

# BIOMETRICS INFORMATION

(You're 95% likely to need this information)

PAMPHLET NO. # 33

DATE: September 16, 1991

SUBJECT: Box Plots

During data analysis, data is often summarized graphically to reveal special features or trends. Box plots are useful for this purpose. They are more powerful than other graphical methods, such as bar graphs or stem and leaf plots, as they contain more information. This pamphlet will describe box-plots and a SAS program for plotting them.

A given set of data can be summarized by several easily-found numbers. They are:

**maximum** : the highest extreme value;

**minimum** : the lowest extreme value;

**mean** : the average of all data values;

**median** : the middle value or the 50th percentile, also known as the 2nd quartile;

**lower hinge** : the 25th percentile value or the 1st quartile;

half-way between the minimum and median;

**upper hinge** : the 75th percentile value or the 3rd quartile;

half-way between the maximum and median.

The  $n$ th percentile value is the data point below which  $n$  percent of the data lies. For example, the 100th percentile value is the maximum, 50th percentile value is the median, and the 0th percentile is the minimum.

As an example, consider a set of DBH data ranked in increasing order:

17.2 17.3 17.5 17.7 18.1 18.3 18.5 19.0 19.1 19.4

19.5 19.6 19.7 19.8 19.9 20.0 20.2 20.3 20.4 20.6

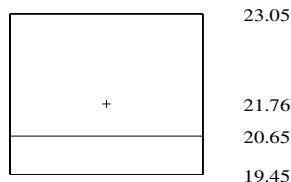
20.7 20.8 21.0 21.2 21.3 21.4 21.9 22.0 22.8 23.0

23.1 23.7 24.2 24.5 24.9 26.5 26.8 30.3 31.1 37.0

There are  $n=40$  data values; maximum=37; minimum=17.2; mean=21.76. Since there are an even number of observations, the median =  $1/2(20\text{th data} + 21\text{st data}) = 1/2(20.6+20.7) = 20.65$ ;

upper hinge =  $1/2(30\text{th data} + 31\text{th data}) = 1/2(23.0+23.1) = 23.05$ ;

lower hinge =  $1/2(10\text{th data} + 11\text{st data}) = 1/2(19.4+19.5) = 19.45$ .

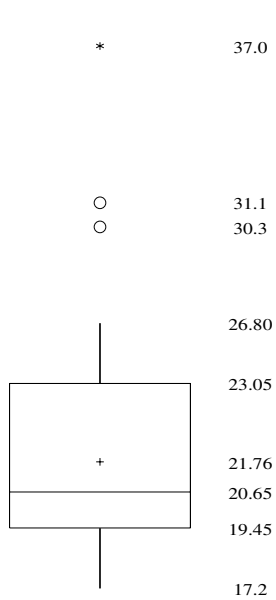


We can display these summary statistics visually by drawing a box that stretches from the lower hinge to the upper hinge, crossing it with a bar at the median and marking the mean with a "+". The box for this data is shown on the left.

Sometimes within a data set, there are values that stray out far beyond the others. A rule for identifying these "outliers" runs as follows:

1. compute **H-Spread**, the distance between the hinges (also called Interquartile Distance, IQD)  
H-spread = upper hinge - lower hinge;
2. compute the **step** which is 1.5 times H-Spread;
3. locate the **Inner Fences** which are 1 step beyond hinges;
4. locate the **Outer Fences** which are 2 steps beyond hinges;
5. **mild outliers** are the values that fall between the inner fence and corresponding outer fence;
6. **extreme outliers** are the values that fall beyond the outer fences.

For normally distributed data, 99.3% of the data should lie within the lower and upper inner fences. Therefore data observed outside of the fences are highly unlikely. To continue with the



example, H-Spread =  $23.05 - 19.45 = 3.6$ ;

step =  $1.5(3.6) = 5.4$ ;

upper inner fence =  $23.05 + 5.4 = 28.45$ ;

upper outer fence =  $28.45 + 5.4 = 33.85$ ;

lower inner fence =  $19.45 - 5.4 = 14.05$ ;

lower outer fence =  $14.05 - 5.4 = 8.65$ ;

These values can be added to the basic box plot above. A "whisker" is drawn from each end of the box to the last data point within the corresponding inner fence. In our example, the lower whisker stops at 17.2, and the upper whisker stops at 26.8, the last actual data points within the calculated lower and upper inner fences. A mild outlier is marked by a "o" and an extreme outlier is marked by a "\*". The complete box-plot is on the left.

