

P019

## ROBUST\_ES: A SAS® Macro for Computing

### Robust Estimates of Effect Size

Jeffrey D. Kromrey

Kevin B. Coughlin

University of South Florida, Tampa, FL

#### ABSTRACT

Effect sizes are useful statistics that complement null hypothesis testing and confidence interval estimation. Because traditional indices of effect size are sensitive to violations of distributional assumptions, many robust effect size indices have been proposed and described in the methodological literature. The macro described in this paper computes the traditional standardized mean difference and six robust indices of effect size for the two-group case: a standardized trimmed-mean difference, gamma, CL, A, delta, and two estimators of W. The macro is written in SAS/IML and although the computations are limited to the case of two independent groups, the macro may be easily modified to compute effect sizes for other data structures (more than two groups, correlated observations, categorical outcome variables). Through building a convenient method to compute multiple indices of effect size, this paper will encourage researchers to provide their audiences with indications of the practical effects of their findings.

#### INTRODUCTION

Over the years, there has been a concerted effort aimed toward encouraging researchers to always provide some indication of effect size in addition to or in place of the results of statistical significance tests. Effect sizes have been viewed as logically consistent with null hypothesis significance testing and as an important compliment. Yet, despite urgings for the regular reporting of effect sizes, these measures are seldom found in published reports, and are seemingly still far from becoming standard practice (Kirk, 1996, Thompson & Snyder, 1997, 1998). Carver (1993) contends, "statistical significance tells us nothing directly relevant to whether the results we found are large or small, and it tells us nothing with respect to whether the sampling error is large or small. This problem can be eliminated by reporting both effect size and standard errors" (p. 291). Additionally, the reporting of effect sizes assists researchers in planning future research (i.e., the determination of sample size for subsequent investigations) as well as facilitating comparison of results across studies through the use of meta-analytic techniques.

Recent debates regarding reporting of research results support the inclusion of both measures of statistical significance, e.g., *p*-values, and measures of practical significance, e.g., effect sizes (Nix & Barnette, 1998; Thompson, 1998). The 5<sup>th</sup> edition of the style manual for publication from the American Psychological Association (APA, 2001) cites the failure to report effect sizes as a defect in research reporting. Further, the report by Wilkinson and the APA Task Force on Statistical Inference (1999) addresses the need for effect size reporting.

One practical impediment to the use and reporting of effect sizes may stem from poor understanding. Choosing among the various possible effect-size estimates is not always apparent (Rosenthal, 1991), and opinions vary regarding the merits of the various possibilities (Crow, 1991; Gorsuch, 1991; McGraw, 1991; Parker, 1995; Rosenthal, 1991; Strahan, 1991). An important consideration is the extent to which indices of effect size calculated from a sample provide information about the magnitude of effect in the population from which the sample was drawn. That is, the statistical bias and sampling error associated with sample effect size indices are attributes that must be taken into account in developing accurate interpretations of observed effect sizes. Further, the valid interpretation of sample effect sizes must include a consideration of the sensitivity of effect size indices to differences in population distribution shape or differences in population variances (Hogarty & Kromrey, 2001).

Previous research (e.g., Hess & Kromrey, 2003; Kromrey & Hess, 2002; Hogarty & Kromrey, 2001) has suggested that the sensitivity of traditional indices of effect size, such as Cohen's *d*, precludes their valid interpretation under variance heterogeneity and non-normality. However, alternative indices of effect have evidenced notably lower levels of bias under such conditions (Hess & Kromrey, 2004; Hogarty & Kromrey, 2001).

#### EFFECT SIZE INDICES

Traditional measures of effect size for the two-group case (Cohen's *d* or Hedge's *g*) may be used to describe differences in means relative to an assumed common standard deviation. Cohen's *d* is given by

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}}$$

where  $\hat{\sigma}$  is a pooled estimate of the common population standard deviation.

Hedges and Olkin (1985) suggested that the  $d$  index evidences a small sample bias, and provided an adjusted effect size estimate,  $g$ , designed to reduce such bias:

$$g = d \left( 1 - \frac{3}{4N - 9} \right)$$

where  $N$  is the total sample size (i.e.,  $N = n_1 + n_2$ ).

Problems with both of these effect size indices arise when the samples are drawn from populations that are non-normal or heterogeneous in variances.

