Paper 43-2010

Two EDF Graphs for Assessing Agreement between Paired Data

David W. Meek, USDA-ARS-MWA NLAE, Ames, IA 50011

ABSTRACT

This paper is intended for those familiar with PROC GPLOT from SAS/GRAPH®[1] and PROCs MEANS, TTEST, and NPAR1WAY in SAS/STAT®. Via macro procedures, two graphs related to the Smirnoff D statistic are developed to aid in the assessment of agreement between paired data. The first is a PROC GPLOT version of the ODS EDF plot in PROC NPAR1WAY; the second is a variation of it. A dataset from a SAS manual is used to exhibit the graphs. Each macro is a stand-alone example; each is included in its own appendix along with a data step that includes the dataset followed by the macro call statement.

INTRODUCTION

In many different fields including science and engineering, a researcher often has to assess agreement between two variables of interest in which the observations are logically paired. For example, a researcher may want to know if a new or less expensive instrument can adequately reproduce the measures from a standard or more expensive instrument for a specified set of values over an operational range. Alternatively, a researcher may want to know if a given model adequately estimates an independent set of observations. Meek (2007) reviews some common graphics and measures for such assessments including a version of Bland and Altman's (1986) mean-difference plot for assessing the mean bias error, a version of a bivariate plot, Berg's (1992) combination of the former two, and Lin's (1989) concordance correlation coefficient. When there is reasonable agreement by various measures between two variables with such paired observations, generally both of the empirical distribution functions (EDFs) are similar. The Smirnov test statistic for two samples of equal size (D) is a common nonparametric test used to evaluate such a similarity or difference in the two EDFs (see e.g., Section 6.3, pp 456-462 and Appendix Table A19, p. 556 in Conover, 1999). The D statistic has unrestrictive distributional assumptions and can detect differences in both variability and central tendency. A graph of the EDF comparison can be very insightful, revealing both the magnitude and patterns of the two EDFs over the combined range of values for the paired observations.

In the SAS 9.2 (TS Level 2M0) version of PROC NPAR1WAY (SAS Inst., 2008), an EDF graph is produced when the ODS options are turned on. The default ODS graph, however, is limited in several ways. When the graph is displayed on the monitor (a.k.a. *CRT*), the resolution is limited to that set on the computer's display. In addition, altered data used to generate the graph are not available as an output set and the features of graph are *as is* unless you go into an editor. This paper presents two alternative graphs generated by SAS® macros. Each macro can produce either a CRT graph or a high resolution (300 dpi) png file. The first macro generates a grayscale variant of the NPAR1WAY ODS graph. The second macro generates a plot of the actual difference between the two EDFs over the range of values for the paired observations. In either case, the code can be altered to capture the EDF data or change the presentation features in the graph. Some ideas used to develop the graphs came from both Cleveland (1993) and Tufte (1983).

THE DATA SET

For simplicity, a small published data set from a SAS/STAT® manual is selected. Here, the EDFs for posttest and pretest paired test scores for a group of 15 individuals are to be examined. The data set is from Example 2, pp. 956-947 in the PROC TTEST documentation (SAS Inst., 1988). Here the posttest/pretest paired differences are examined with PROC MEANS to see if the mean difference is equal to zero. It is not! The mean difference is 7.93±2.56 (Pr<0.0079). If the T-test was not previously run, this bias would not be known. An EDF graph, however, could reveal it along with other noteworthy features over the range of differences.

THE ODS EDF GRAPHIC WITH PROC NPAR1WAY

Now consider the default plot produced by PROC NPAR1WAY for the posttest/pretest data. In this and the following examples, every pair with a missing value is to be excluded (note code) but here there are not any. The SAS code that produces an EDF plot is as follows:

```
options ps=50 ls=78 nodate notes nonumber;*ODS_EDFgraph_ex.sas;
title1 "SAS Proc T-Test Paired Data Example";
data a0;
input id pretest posttest @@;
   dif = posttest-pretest;if dif ne .;*exclude missing values;
lines;
01 80 82 02 73 71 03 70 95 04 60 69 05 88 100 06 84 71 07 65 75 08 37 60
09 91 95 10 98 99 11 52 65 12 78 83 13 40 60 14 79 86 15 59 62
; run;
```

[1] The mention of a trade name is for informational purposes only and is not an endorsement by the USDA-ARS.

```
/* Rearrange the data. */
data a01; set a0;  test = pretest;  TYPE = "Pretest"; keep type test; run;
data a02; set a0;  test = posttest;  TYPE = "Posttest";  keep type test;  run;
data a0s;  set a01 a02;  run; quit;

/* Run the analysis and produce the ODS plot. */
ods graphics on;  proc npar1way anova d edf data=a0s plots=edfplot;  class
type;  var test;  run;  ods graphics off; quit;
```