In SAS, PROC UNIVARIATE can be used to calculate the required summary statistics and to generate a printer box plot. For example:

```
data work;
  infile 'dbh.dat';
  input dbh;
proc univariate data=work plot;
  var dbh;
run;
```

The SAS box-plot has whiskers extending from the hinges to a distance of 1.5 IQD. i.e., the whiskers do not necessarily end at an observation. This is undesirable as a whisker ending at a blank spot is misleading. Also the resulting box-plots are not of report quality. To correct these shortcomings, a SAS MACRO program has been written to generate box-plots using SAS/GRAPH on a Postscript device. This macro program is easy to use. The data must be in a SAS data set, then the macro BOX can be called up in the data step with the following arguments:

`dset` = name of the SAS data set;  
`byvar` = the list of by-variable(s);  
`yvar` = the name of the y-variable for the box-plot;  
`ylabel` = label for the y-axis; `yvar` will be used if a missing value (.) is specified.  
`title` = box-plot title; the default title is "BOX PLOT".

This program will produce box-plots for the by-variables specified.

As an example, suppose we have a set of DBH data for trees of different species and crown class (`crcl`). The following SAS code will generate box-plots for DBH by location, species and `crcl`:

```

data work1;
  infile 'dbh.dat';
  input location$ species$ crcl dbh;
  %box(work1,location species crcl,dbh,.,Box-Plot Example)
run;
  
```

The resulting box-plots are shown in figure 1a. The printer box-plots from PROC UNIVARIATE are shown in figure 1b for comparison. Observe that the second box-plot, corresponding to `crcl` 2, DF, and CO, has the median line closer to the upper hinge. This suggests that the data are skewed to the right, that is, there are more data with large values. Also, the means of the last three plots display an apparent linear downward trend, implying a linear relationship for DBH with respect to crown class.

A copy of this Macro-Box-Plot program can be obtained from the Biometrics Section of the Forest Science Research Branch. Minor changes in the program would be required if you have a printer device other than Postscript. Please send in a blank disk specifying the pamphlet number and the name of the program required. Also indicate the type of device you have available.

### Reference:

Tukey, John W., 1977, *Exploratory Data Analysis*, Addison-wesley Publishing company, Inc., CA.

CONTACT: Vera Sit  
356-0435

---

### NEW PROBLEM

---

The following is a set of data on cost per experiment (in \$10,000).

```

1.08 0.84 1.41 0.99 0.82 0.89 0.38 1.05 1.19 0.65
1.09 1.03 0.81 0.55 0.71 1.89 0.47 0.59 1.22 1.27
1.02 1.09 1.02 0.86 1.23 1.23 0.85 1.02 1.25 0.80
  
```

Rank the data in increasing order.

Find the mean, median, lower and upper hinges, H-spread, and step.

Draw a box plot for this set of data.