As an alternative to Cohen's  $d$  or Hedges'  $g$ , robust estimators of location and scale (such as trimmed means and Winsorized variances) may be useful in computing effect size indices (Yuen, 1974; Hedges & Olkin, 1985). This approach is appealing because of its computational ease and strong theoretical properties. The properties of such an index (*trimmed-d*) have been investigated in the context of meta-analytic tests of homogeneity (Hogarty & Kromrey, 1999), a context in which the index evidenced excellent Type I error control and reasonably large statistical power. The trimmed mean for a sample of scores is obtained by eliminating the highest and lowest  $k$  scores from the sample before the mean is computed.

$$\bar{x}_t = \frac{X_{k+1} + X_{k+2} + \dots + X_{n-k}}{n - 2k}$$

Similarly, the Winsorized variance is the sample variance computed by replacing the lowest  $k$  values by the  $(k + 1)$ th value, and replacing the highest  $k$  values by the  $(n - k)$ th value.

$$S_w^2 = \frac{(k+1)(X_{k+1} - \bar{x}_w)^2 + (X_{k+2} - \bar{x}_w)^2 + \dots + (k+1)(X_{n-k} - \bar{x}_w)^2}{n - 2k}$$

where  $\bar{x}_w$  is the Winsorized mean:

$$\bar{x}_w = \frac{(k+1)(X_{k+1}) + X_{k+2} + \dots + (k+1)(X_{n-k})}{n}$$

For the macro provided in this paper,  $k$  was set to the number of observations corresponding to 10% of the respective sample. The *trimmed-d* effect size index is given by

$$\text{Trimmed-d} = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{\hat{\sigma}_w^2}}$$

where the  $\bar{X}_{ti}$  are the sample trimmed means and  $\hat{\sigma}_w^2$  is the Winsorized variance.

Non-parametric indices of effect size have been suggested by several authors. For example, Kraemer and Andrews (1982) have suggested an index,  $\gamma_1^*$ , based on the degree of overlap between samples. Specifically,

$$\gamma_1^* = \Phi^{-1}(q^*)$$

where  $q^*$  is the sample proportion of scores in one group that are less than the median score of the other group, and  $\Phi^{-1}$  is the inverse of the standard normal cumulative distribution function.

The  $\gamma_1^*$  index is, therefore, the normal deviate that corresponds to the proportion  $q^*$ . In practice, if the observed  $q^* = 0$  or 1 (for which the inverse transformation yields negative or positive infinity) the proportion is replaced with  $1/(n + 1)$  or  $n/(n + 1)$ , respectively.

Using a similar type of logic, McGraw and Wong (1992) have proposed a “common language” effect size statistic ( $CL$ ) that expresses the relative frequency with which a score sampled from one distribution will be greater than a score sampled from a second distribution. The  $CL$  statistic is given by

$$CL = \Phi(z^*), \text{ with } z^* = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}$$

The  $CL$  index is thus the proportion of the standard normal distribution that is less than  $z^*$ . Although the  $CL$  index is calculated using sample means and variances (statistics known to be sensitive to non-normality) and the cumulative normal distribution is used to convert  $z^*$  into a proportion, simulation results reported by McGraw and Wong (1992) suggest that this index is relatively robust to violations of normality and homogeneity of variance.

Vargha and Delaney (2000) suggested a measure of stochastic superiority as a different generalization of  $CL$  that applies to distributions that are at least ordinally scaled. This measure, designated as  $A$ , is given by

$$A = P(x_1 > x_2) + .5P(x_1 = x_2)$$

where  $P(X)$  is the probability of event  $X$ .

A sample estimate of this population effect size is given by

$$\hat{A} = \frac{\#(x_1 > x_2) + .5\#(x_1 = x_2)}{n_1 n_2}$$

where  $x_{i1}$  is a member of sample one,  $x_{i2}$  is a member of sample two, and  $n_j$  is the sample size in group  $j$ . The  $\#$  operator denotes the number of occurrences of the event.

A related index, the delta statistic, was proposed by Norman Cliff (1993, 1996), for testing null hypotheses about group differences on ordinal level measurements. The delta statistic is used to test equivalence of probabilities of scores in each group being larger than scores in the other (the property that Cliff (1993) referred to as “dominance”). A sample estimate of the parameter can be obtained by enumerating the number of occurrences of a sample one member having a higher response value than a sample two member, and the number of occurrences of the reverse. This gives the sample statistic

$$delta = \frac{\#(x_{i1} > x_{i2}) - \#(x_{i1} < x_{i2})}{n_1 n_2}$$

This statistic is most easily conceptualized by considering the data in an arrangement called a dominance matrix. This  $n_1$  by  $n_2$  matrix has elements taking the value of 1 if the row response is larger than the column response, -1 if the row response is less than the column response, and 0 if the two responses are identical. The sample value of  $delta$  is simply the average value of the elements in the dominance matrix.

Finally, Wilcox and Muska (1999) have suggested a non-parametric analogue of  $\omega^2$  that estimates the degree of certainty with which an observation can be associated with one population rather than the other. That is, the effect size index  $W$  represents the probability of correctly classifying an observation into one of the two groups. Wilcox and Muska (1999) used a non-parametric classification rule based on a kernel density estimator and compared four methods of estimating  $W$  (a naïve estimator, a leave-one-out cross-validation estimator, a basic bootstrap estimator, and a .632 bootstrap estimator). Although all four estimators evidenced relatively small degrees of statistical bias, the .632 estimator was recommended as providing the best overall performance.