The resultant graph shows the two EDFs over the combined range of test scores; notice that the maximum EDF difference is 0.267 at test = 59, while the maximum test score difference is 23 and occurs at EDF = 0 (Figure1, below). Also notice the EDF increments are in steps of 1/15. While the two-sided D statistic (denoted KSa in the graph) is insignificant (Pr > 0.66), the plot is revealing. Recall that based on the reported T-test, on average, the posttest scores are significantly higher than the pretest scores. This difference as shown in Figure 1, however, is not constant over the range of test scores; rather it tends to be larger near low values and closer near the center. Such insight may be of considerable interest to a researcher.

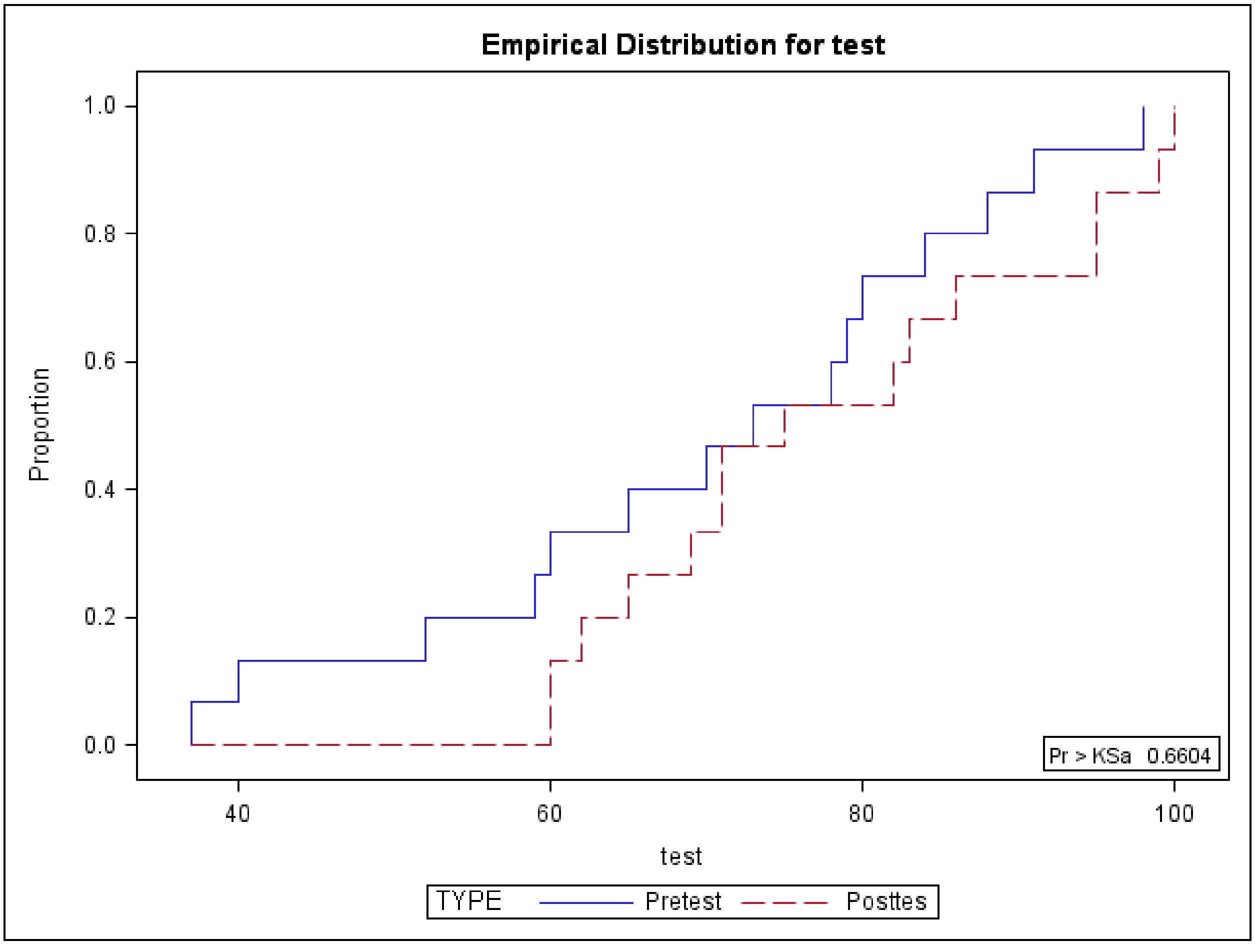

Figure 1. The default ODS graph produced in PROC NPAR1WAY.