Box-Plot Example

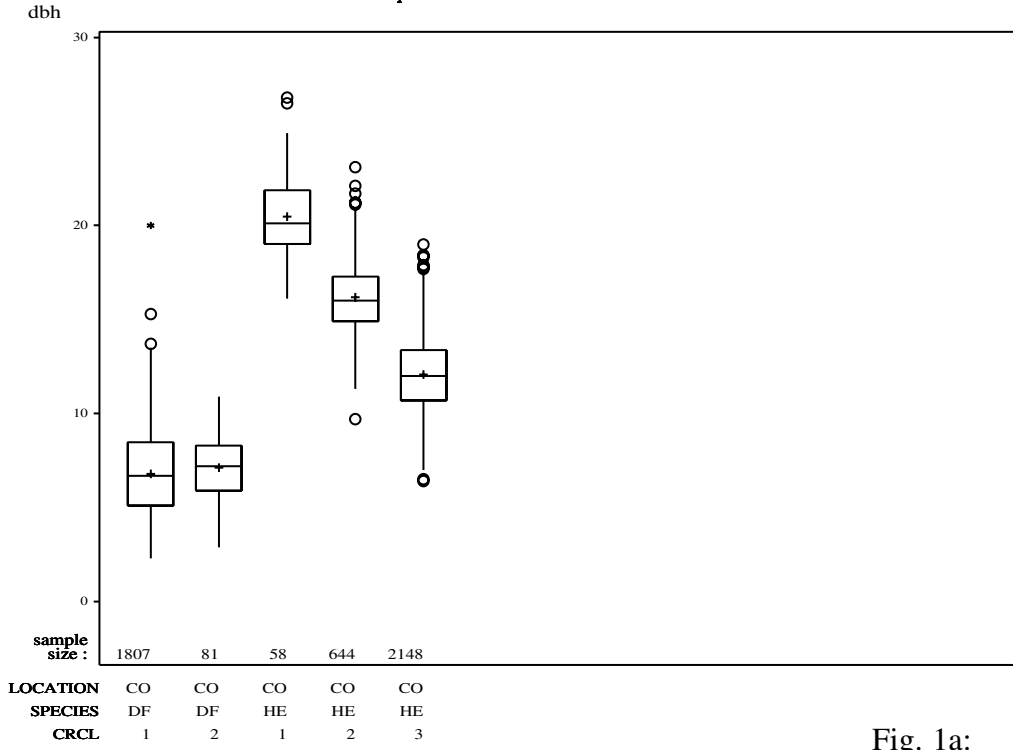


Fig. 1a:  
Macro Box-Plots

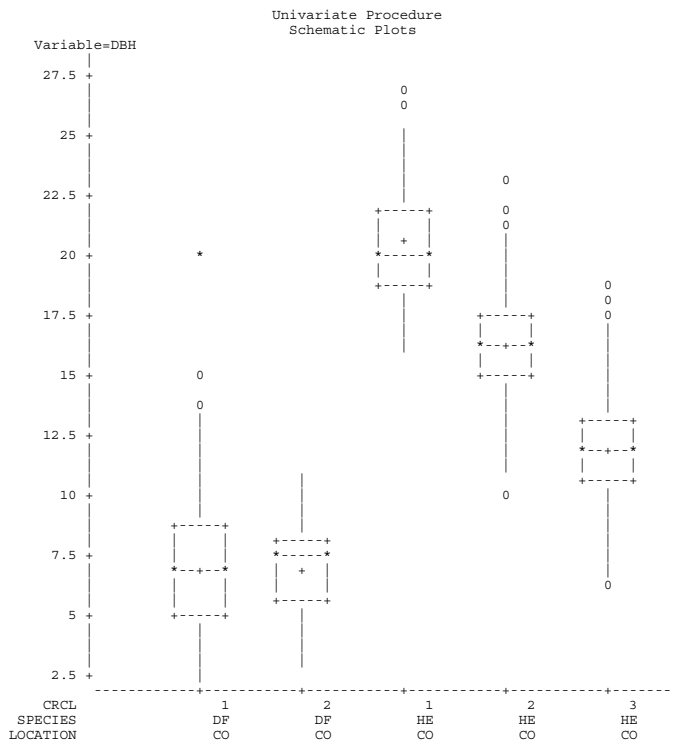


Fig 1b:  
Printer Box-Plots

## SAS MACRO PROGRAM

```

/*macro SETUP identifies the by variables & definid the y-label.*/
%macro setup;
  %if &ylabel=. %then %let ylabel=&yvar;
  %let i=1;
  %global var&i;
  %let var&i=%scan(&byvar,&i);
  %do %until (&&var&i= );
    %let var&i=%upcase(&&var&i);
    %let i=%eval(&i+1);
    %global var&i;
    %let var&i=%scan(&byvar,&i);
  %end;
  %global nby;
  %let nby=%eval(&i-1);
%mend setup;

/* macro STATS sorts the data & compute the required statistics */
%macro stats;
  proc sort data=&dset;
    by &byvar &yvar;
  proc univariate data=&dset plot ;
    by &byvar;
    var &yvar;
    output out=stats nobs=n mean=mean median=median max=max min=min
    q1=lh q3=uh qrange=iqd;
  %mend stats;

/*macro AXES creates a pseudo data set for setting up the axes for PROC GPLOT*/
%macro axes;
  proc means data=stats noprint;
    var max min;
    output out=axes n=n max=high m1 min=m2 low;
  data axes(keep=xa ya);
    set axes;
    xa=0; ya=high; output;
    xa=n*3+1;
    if (xa<31) then xa=31; ya=low; output;
%mend axes;

/* macro POINTS computes the 5 major stats. for the box-plot and classifies the
outliers as mild or extreme */
%macro points;
  data stats(drop=max min iqd whisker);
    set stats;
    by &byvar;
    whisker=1.5*iqd;
    lw=lh-whisker; uw=uw+whisker; lout=lw-whisker; uout=uw+whisker;
    if (min > lw) then lw=min;
