sum((1-u[u<=1]^2)*(1-5*u[u<=1]^2))
```

Run Program

Results and Conclusion

```
> ## Classical Mean
> mean(x)
[1] 22.40435
> mean(y)
[1] 11.00909
>
> ## Robust Mean (M-estimator)
> mest(x)
[1] 23.14211
> mest(y)
[1] 9.687215
>
> ## Classical Pooled Variance
> std.dev.pooled.x.y<-std.dev.pooled(var(x)^.5, var(y)^.5, length(x), length(y))
> std.dev.pooled.x.y
[1] 15.36189
>
> ## Pooled Robust variance
> robust.std.dev.pooled.x.y<-std.dev.pooled(bivar(x)^.5, bivar(y)^.5, length(x), length(y))
> robust.std.dev.pooled.x.y
[1] 14.3583
>
```

The resulting M-estimators suggest that the population control group mean is downwardly biased (23.24 - robust; 22.40 - classical) and the experimental population group mean is biased upwardly (9.69 - robust; 11.01 - classical). Additionally, the robust pooled scale estimate is smaller than the classical pooled scale estimate (14.36 - robust; 15.36 - classical). Using these estimates to calculate Cohen's d measure indicates that the effect size is downwardly biased. Cohen's d based on classical estimators suggests a medium to large effect size (.74), whereas Cohen's d based on robust estimators suggests a very large effect size (.94). In terms of sample size planning for future experiments, the robust Cohen's d would suggest that a much smaller sample size would be needed to achieve the same power for a smaller effect size using non-robust estimators of location and scale - a considerable savings in terms of data that needs to be collected.

```
> ## Cohens d (non robust)
> cohen.d<-(mean(x)-mean(y))/std.dev.pooled.x.y
> cohen.d
[1] 0.7417875
>
> ## Robust Cohen's d
> robust.d<-(mest(x)-mest(y))/robust.std.dev.pooled.x.y
> robust.d
[1] 0.937081
>
```

References

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

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

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

Hoaglin, D. C., Mosteller, F., & Tukey, J.W. (1983). Understanding robust and exploratory data analysis. New York: Wiley.

Lax, D.A. (1985). Robust estimators of scale: finite sample performance in long-tailed symmetric distributions. Journal of the American Statistical Association, 80(391), 736-741.

Wilcox, Rand R. (1997). Introduction to Robust Estimation and Hypothesis Testing. Academic Press, New York.