THE FIRST MACRO

Both macros in this paper are organized in the form of a complete stand-alone example. Each starts with the macro code, followed by the data step (the same one as shown in ODS example for the EDF graph). A macro call statement then completes each one.

As an alternative to the ODS graph, the first macro produces a gray-scale version of the plot shown in Figure 1. The SAS® code, however, is open and can be changed. Users, hence, may modify the code to suit the needs and requirements of each given analysis. The dataset name used in the plot is *ma7*. In addition, both a CRT plot and a 300 dpi png file option are possible outputs. Using the same data set from the ODS example, the resulting graph from the macro's 300 dpi png file option is shown in Figure 2 (below). The SAS® code used to create the graph is

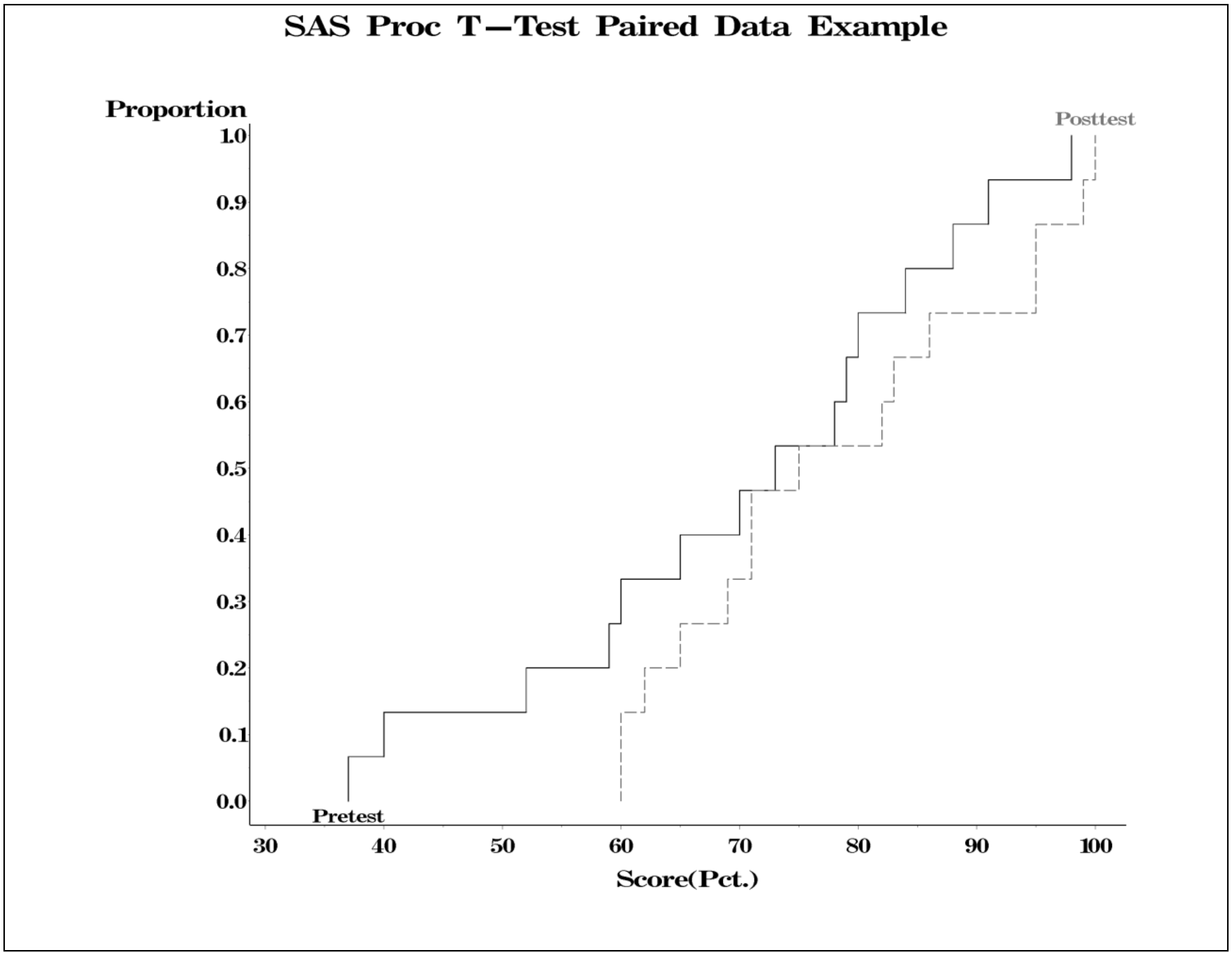

Figure 2. The EDF plot produced by the first macro.