## MACRO ROBUST\_ES

A SAS/IML macro was designed to compute the traditional standardized mean difference effect sizes ( $d$  and  $g$ ) and six robust indices of effect size for the two-group case: the *trimmed-d* effect size,  $\gamma_1^*$ ,  $CL$ ,  $A$ ,  $delta$ , and two estimators of  $W$ . The macro was developed to provide researchers with an easily accessible tool for calculating these

effect sizes. Arguments supplied to the macro include the name of the SAS dataset that contains the two samples of observed scores, the name of the grouping variable and the values of this variable for the two groups, and the names of the dependent variable(s). By default, the macro uses the latest SAS data set created. The grouping variable may be either alphanumeric or numeric and any number of dependent variables may be analyzed with a single call to the macro.

The macro passes the SAS dataset name and the variables to PROC IML for analysis. Within the macro, a do-loop executes all operations on each of the dependent variables specified in the macro arguments. The macro is written with subroutines (modules) for operations that are required multiple times during the analysis. The output from the macro includes a table to present the calculated indices of effect size, as well as descriptive information about the data analyzed. Of course, the macro syntax may be easily modified to write the output to a disk file or to send the descriptive statistics and effect sizes back to regular SAS for further analyses.

## ROBUST\_ES Code

```
%macro ROBUST_ES (DATASET=_LAST_,DVLIST=X1,GROUPVAR=GROUP,VALUE1=1,VALUE2=2);
```

```
proc iml;
```

```
START TRIMMIT(XX,trimpct,trim,T_mean,W_var);
```

```
* +-----+
```

```
  Compute trimmed mean and Winsorized variance
  INPUT: XX = column vector of observed variable values
         TRIMPCT = percent of observations to trim from each tail

  OUTPUT: TRIM = number of observations trimmed from each tail
         T_MEAN = trimmed mean
         W_VAR = Winsorized variance
```

```
+-----+;
```

```
n_obs = NROW(XX);
trim = round((trimpct/100)#n_obs + 0.5);
```

```
* Vector XT is the "trimmed" version of vector XX;
```

```
XT = J(n_obs - 2*trim,1,0);
do t = 1 to (n_obs - 2*trim);
  XT[t] = XX[t+trim];
end;
```

```
* Compute trimmed mean and Winsorized mean;
```

```
T_mean = 0;
W_mean = 0;
do t = 1 to (n_obs - 2*trim);
  if (t = 1 | t = n_obs - 2*trim) then wt = trim + 1;
  if (t ^= 1 & t ^= n_obs - 2*trim) then wt = 1;
  W_mean = W_mean + wt*XT[t];
  T_mean = T_mean + XT[t];
end;
W_mean = W_mean/n_obs;
T_mean = T_mean/(n_obs - 2*trim);
```

```
* Compute Winsorized variance;
```

```
W_var = 0;
do t = 1 to (n_obs - 2*trim);
  if (t = 1 | t = n_obs - 2*trim) then wt = trim + 1;
  if (t ^= 1 & t ^= n_obs - 2*trim) then wt = 1;
  W_var = W_var + wt*(XT[t] - W_mean)**2;
end;
W_var = W_var/(n_obs - 2*trim);
```

```
FINISH;
```

```
START BUBBLE(x,n,c);
```

```
* +-----+
```

```

Simple bubble sort on column of a matrix
INPUT: X = matrix to be sorted
      N = number of rows in the matrix
      C = column of matrix by which to sort

OUTPUT: X = matrix sorted by column C
+-----+;
do i = 1 to n;
  do j = 1 to n-1;
    if x[J,C] > x[J+1,C] then do;
      temp = x[J+1,];
      x[J+1,] = x[J,];
      x[J,] = temp;
    end;
  end;
end;
FINISH;

START BOOTSTRP(ORIG_X,BOOTN,B_X);
* +-----+
  Draw a bootstrap sample
  INPUT: ORIG_X = observed matrix from which to sample
        BOOTN = scalar size of the bootstrap sample to draw

  OUTPUT: B_X = bootstrap sample of size BOOTN drawn from ORIG_X
+-----+;
BIG_N = NROW(ORIG_X);
do i = 1 to BOOTN;
  * +-----+
    Randomly select rows from the matrix ORIG_X to
    create the matrix B_X.
    Sampling is with replacement.
  +-----+;
  ranrow = round(ranuni(0)*(BIG_N + 0.49999));
  if ranrow = 0 then ranrow = 1;
  if i = 1 then do;
    B_X = ORIG_X[ranrow,];
  end;
  if i > 1 then do;
    B_X = B_X//ORIG_X[ranrow,];
  end;
end;
free big_n i ranrow;
FINISH;

START GET_H(VEC_IN,H_EST);
* +-----+
  Compute H for kernel density estimation
  INPUT: VEC_IN = column vector of observations

  OUTPUT: H_EST = scalar estimate of  $h = 1.2(Q75 - Q25)/n^{(1/5)}$ 
+-----+;
n_obs = nrow(VEC_IN);
run bubble(VEC_IN,n_obs,1);
n_25 = round(n_obs/4);
n_75 = round((3#n_obs)/4);
H_EST = (1.2#(VEC_IN[n_75,1] - VEC_IN[n_25,1])) / (n_obs##(1/5));
if h_est = 0 then h_est = .05;
free n_25 n_75;
FINISH;