    if (max < uw) then uw=max;
    if last.&&var&nby then output;

```

```

data comb(drop=ylast lxi lmi uxi umi);
  merge stats &dset;
  by &byvar;
  array lxout{20} lxout1-lxout20; array lmout{20} lmout1-lmout20;
  array umout{20} umout1-umout20; array uxout{20} uxout1-uxout20;
  if (first.&&var&nby) then do;
    do i=1 to 20;
      lxout{i}=.;lmout{i}=.;umout{i}=.;uxout{i}=.;
    end;
    lxi=0;lmi=0;uxi=0;umi=0;umf=0;uxf=0;lmf=0;lsx=0;
  end;
  retain lxout1-lxout20 lmout1-lmout20 umout1-umout20 uxout1-uxout20 ylast
  umf uxf lmf lxf;
  if (&yvar > uw) then do;
    if (ylast<uw) then uw=ylast;
    if (&yvar < uout) then do;
      if (umi<20) then do;
        umi+1; umout{umi}=&yvar;
      end;
    else umf=1;
  end;
  else if (uxi<20) then do;
    uxi+1; uxout{uxi}=&yvar;
  end;
  else uxf=1;
end;
if (&yvar < lw) then do;
  if (&yvar < lout) then do;
    if (lxi<20) then do;
      lxi+1; lxout{lxi}=&yvar;
    end;
  else lxf=1;
end;
else if (lmi<20) then do;
  lmi+1; lmout{lmi}=&yvar;
end;
else lmf=1;
end;
else if (ylast<lw) then lw=&yvar;
ylast=&yvar;
if last.&&var&nby then output;
%mend points;
/*macro ANNOTE creates the annotated data set for drawing the box-plots */
%macro annote;
data annote;
  set comb; by &byvar;
  array lxout{20} lxout1-lxout20; array lmout{20} lmout1-lmout20;
  array umout{20} umout1-umout20; array uxout{20} uxout1-uxout20;
  length function $8 text $200;
  x1=1+3*( _n_-1); x2=x1+1; x3=x1+2; num=put(n,6.); xsys='2'; ysys='2';
  hsys='6';