listed in the appendix under the title ***MACRO 1: EDF GRAPH FOR PAIRED OBSERVATIONS****.* The png file option call statement used, however, is as follows:

> %mks1(a0, pretest, posttest, namex="Pretest", namey="Posttest", xtitle="Score(Pct.)", gtitle="SAS Proc T-Test Paired Data Example", gtype=0);

It is the same as in the second commented out call statement. By default the filename with path location is defined by the *fname="C:\TEMP\edfplot1.png*“ statement in the top line of the macro. Users will need to define their own path and file name. If desired, the file type may be changed and some features may be modified in many graphics editors.

THE SECOND MACRO

As previously stated, this example is organized in the same format as the first. The SAS code used to create the graph is listed in the appendix under the title ***MACRO 2: EDF DIFFERENCE GRAPH FOR PAIRED OBSERVATIONS***. Again, the data and SAS® code are open and may be user modified. The dataset name used in the plot is *ma10.* Both a CRT plot and a 300 dpi png file option are, once again, alternative outputs as well. The png file option call statement used is as follows:

> %mks2(a0, pretest, posttest, xtitle="Score (Pct.)", gtitle="The Difference in EDF's", dref=0.25, gtype=0);

This macro is offered as another alternative to the ODS graph. This one plots the EDF difference over the range of scores. Conceptually, the graph is a residual plot for Figure 1 or 2. The resulting plot generated by the macro is shown in Figure 3 below.