```

```

START GET_F(VEC_IN,H_EST,TARGET_X,F_X);
* +-----+
  Compute kernel density estimate for TARGET_X
  INPUT: VEC_IN = column vector of observations
         H_EST = estimate of the constant h
         TARGET_X = value of X for which estimate is desired

  OUTPUT: F_X = estimate of probability density
+-----+;
n_obs = nrow(VEC_IN);
* run bubble(VEC_IN,n_obs,1);
AA = 0;
BB = 0;
do i = 1 to n_obs;
  if VEC_IN[i,1] <= (TARGET_X + H_EST) then AA = AA + 1;
  if VEC_IN[i,1] < (TARGET_X - H_EST) then BB = BB + 1;
end;
F_X = (AA - BB) / (2#n_obs#H_EST);
free AA BB n_obs;
FINISH;

START GET_QAP(VEC1,VEC2,Q_AP,Q_1,Q_2);
* +-----+
  Compute the "apparent" estimator of the effect size Q
  INPUT: VEC1, VEC2 = column vectors of observations

  OUTPUT: Q_AP = estimate of Q
         Q_1 = estimate using only VEC1 observations
         Q_2 = estimate using only VEC2 observations
+-----+;
run GET_H(VEC1,H_1);
run GET_H(VEC2,H_2);
n_1 = nrow(vec1);
n_2 = nrow(vec2);
eta_1 = 0;
eta_2 = 0;
* +-----+
  For each observation in VEC1, compute probability
  densities for the two sample distributions. Count the
  number of observations for which sample 1 is higher
+-----+;
do i = 1 to n_1;
  test_x = VEC1[i,1];
  run GET_F(VEC1,H_1,test_x,F_X1);
  run GET_F(VEC2,H_2,test_x,F_X2);
  if F_X1 > F_X2 then eta_1 = eta_1 + 1;
end;
* +-----+
  For each observation in VEC2, compute probability
  densities for the two sample distributions. Count the
  number of observations for which sample 2 is higher
+-----+;
do i = 1 to n_2;
  test_x = VEC2[i,1];
  run GET_F(VEC1,H_1,test_x,F_X1);
  run GET_F(VEC2,H_2,test_x,F_X2);
  if F_X2 > F_X1 then eta_2 = eta_2 + 1;
end;
* +-----+
  Compute apparent value of Q = mean number of correct
  decisions about group membership.

```

```

+-----+;
Q_AP = (eta_1 + eta_2) / (n_1 + n_2);
Q_1 = eta_1 / n_1;
Q_2 = eta_2 / n_2;
FINISH;

START Q_BOOT(VEC1,VEC2,n_boot,epsi_1,epsi_2);
* +-----+
  Compute the "bootstrap" estimator of the effect size Q
  INPUT: VEC1, VEC2 = column vectors of observations
         n_boot = n of bootstrap samples to draw
  OUTPUT: epsi_1, epsi_2 = the epsilon components of the 632 bootstrap
+-----+;
n_1 = nrow(vec1);
n_2 = nrow(vec2);
vec_1A = J(n_1,1,0)||vec1;
vec_2A = J(n_2,1,0)||vec2;
do i = 1 to n_1;
  vec_1A[i,1] = i;
end;
do i = 1 to n_2;
  vec_2A[i,1] = i;
end;

eta_b1 = J(n_1,1,0);
eta_b2 = J(n_2,1,0);
n_b1 = J(n_1,1,0);
n_b2 = J(n_2,1,0);

do i = 1 to n_boot;
  run BOOTSTRP(VEC_1A,N_1,Boot_1);
  run BOOTSTRP(VEC_2A,N_2,Boot_2);

  do j = 1 to n_1;
    in_samp = 0; * observation "j" is not in the bootstrap sample;
    do k = 1 to n_1;
      if vec_1A[j,1] = Boot_1[k,1] then in_samp = 1; * OOPS! it is!;
    end;

    if in_samp = 0 then do;
      run GET_H(Boot_1,H_1);
      run GET_H(Boot_2,H_2);

      test_x = vec_1A[j,2];
      run GET_F(Boot_1,H_1,test_X,F_X1);
      run GET_F(Boot_2,H_2,test_X,F_X2);
      if F_X1 > F_X2 then eta_b1[j,1] = eta_b1[j,1] + 1;
      n_b1[j,1] = n_b1[j,1] + 1;
    end;
  end;
do j = 1 to n_2;
  in_samp = 0; * observation "j" is not in the bootstrap sample;
  do k = 1 to n_2;
    if vec_2A[j,1] = Boot_2[k,1] then in_samp = 1; * OOPS! it is!;
  end;

  if in_samp = 0 then do;
    run GET_H(Boot_1,H_1);
    run GET_H(Boot_2,H_2);

    test_x = vec_2A[j,2];