```

```

function='move'; x=x1; y=lh; output;
function='draw'; x=x3; y=lh; output;
function='draw'; x=x3; y=uh; output;
function='draw'; x=x1; y=uh; output;
function='draw'; x=x1; y=lh; output;
function='move'; x=x2; y=uh; output;
function='draw'; x=x2; y=uw; output;
function='move'; x=x2; y=lh; output;
function='draw'; x=x2; y=lw; output;
function='move'; x=x1; y=median; output;
function='draw'; x=x3; y=median; output;
function='symbol'; x=x2; y=mean; size=1.5; text='plus'; output;
do i=1 to 20;
  x=x2; size=1.2;
  if (lmout{i}<>.) then do;
    function='symbol';
    y=lmout{i};text='circle'; output;
  end;
  if (umout{i}<>.) then do;
    function='symbol';
    y=umout{i};text='circle'; output;
  end;
  if (lxout{i}<>.) then do;
    function='symbol'; y=lxout{i};
    text='star'; output;
  end;
  if (uxout{i}<>.) then do;
    function='symbol'; y=uxout{i};
    text='star'; output;
  end;
end;
function='label'; ysys='6'; x=x2; y=16; position='4';size=1.2; text=num;
output;
%do j=1 %to &nby;
  %let id=%eval(&j);
  y=15-2*&i; ysys='6'; yf=y; xtext=&&var&j;
  function='label'; xsys='2'; x=x2;
  position='4';text=xtext; output;
  function='label'; xsys='6'; x=10;
  position='4';text="&&var&j"; output;
%end;
function='label'; x=9; y=17.2; size=1.3; text='sample'; output;
function='label'; y=16; text='size :'; output;
xsys='2'; y=17.5; x=x2+0.5; flag=0; size=0.8;
if (lmf=1) then do;
  flag=1; position='a';text='1'; output;
end;
if (umf=1) then do;
  flag=1; position='c';text='2'; output;
end;
if (lxf=1) then do;
  flag=1; position='d';text='3'; output;
end;

```

```

if (uxf=1) then do;
  flag=1; position='f';text='4'; output;
end;
if (flag=1) then do;
  xsys='6'; x=10; y=yf-2.5; position='6'; size=1.3;
  function='label';
  text='NOTE:1&2indicate that there are over 20
mildoutliers in the lower and upper regions respectively.';
  output; x=15.3; y=yf-4.5; position='6'; text='3 & 4 indicate that there
are over 20 extremeoutliers in the lower and upper regions respectively.';
  output;
end;
keep function x y size text position xsys ysys hsys;proc print;
%mend annotate;

/* macro GRAPH sets up the graphic options and the plotting area */
%macro graph;
  filename graph 'box.ps';
  goptions reset=all noprompt device=ps vpos=75 hpos=100vsize=7.5in hsize=10in
  rotate=landscape chartype=5 gsfmode=replace gsfname=graph;
  symbol1 v=none;
  axis1 minor=none origin=(10,20) length=50 label=none;
  axis2 minor=none origin=(10,15) length=80 label=none major=none value=none;
%mend graph;

/*macro GPLOT plots the box-plots using the annotated data set */
%macro gplot;
  proc gplot data=axes;
    plot ya*xa/frame vaxis=axis1 haxis=axis2 annotate=annotate;
    title1 h=2 m=(20,73) cells "&title";
    title2 h=1.5 m=(3,71) cells "&ylabel";
  %mend gplot;

/*macro BOX is the main macro that calls up the other macros to draw box plots
of the y-variable 'yvar' by 'byvar' for the data set 'dset'. User can also
specifiy a label for the y-axis, 'ylabel', and a title for the plot, 'title'.*/
%macro box (dset,byvar,yvar,ylabel,title=Box Plot);
  %setup; /* identify by-var and ylabel */
  %stats; /* compute statistics */
  %axes; /* set up x- y-axes */
  %points; /* compute box-plot stats and classify outliers */
  %annotate; /* create annotated data set */
  %graph; /* set up graphic options */
  %gplot; /* draw box-plots with annotated data set */
  run;
%mend box;

/* Main program to call up macro BOX to draw box-plots for data set WORK1. */
data work1;
  infile 'pamp2.dat';
  input location$ species$ crcl dbh;
  %box(work1,location species crcl,dbh,DBH,title=Box Plot for Pamp2)
run;

quit;

```



---

 ANSWER FOR PAMPHLET #32
 

---

The specific answer to the problem in pamphlet #32 is shown at the top of the next page. As a specific example, the data from the previous pamphlet has been used with the important change of calling person 1, block 1, person2, block 2, etc. Also, the program below has a few extra wrinkles in it. Another way to obtain the repeated sums of squares and F-values is by calculating the contrasts in the data step and running separate ANOVA's on them. This is described in *Analysing data with repeated observations on each experimental unit* by J. G. Rowell and D. E. Walters in *J. agric. Sci.*, 1976, 87: 423-432. Thanks go to B. Wikeem for bringing this paper to my attention.