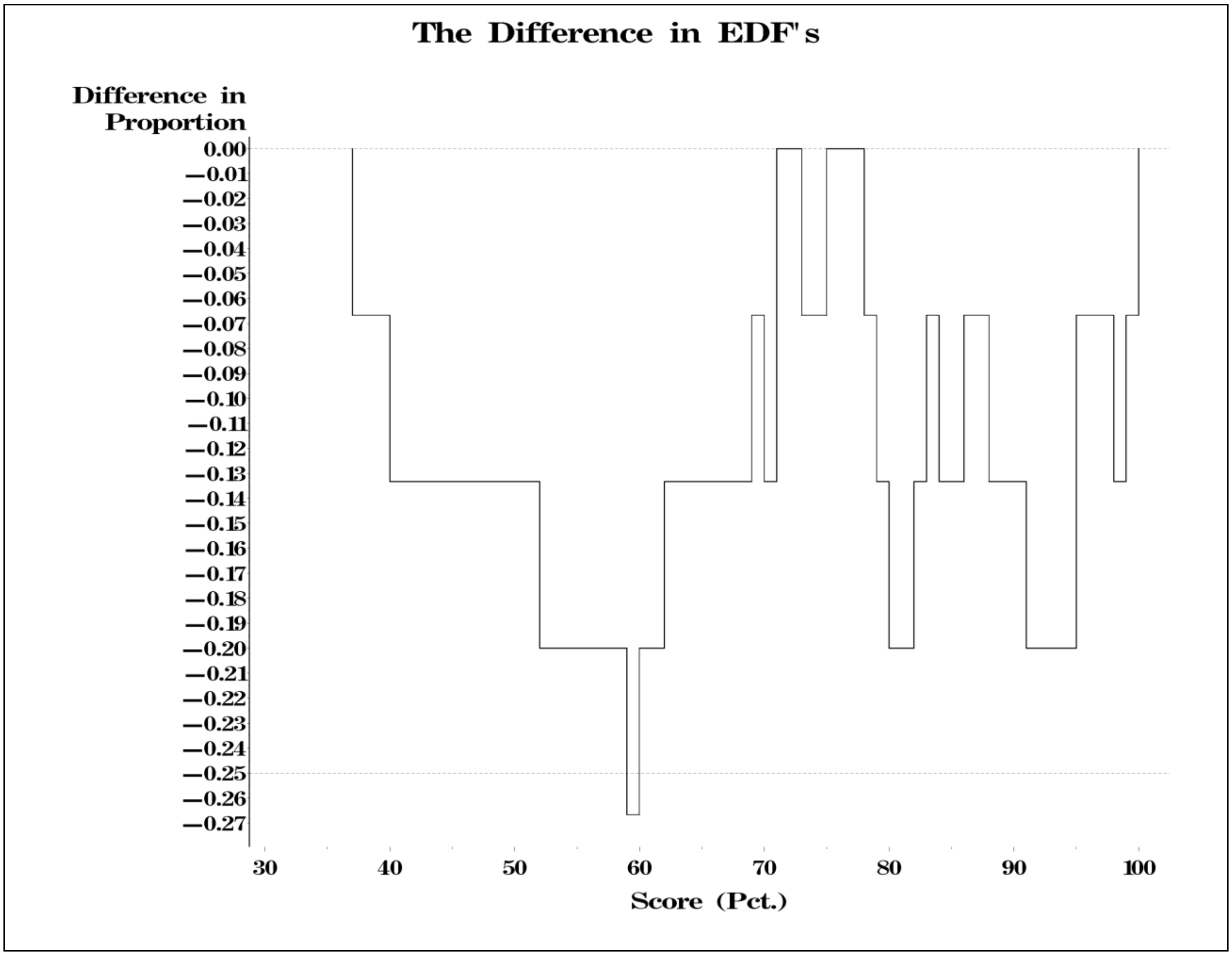

Figure 3. The EDF difference plot produced by the second macro.

The default file name with path location is fname="C:\TEMP\dedfplot1.png". Notice that one can readily observe some salient features. All differences are ≤0 and consequently show an overall systematic difference between the EDFs! Due to the discrete nature of the EDFs there are modes for the most common differences – here 0,-1/15, -2/15, and -3/15. Finally, for 59 ≤ score ≤ 60 the input reference difference level is exceeded. Such characteristics may be of considerable interest to a given analyst in a particular study. Know that if the maximum absolute difference is less than the input reference value (*dref*), it may not appear in the plot. In addition, notice the input reference difference (*dref = 0.25*) is not from a table of D values but arbitrarily set to a quartile – here, a difference of possible practical interest.

DISCUSSION AND CONCLUSIONS

To help researchers at the author's institution, these macros were developed to aid in the assessment of agreement between paired observations from many different kinds of research. As with these researchers, each new user has the responsibility for the proper assessment and use of the macros. Also each new user should have a thorough knowledge of each and every data set to be examined. In turn, these graphs should help an analyst gain insight into the distributional similarities or differences between paired observations. No problems have been observed in any tested data sets. The code may be user modified and the plotted values obtained to be examined or saved. Line types and colors or labels can be readily changed. Although this author prefers analysis results like Pr or D values in the figure caption, with some extra coding they could be inserted in the graph.