```

```

        run GET_F(Boot_1,H_1,test_X,F_X1);
        run GET_F(Boot_2,H_2,test_X,F_X2);
        if F_X2 > F_X1 then eta_b2[j,1] = eta_b2[j,1] + 1;
        n_b2[j,1] = n_b2[j,1] + 1;
    end;
end;
end;

* +-----+
  Compute epsilon values for computation of
  632 bootstrap Q = mean number of correct decisions
  about group membership for observations not in
  bootstrap sample.
+-----+;
eta1 = 0;
do i = 1 to n_1;
    eta1 = eta1 + (eta_b1[i,1] / n_b1[i,1]);
end;
eta2 = 0;
do i = 1 to n_2;
    eta2 = eta2 + (eta_b2[i,1] / n_b2[i,1]);
end;
epsi_1 = eta1 / n_1;
epsi_2 = eta2 / n_2;
FINISH;

* +-----+
  Read data from regular SAS into PROC IML
+-----+;
use &dataset;

read all var {&dvlist} into all_dataA;
read all var {&dvlist} where (&groupvar = &value1) into dat_1A;
read all var {&dvlist} where (&groupvar = &value2) into dat_2A;

n_vars = ncol(all_dataA);
n_all = nrow(all_dataA);
ones = j(1,n_all,1);

* +-----+
  Extract SAS Names of Variables for Printed Output from Macro
+-----+;
varnames = symget('dvlist');
varnamev = J(1,n_vars,'XXXXXXXXXXXXXXXXXX');
noblank = 0;
numchars = 0;
vseq = 1;
do i = 1 to length(varnames);
    if noblank = 0 then do;
        if substr(varnames,i,1) = ' ' then noblank = 0;
        if substr(varnames,i,1) ^= ' ' then noblank = i;
    end;
    if noblank ^= 0 then do;
        if substr(varnames,i,1) = ' ' | i = length(varnames) then do;
            numchars = i - noblank;
            if i = length(varnames) then numchars = numchars + 1;
            varnamev[1,vseq] = substr(varnames,noblank,numchars);
            vseq = vseq + 1;
            noblank = 0;
        end;
    end;
end;
end;

```

```

end;

* +-----+
  Begin do-loop to analyze each variable sent to the macro
+-----+;
do variable = 1 to n_vars;
  all_dat = all_datA[,variable];
  dat_1 = dat_1A[,variable];
  dat_2 = dat_2A[,variable];

  * +-----+
    Computation of n, mean, standard deviation for each group
  +-----+;
  n_var = ncol(all_dat); n_all = nrow(all_dat);
  ones = j(1,n_all,1);
  sum = ones*all_dat;
  n_1 = nrow(dat_1); n_2 = nrow(dat_2);
  ones_1 = j(1,n_1,1); ones_2 = j(1,n_2,1);
  sum_1 = ones_1*dat_1; sum_2 = ones_2*dat_2;
  mean_1 = (1/n_1)*sum_1; mean_2 = (1/n_2)*sum_2;
  mnm_1 = ones_1`*mean_1; mnm_2 = ones_2`*mean_2;
  dev_1 = dat_1 - mnm_1; dev_2 = dat_2 - mnm_2;
  sscp_1 = dev_1`*dev_1; sscp_2 = dev_2`*dev_2;
  ss_1 = vecdiag(sscp_1); ss_2 = vecdiag(sscp_2);
  cov_matrix_1 = 1/(n_1 - 1)*sscp_1; cov_matrix_2 = 1/(n_2 - 1)*sscp_2;
  s_1 = vecdiag(cov_matrix_1); s_2 = vecdiag(cov_matrix_2);
  num_s_p = ss_1 + ss_2;
  n_all = (n_1 + n_2) -2;
  ss_rows = nrow(num_s_p);
  n_all_vctr = j(ss_rows,1,n_all);
  var_p_mx = num_s_p*(1/n_all_vctr)`;
  s_p_mx = sqrt(var_p_mx);
  s_pooled = vecdiag(s_p_mx);

  * +-----+
    Cohen d effect size
  +-----+;
  delta_mean = mean_1 - mean_2;
  d_mtrx = (1/s_pooled)*delta_mean;
  Cohen_d = vecdiag(d_mtrx);

  * +-----+
    Hedges g effect size
  +-----+;
  d_const = 1 - (3/((4*n_all)-9));
  d_const_v = j(1,n_var,d_const);
  hg_matrix = Cohen_d*d_const_v;
  Hedges_g = vecdiag(hg_matrix);

  * +-----+
    Common Language effect size
  +-----+;
  com_1 = probnorm(abs(mean_1 - mean_2)`*1/(sqrt(s_1 +s_2)));

  * +-----+
    Cliff Delta effect size
  +-----+;
  val_1 = dat_1*(1/dat_2)`;
  ones_val_1 = j(n_1,n_2,1);
  val_1_gt_2 = val_1 > ones_val_1;
  val_1_lt_2 = val_1 < ones_val_1;