REFERENCES

Berg, R.L. 1992. First Place Best Presentation of Data-Monochrome, P. 1521-1527. Sugi 17 - Proceedings of the Seventeenth Annual SAS Users ® Group International Conference. Apr. 12-15, 1992. Honolulu, HI.

Bland, J.M. and Altman, D.G. 1986. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet I: 307-310.

Cleveland, W.S. 1993. Visualizing Data. Hobart Press, Summit, NJ., pp. 360.

Conover, W.J. 1999. Practical Nonparametric Statistics. J. Wiley & Sons, New York.

Lin, L.I. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 45: 255-268.

Meek, D. 2007. Two Macros for Producing Graphs to Assess Agreement between Two Variables, Paper D6. [CD-ROM] *in* A. Katschke (ed.) Data Visualization and Graphics Section, 2007 MidWest SAS Users Group Annual Conference Proc., Des Moines, IA. 28–30 Oct. 2007. SAS Inst., Cary, NC.

SAS Inst. 1988. Ch. 33, The TTEST Procedure, pp. 941-947. In: SAS/STAT® User's Guide, Release 6.03 Ed. SAS Inst. Inc., Cary, NC.

SAS Inst. 2008. The NPAR1WAY Procedure, SAS/STAT® 9.2 User's Guide, SAS Inst. Inc., Cary, NC. (The download link: http://support.sas.com/documentation/cdl/en/statugnpar1way/61813/PDF/default/statugnpar1way.pdf).

Tufte, E.R. 1983. The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT. 197 pp.

ACKNOWLEDGMENTS

This work was supported by the USDA-ARS National Laboratory for Agriculture and the Environment, Ames, IA, Dr. J.L. Hatfield, Director. The author thanks the following colleagues for their comments: Dr. R.P. Ewing, Iowa State University, Ames, IA; Dr. B.J. Fergen, USDA-APHIS, Ames, IA; and Dr. J. Singer, USDA-ARS NLAE, Ames, IA.

CONTACT INFORMATION

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

Name: David Meek
Enterprise: USDA-ARS-MWA NLAE
2110 University Blvd.
Ames, IA 50011
Work phone: (515) 294-2246
Fax: (515) 294-8125
Email: Dave.Meek@ars.usda.gov
Web: None

MACRO 1: EDF GRAPH FOR PAIRED OBSERVATIONS

```
/* SAS MACRO. */
%macro mks1(datafile,xvar,yvar,namex="x",namey="y",xtitle="X",gtitle=" ",
            gtype=1, fname="C:\TEMP\edfplot1.png");*mks1.sas;


/*    This macro generates a gplot graph for paired observations.  It shows
   the Empirical Distribution Functions (EDF) for two variables over the range
   of the common values.  It is commonly shown with a Smirnov D test statistic
   for comparing the distributions of two samples of equal sample size.  The
   required inputs are as follows:
      datafile - The name of the data file with the paired observations.
          xvar - The name the independent or standard variable.
          yvar - The name the dependent variable.
   The optional inputs are as follows:
         xname - The user input name for the x variable's EDF (default = "x").
         yname - The user input name for the y variable's EDF (default = "y").
        xtitle - The axis label for the independent variable (default: xtitle="x").
        gtitle - The graph title (default: gtitle=" ").
         gtype - Insert 1 for MONITOR (CRT)or 0 for 300 dpi png file.
         fname - Insert your own file name and path.
   This macro was written by David Meek, USDA-ARS-MWA
   National Laboratory for Agriculture and the Environment
   2110 University Blvd., Ames, IA,
   50011. TEL: 515-294-2246; Email: Dave.Meek@ars.usda.gov.                       */

/* This part below, up to the graphics routines, generates the edfs and rearranges
   the data so they can be plotted. */
```

```
data _null_; * Count the number of paired observations;
 if 0 then set &datafile point=_N_ nobs=nraw;
 call symput('nraw', left(put(nraw,8.))); stop;
run; quit;
proc sort data=&datafile; by &yvar; run; quit;
data ma1; set &datafile;
  edf = (_N_-1)/&nraw;
 keep edf &yvar;
run; quit;
data ma1a; set &datafile;
  edf = (_N_-1)/&nraw;
   y1 = lag1(&yvar);
 keep edf y1;
 rename y1=&yvar;
run; quit;
proc sort data=&datafile; by &xvar; run; quit;
data ma2; set &datafile;
 edf = (_N_-1)/&nraw;
 keep edf &xvar;
run; quit;
data ma2a; set &datafile;
 edf = (_N_-1)/&nraw;
  x1 = lag(&xvar);
 keep edf x1;
 rename x1=&xvar;
run; quit;
data ma3; set ma1 ma1a; run; quit;
proc sort data=ma3; by &yvar edf; run; quit;
data ma4; set ma2 ma2a; run; quit;
proc sort data=ma4; by &xvar edf; run; quit;
data ma5; merge ma3 ma4; by edf; call symput('lastobs', 2*&nraw); run; quit;
data ma6; set ma5; if _N_ = &lastobs; edf=1; run; quit;
data ma7; set ma5 ma6; if ((&xvar=.) and (&yvar=.)) then delete; run; quit;

/* The part below generates the EDF Comparison Graph on the MONITOR (or makes the
   file). */

proc means data=&datafile noprint; var &xvar &yvar;
output out = mcmax min = xmin ymin max=xmax ymax;
run; quit; * Find label positions;
data mlabels; set mcmax;
%annomac;
%dclanno;
%system(2,2,2);
length text $ 32;
when ="A";
%move(xmin,0);
%label(xmin,0,left(trim(&namex)),black ,0,0,0.0250,centb,E);
%move(ymax,1);
%label(ymax,1,left(trim(&namey)),gray77,0,0,0.0250,centb,2);
run; quit;
%if &gtype=1 %then %do;
  goptions reset=all gunit=pct ftext=centb htext=3 aspect=1 cback=white
rotate=landscape
  device=win noprompt autofeed ftext=centb hsize=27.94 cm vsize=21.59 cm target=win
;*colors=?;
%end;
%else %do;
  filename tout &fname;
  goptions reset=all gunit=pct ftext=centb htext=3 aspect=1 cback=white
rotate=landscape
  device=zpng gsfname=tout gsfmode=replace noprompt autofeed ftext=centb xmax=27.94 cm
  xpixels=3300  ymax=21.59 cm  ypixels=2550; * colors=?;
  /* If your SAS version is pre 9.2.2 then use device=png. */
%end;

proc gplot data=ma7; /* Produce the Plot. */
```

```
title1 j=c h=0.5 f=centb c=black
            h=0.5 f=centb a = 90
            h=0.5 f=centb a =-90;
  footnote1 j=l h=0.5 f=centb
            m=(9,+0)
            m=(9,-1.12);
title2 j=c h=2.75 f=centb c=black &gtitle;
    symbol1 c=black  l=1 h=2.00 w=2.00  v=none  I=join;
    symbol2 c=gray66 l=4 h=2.00 w=2.00  v=none  I=join;
   axis1 label=(R=0 a=0 c=black f=centb h=2.25 j=r "Proportion")
         value=(c=black f=centb h=2.00)
         order=(0 to 1 by 0.1)
         minor=(n=1)
         width=2
         length=16.0 cm
         offset=(2.50 pct, 1.25 pct);
   axis2 label=(c=black f=centb h=2.25 &xtitle)
         value=(c=black f=centb h=2.00)
         minor=(n=1)
         width=2
         length=20.0 cm
         offset=(1.25 pct, 2.50 pct);
   plot edf*&xvar=1 edf*&yvar=2/ overlay noframe vaxis=axis1 haxis=axis2
annotate=mlabels;
run; quit;
%mend mks1;

/* Data Step. */
options ps=50 ls=78 nodate notes nonumber;  *edf_plot_ex.sas;
title1 "SAS Proc T-Test Paired Comparison Example";
data a0;
input id pretest posttest @@;
 dif = posttest-pretest; if dif ne .;*exclude missing values;
lines;
01 80 82 02 73 71 03 70 95 04 60 69 05 88 100 06 84 71 07 65 75 08 37 60
09 91 95 10 98 99 11 52 65 12 78 83 13 40 60 14 79 86 15 59 62
; run;

/* Macro call. */
%mks1(a0,pretest,posttest,namex="Pretest",namey="Posttest",xtitle="Score
(Pct.)",gtitle="SAS Proc T-Test Paired Data Example"); quit; * This call produces a
CRT Plot.;

/* %mks1(a0,pretest,posttest,namex="Pretest",namey="Posttest",xtitle="Score (Pct.)",
gtitle="SAS Proc T-Test Paired Data Example",gtype=0); quit;* This call produces a 300
dpi png file. */
```