```

```

diff_1 = val_1_gt_2 - val_1_lt_2;
d_mtrx = ones_1*diff_1;
sum_d_mtrx = sum(d_mtrx`);
delta_index = sum_d_mtrx*(1/(n_1*n_2));

* +-----+
  Stochastic Superiority effect size
+-----+;
Stochastic_A = (delta_index + 1)/2;

* +-----+
  Trimmed d effect size
+-----+;
run trimmit(dat_1,10,t_count1,T_mean1,W_var1);
run trimmit(dat_2,10,t_count2,T_mean2,W_var2);

trim_es = (T_mean1 - T_mean2)/
          Sqrt((W_var1*(n_1 - 2*t_count1) + W_var2*(n_2 - 2*t_count2))/
              ((n_1 - 2*t_count1) + (n_2 - 2*t_count2)));

* +-----+
  Gamma effect size
+-----+;
run bubble(dat_1,n_1,1);
run bubble(dat_2,n_2,1);

if 0.5*n_1 = round(0.5*n_1) then even=1;
if 0.5*n_1 ^= round(0.5*n_1) then even=0;

if even=0 then do;
  median1 = dat_1[0.5*n_1 + 0.5,1];
end;
if even=1 then do;
  median1= 0.5*(dat_1[(0.5*n_1),1] + dat_1[(0.5*n_1 + 1),1]);
end;

countlss=0;
do g = 1 to n_2;
  if dat_2[g,] < median1 then countlss = countlss + 1;
end;
if (countlss > 0 & countlss < n_2) then gamma1 = probit(countlss/n_2);
if countlss = 0 then gamma1 = probit(1/(n_2+1));
if countlss = n_2 then gamma1 = probit(countlss/(n_2+1));
* Get Group 2 median just for reporting;
if 0.5*n_2 = round(0.5*n_2) then even=1;
if 0.5*n_2 ^= round(0.5*n_2) then even=0;

if even=0 then do;
  median2 = dat_2[0.5*n_2 + 0.5,1];
end;
if even=1 then do;
  median2= 0.5*(dat_2[(0.5*n_2),1] + dat_2[(0.5*n_2 + 1),1]);
end;

* +-----+
  Wilcox & Muska W effect size
+-----+;
run GET_QAP(dat_1,dat_2,Q_APS,Q_1S,Q_2S);
run Q_BOOT(dat_1,dat_2,1000,epsi_1s,epsi_2s);
Q632X = .368#Q_1S + .632#epsi_1s;
Q632Y = .368#Q_2S + .632#epsi_2s;

```

```

Q632 = (n_1#Q632X + n_2#Q632Y) / (n_1+n_2);

* +-----+
  Assemble effect sizes and descriptive statistics
  into vectors for printed output
+-----+;
if variable = 1 then do;
  d_vec = Cohen_d;
  g_vec = Hedges_g;
  i_vec = delta_index;
  c_vec = com_1;
  t_vec = trim_es;
  gm_vec = gamma1;
  Q1_vec = Q_APS;
  Q2_vec = Q632;
  SA_vec = Stochastic_A;
  nn_vec1 = n_1;
  nn_vec2 = n_2;
  mn_vec1 = mean_1;
  mn_vec2 = mean_2;
  sd_vec1 = sqrt(s_1);
  sd_vec2 = sqrt(s_2);
  md_vec1 = median1;
  md_vec2 = median2;
  wv_vec1 = w_var1;
  wv_vec2 = w_var2;
end;

if variable > 1 then do;
  d_vec = d_vec||Cohen_d;
  g_vec = g_vec||Hedges_g;
  i_vec = i_vec||delta_index;
  c_vec = c_vec||com_1;
  t_vec = t_vec||trim_es;
  gm_vec = gm_vec|gamma1;
  Q1_vec = q1_vec|Q_APS;
  Q2_vec = Q2_vec|Q632;
  SA_vec = SA_vec|Stochastic_A;
  nn_vec1 = nn_vec1|n_1;
  nn_vec2 = nn_vec2|n_2;
  mn_vec1 = mn_vec1|mean_1;
  mn_vec2 = mn_vec2|mean_2;
  sd_vec1 = sd_vec1|sqrt(s_1);
  sd_vec2 = sd_vec2|sqrt(s_2);
  md_vec1 = md_vec1|median1;
  md_vec2 = md_vec2|median2;
  wv_vec1 = wv_vec1|w_var1;
  wv_vec2 = wv_vec2|w_var2;
end;
end;

* +-----+
  Printed output from macro
+-----+;
file print;
put @1 'Descriptive Statistics:'//
@33 'Group 1' @83 'Group 2'/
@14 '-----'
@64 '-----'
@38 'Standard Winsorized' @87 'Standard Winsorized'/
@1 'Variable      N      Mean      Median Deviation Variance'

```

```

@66 'N      Mean      Median      Deviation      Variance' /
@1 '-----' /;
-----' /;
do i = 1 to n_vars;
  pname = varnamev[1,i];
  n1 = nn_vec1[1,i];
  n2 = nn_vec2[1,i];
  mn1 = mn_vec1[1,i];
  mn2 = mn_vec2[1,i];
  sd1 = sd_vec1[1,i];
  sd2 = sd_vec2[1,i];
  md1 = md_vec1[1,i];
  md2 = md_vec2[1,i];
  wv1 = wv_vec1[1,i];
  wv2 = wv_vec2[1,i];
  file print;
  put @1 pname @14 n1 best5. @22 mn1 best5. @30 md1 best5. @39 sd1 best5. @51 wv1 best5.
    @64 n2 best5. @71 mn2 best5. @79 md2 best5. @88 sd2 best5. @100 wv2 best5. /;
end;
file print;
put @1 '-----' /;
-----' /;

@1 'Effect Sizes:' // @99 'W' /
@1 'Variable      Cohen      Hedges      Trimmed      Common      Stochastic      Cliff      --
-----' /
@17 'd          g          d          Gamma      Language      Superiority (A)      Delta      Naive      .632' /
@1 '-----' /;
-----' /;
do i = 1 to n_vars;
  pname = varnamev[1,i];
  dd = d_vec[1,i];
  gg = g_vec[1,i];
  ii = i_vec[1,i];
  cc = c_vec[1,i];
  tt = t_vec[1,i];
  gm = gm_vec[1,i];
  Q1 = Q1_vec[1,i];
  Q2 = Q2_vec[1,i];
  AA = SA_vec[1,i];
  file print;
  put @1 pname @14 dd best5. @24 gg best5. @35 tt best5. @45 gm best5. @54 cc best5.
    @68 AA best5. @83 ii best5. @92 Q1 best5. @101 Q2 best5. /;
end;
file print;
put @1 '-----' /;
-----' /;
quit;

%mend ROBUST_ES;

```

## EXAMPLE OF MACRO ROBUST\_ES

The easiest way in which the macro ROBUST\_ES may be used is to simply create a SAS dataset that inputs the sample data. The macro is then called, using as arguments the name of the dataset and the names of the relevant variables. Summary data from 20 observations on three variables are used to illustrate the macro. The observed data are read into the SAS dataset SCORES.

```

data scores;
  input SEX GREQ GREV AGE;

```

```

cards;
1      800      310      37
1      670      280      32
1      680      570      31
1      460      570      40
1      590      610      28
1      790      600      32
1      780      620      27
1      780      690      34
1      710      530      38
1      660      480      28
2      440      530      39
2      360      430      50
2      670      740      40
2      690      550      51
2      460      440      37
2      280      330      47
2      610      650      32
2      490      510      41
2      670      690      44
2      630      780      38
;

```

The following call to the macro identifies the SAS dataset to be used for analysis (SCORES), the names of the dependent variables (GREQ, GREV, and AGE) and the name and values of the grouping variable (SEX).

```

%robust_es (DATASET = SCORES, DVLIST = GREQ GREV AGE, GROUPVAR = SEX, VALUE1 = 1, VALUE2 = 2);
run;

```

### OUTPUT FROM MACRO ROBUST\_ES

Table 1 provides an example of the tabled output produced by the macro ROBUST\_ES. The top section of the output provides a selection of basic descriptive statistics for each variable in each group (sample size, mean, median, standard deviation, and Winsorized variance). Because three variables were included in the DVLIST = argument when the macro was called, three rows of descriptive statistics are provided.

The bottom section of the output provides the effect sizes for each of the dependent variables. The standardized mean differences (Cohen's *d* and Hedges' *g*) are provided first, followed by the *trimmed-d* effect size based on trimmed means and Winsorized variances. These three effect sizes estimate the difference between the two population means in standard deviation units. The remaining indices of effect size ( $\gamma_1^*$ , *CL*, *A*, *delta*, and the two estimators of *W*) estimate the degree of overlap between the two population distributions, based upon the degree of overlap in the sample distributions. The  $\gamma_1^*$  and *CL* effect sizes use the cumulative normal distribution function in deriving their values, while the remaining indices use simple functions of the observed data. Because of differences in scale of these effect size indices, the reader is cautioned that comparisons of values across the indices are not meaningful. That is, a Cohen's *d* value of 1.276 cannot be meaningfully compared with a Cliff's *delta* value of 0.65 because the scales are not equivalent.