MACRO 2: EDF DIFFERENCE GRAPH FOR PAIRED OBSERVATIONS

```
/* SAS MACRO. */
%macro mks2(datafile, xvar, yvar, xtitle="X", gtitle=" ", dref=1, gtype=1,
fname="C:\temp\dedfplot1.png");*mks2.sas;

/*      This macro generates a gplot graph for paired observations of the
   difference of the Empirical Distribution Functions (EDF) for two variables
   over the range of the common values. It is related to the EDF plot that is
   commonly shown with a Smirnov D test statistic for comparing the distributions
   of two sets of equal sample size.

   The required inputs are as follows:

      datafile - The name of the data file with the paired observations.
          xvar - The name the independent or standard variable.
          yvar - The name the dependent variable.

   The optional inputs are as follows:
        xtitle - The axis label for the independent variable (default: xtitle="x").
        gtitle - The graph title (default: gtitle=" ").
          dref - Input value for the Kolmogorov-Smirnov D statistic (default=1).
         gtype - Insert 1 for MONITOR (CRT) or 0 for 300 dpi png file.
         fname - Insert your own file name and path.

   This macro was written by David Meek, USDA-ARS-MWA
   National Laboratory for Agriculture and the Environment
   2110 University Blvd., Ames, IA,
   50011. TEL: 515-294-2246; Email: Dave.Meek@ars.usda.gov.                         */

/* This part below, up to the graphics routines, generates the edfs and rearranges the
   data so they can be plotted. */

data _null_; * Count the number of paired observations;
 if 0 then set &datafile point=_N_ nobs=nraw;
 call symput('nraw', left(put(nraw,8.))); stop; run; quit;
proc sort data=&datafile; by &yvar; run; quit;

proc sort data=&datafile; by &yvar; run; quit;
data ma1; set &datafile;
  edf = (_N_-1)/&nraw;
 keep edf &yvar;
run; quit;
data ma1a; set &datafile;
  edf = (_N_-1)/&nraw;
  y1 = lag1(&yvar);
 keep edf y1;
 rename y1=&yvar;
run; quit;
proc sort data=&datafile; by &xvar; run; quit;
data ma2; set &datafile;
  edf = (_N_-1)/&nraw;
 keep edf &xvar;
run; quit;
data ma2a; set &datafile;
  edf = (_N_-1)/&nraw;
  x1 = lag(&xvar);
 keep edf x1;
rename x1=&xvar;
run; quit;
data ma3; set ma1 ma1a; run; quit;
proc sort data=ma3; by &yvar edf; run; quit;
data ma4; set ma2 ma2a; run; quit;
proc sort data=ma4; by &xvar edf; run; quit;
data ma5; merge ma3 ma4; by edf; call symput('lastobs', 2*&nraw); run; quit;
data ma6; set ma5; if _N_ = &lastobs; edf=1; run; quit;
data ma7; set ma5 ma6; * if ((&xvar=.) and (&yvar=.)) then delete; run; quit;
data ma8; set ma7; rename edf=edfx &xvar=xm; keep edf &xvar; run; quit;
data ma9; set ma7; rename edf=edfy &yvar=xm; keep edf &yvar; run; quit;
data ma10; merge ma8 ma9; by xm; retain lastx . lasty .;
 if _N_ = 1 and lasty = . then lasty = 0;
 if _N_ = 1 and lastx = . then lastx = 0;
 if edfy = . then edfy = lasty; else lasty = edfy;
 if edfx = . then edfx = lastx; else lastx = edfx;
 edf_dif = edfy - edfx;
 uref = &dref;
 lref = -uref;
 call symput('dlb', uref);
 call symput('dub', lref);
 drop uref lref lasty lastx edfy edfx;
run; quit;

/* The part below generates the EDF Difference Graph on the MONITOR (or makes the
   file). */
```

```
%if &gtype=1 %then %do;
  goptions reset=all gunit=pct ftext=centb htext=3 aspect=1 cback=white
   rotate=landscape device=win noprompt autofeed ftext=centb hsize=27.94 cm
   vsize=21.59 cm target=win;* colors=?;
%end;
%else %do;
  filename tout &fname;
  goptions reset=all gunit=pct ftext=centb htext=3 aspect=1 cback=white
   rotate=landscape device=zpng gsfname=tout gsfmode=replace noprompt autofeed
   ftext=centb xmax=27.94 cm  xpixels=3300  ymax=21.59 cm  ypixels=2550;* colors=?;
  /* If the version is before v 9.2.2 use device=png. */
%end;
proc gplot data=ma10;
title1 j=c h=0.5 f=centb c=black ' '
            h=0.5 f=centb a = 90 ' '
            h=0.5 f=centb a =-90 ' ';
  footnote1 j=l h=0.5 f=centb ' '
            m=(9,+0) ' '
            m=(9,-1.12) ' ';
title2 j=c h=2.75 f=centb c=black &gtitle ;
   symbol1 c=black h=2 w=2 l=1 v=none  I=join;
   axis1 label=(R=0 a=0 c=black f=centb h=2.25 "Difference in" j=r "Proportion")
         value=(c=black f=centb h=2.00)
         minor=(n=1)
         width=2
         length=16.2 cm
         offset=(2.50 pct, 1.25 pct);
   axis2 label=(c=black f=centb h=2.25 &xtitle)
         style=0
         value=(c=black f=centb h=2.00)
         minor=(n=1)
         width=2
         length=21 cm
         offset=(1.25 pct, 2.50 pct);
  plot edf_dif*xm=1  / noframe nolegend
                       vaxis=axis1 vref = &dlb, 0, &dub cvref=gray66 lvref=2
                       haxis=axis2 ;
run; quit;
%mend mks2;

/* Data Step. */
options ps=50 ls=78 nodate notes nonumber;*EDF_DIFF_Plot_ex.sas;
title1 "SAS Proc T-Test Paired Comparison Example";
data a0;
input id pretest posttest @@;
 dif = posttest-pretest; if dif ne .;*exclude missing values;
lines;
01 80 82 02 73 71 03 70 95 04 60 69 05 88 100 06 84 71 07 65 75 08 37 60
09 91 95 10 98 99 11 52 65 12 78 83 13 40 60 14 79 86 15 59 62
; run;

/* Macro call. */
%mks2(a0, pretest, posttest, xtitle="Score (Pct.)",gtitle="The Difference in
EDFs",dref=0.25); quit;* This call produces a CRT Plot.;

/* %mks2(a0,pretest,posttest,xtitle="Score (Pct.)",gtitle="The Difference in EDFs",
dref=0.25, gtype=0,fname="C:\TEMP\dedfplot1.png"); quit;* This call produces a 300 dpi
png file. */
```