Table 1  
*Sample Output from Macro.*

Descriptive Statistics:

Variable	Group 1					Group 2				
	N	Mean	Median	Standard Deviation	Winsorized Variance	N	Mean	Median	Standard Deviation	Winsorized Variance
GREQ	10	692	695	107	17433	10	530	550	144.2	26360
GREV	10	526	570	133.9	4193	10	565	540	146.5	29793
AGE	10	32.7	32	4.498	19.93	10	41.9	40.5	6.045	39.33

Effect Sizes:

Variable	Cohen d	Hedges g	Trimmed d	Gamma	Common Language	Stochastic Superiority (A)	Cliff Delta	W	
								Naïve	.632
GREQ	1.276	1.215	0.991	1.335	0.817	0.825	0.65	0.7	0.258
GREV	-0.28	-0.26	0.563	0.253	0.578	0.45	-0.1	0.75	0.276
AGE	-1.73	-1.64	-1.71	-1.34	0.889	0.105	-0.79	0.8	0.294

## CONCLUSION

The use of effect sizes has grown in popularity in recent years (although such application remains far from universal). Renewed debates regarding the over-reliance on hypothesis testing, emphasizing the often misleading nature and inappropriate use of such tests (Nickerson, 2000), may be partially responsible for this increased interest. Because the use of effect sizes, in many instances, provides useful information to supplement traditional inferential statistics, advocacy for their use is appropriate. As the reporting and interpretation of effect sizes become more commonplace, researchers must remain mindful of the limitations of certain indices. For example, Wilcox and Muska (1999) have pointed out the important distinction among indices that reflect differences in location and those that represent more global differences in distributions. Further, Fern and Monroe (1996) have delineated the variety of research factors (e.g., designs, operational details, measurement reliability, sample characteristics) that must be considered in the appropriate interpretation of observed effect sizes.

The macro ROBUST\_ES is provided to facilitate researchers' calculation and use of both traditional and robust effect size indices. Although the macro, as provided, is limited to the case of two independent groups, the SAS/IML code can be easily modified to provide effect size estimates for other data structures (e.g., correlated observations, more than two groups).

## REFERENCES

- American Psychological Association (2001). *Publication manual of the American Psychological Association* (5<sup>th</sup> ed.). Washington, DC: Author.
- Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61, 230-258.
- Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114, 494-509.
- Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, 31, 331-350.
- Crow, E. L. (1991). Response to Rosenthal's comment "How are we doing in soft psychology?" *American Psychologist*, 46, 1083.
- Fern, E. F. & Monroe, K. B. (1996). Effect size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, 23, 89-105.
- Gorsuch, R. L. (1991). Things learned from another perspective (so far). *American Psychologist*, 53, 800-801.
- Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.
- Hess, M. & Kromrey, J.D. (2003, February). *Confidence Bands for Standardized Mean Differences: A Comparison of Nine Techniques Under Non-normality and Variance Heterogeneity*. Paper presented at the Eastern Educational Research Association, Hilton Head, NC.
- Hess, M. & Kromrey, J. D. (2004, April). *Robust Confidence Intervals for Effect Sizes: A Comparative Study of Cohen's d and Cliff's Delta Under Non-normality and Heterogeneous Variances*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.
- Hogarty, K. Y. & Kromrey, J. D. (1999, August). *Traditional and Robust Effect Size Estimates: Power and Type I Error Control in Meta-Analytic Tests of Homogeneity*. Paper presented at the annual meeting of the American Statistical Association, Baltimore.
- Hogarty, K. Y. & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the annual meeting of the American Educational Research Association, Seattle.
- Kirk, R. E. (1996). Practical Significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.
- Kraemer, H. C. & Andrews, G. A. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin*, 91, 404-412.
- Kromrey, J. D. & Hess, M. (2002, April). *Interval Estimates of Effect Size: An Empirical Comparison of Methods for Constructing Confidence Bands Around Standardized Mean Differences*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

- McGraw, K. O. (1991). Problems with BESD: A comment on Rosenthal's "How are we doing in soft psychology?" *American Psychologist*, 46, 1084-1086.
- McGraw, K. O. & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111, 361-365.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241-301.
- Nix, T.W. & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, 5(2), p. 3-14.
- Parker, S. (1995). The "difference of means" may not be the "effect size." *American Psychologist*, 50, 1101-1102.
- Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. *American Psychologist*, 46, 1086-1087.
- Strahan, R. F. (1991). Remarks on the binomial effect size display. *American Psychologist*, 46, 1083-84.
- Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools*, 5, 33-38.
- Thompson, B. & Snyder, P.A. (1997). Statistical significance testing practices in the *Journal of Experimental Education*. *Journal of Experimental Education*, 66, 75-83.
- Thompson, B. & Snyder, P.A. (1998). Statistical significance and reliability analyses in recent JCD research articles. *Journal of Counseling and Development*, 76, 436-441.
- Vargha, A. & Delaney, H.D. (2000). A critique and improvement of the CL Common Language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25, 101-132.
- Wilcox, R. R. & Muska, J. (1999). Measuring effect size: A non-parametric analogue of  $\omega^2$ . *British Journal of Mathematical and Statistical Psychology*, 52, 93-110.
- Wilkinson & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.
- Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, 61, 165-170.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the first author at:

Jeffrey D. Kromrey  
University of South Florida  
4202 East Fowler Ave., EDU 162  
Tampa, FL 33620  
Work Phone: (813) 974-5739  
Fax: (813) 974-4495  
Email: [kromrey@tempest.coedu.usf.edu](mailto:kromrey@tempest.coedu.usf.edu)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.