Note that the repeated measures design uses different error terms for the Block F-tests. Also, since the REPEATED statement does not allow the specification of an error term, the default error term must be the right one. This means, for instance, that if there is sub-sampling within each plot (t.u.) then plot means should be used in the ANOVA.

```

/* Messy332.sas */
data long (keep=B D T y)
  rep (keep=B D total linear quadr cubic t1-t4);
  B + 1;
  do D = 1 to 3;
    input t1-t4 @@;
    total = ( t1 + t2 + t3 + t4)/sqrt(4);
    linear = (-3*t1 - t2 + t3 +3*t4)/sqrt(20);
    quadr = ( t1 - t2 - t3 + t4)/sqrt(4);
    cubic = (-1*t1+3*t2-3*t3 + t4)/sqrt(20);
    output rep;
    array tm{4} t1-t4;
    do T = 1 to 4;
      y = tm{T};
      output long;
    end; end;
  label B = 'Block' D = 'Vegetation Treatment' T = 'Time';
cards;
72 86 81 77 85 86 83 80 69 73 72 74
78 83 88 81 82 86 80 84 66 62 67 73
71 82 81 75 71 78 70 75 84 90 88 87
72 83 83 69 83 88 79 81 80 81 77 72
66 79 77 66 86 85 76 76 72 72 69 70
74 83 84 77 85 82 83 80 65 62 65 61
62 73 78 70 79 83 80 81 75 69 69 68
69 75 76 70 83 84 78 81 71 70 65 63
; * book has last entry of 65 instead of 63!!;
run;

title 'Data taken from Messy Data, page 332';
proc means nway data=long noprint;
  class D T; var y;
output out=means n=num mean= mean; run;
proc print; run;
proc glm data=rep;

title2 'Method described by Rowell & Walters';
  class B D;
  model total linear quadr cubic = B D ;
  means D ;

```

```

run;
title2 'Usual split-plot analysis';
proc glm data=long;
  class B D T;
  model y = B|D T T*D;
  test h = D e = B*D;
  contrast 'Linear: time' T -3 -1 1 3;
  contrast 'Quadr: time' T 1 -1 -1 1;
  contrast 'Cubic: time' T -1 3 -3 1;
  contrast 'Linear: txd' T * D -3 -1 1 3 0 0 0 0 3 1 -1 -3,
                        T * D -3 -1 1 3 6 2 -2 -6 -3 -1 1 3;
  contrast 'Quadr: txd' T * D 1 -1 -1 1 0 0 0 0 -1 1 1 -1,
                        T * D 1 -1 -1 1 -2 2 2 -2 1 -1 -1 1;
  contrast 'Cubic: txd' T * D -1 3 -3 1 0 0 0 0 1 -3 3 -1,
                        T * D -1 3 -3 1 2 -6 6 -2 -1 3 -3 1;
lsmeans D T T*D;
run;

title2 'Usual Repeated Measures Analysis';
proc glm data=rep;
  class B D;
  model t1 t2 t3 t4 = B D/nouni;
  repeated T 4 (1 2 3 4) polynomial/summary nom printe;
run;

```

The final ANOVA table for both analyses is:

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Squares	F-values	
				Split-plot	Rep. Meas.
B	7	458.74	65.53	8.80	0.49
D	2	1333.00	666.50	4.97	4.97
B x D	14	1879.17	134.23	18.02	--
T	3	289.61	96.54	12.96	18.51
Linear	1	11.10	11.10	1.49	1.31
Quadratic	1	243.84	243.84	32.74	72.00
Cubic	1	34.67	34.67	4.65	9.17
T x D	6	527.42	87.90	11.80	16.85
D x Linear	2	83.72	41.86	5.62	4.94
D x Quadratic	2	397.75	198.88	26.70	58.72
D x Cubic	2	45.95	22.98	3.08	6.07
Pooled Error	63	469.22	7.45	--	--
T x B:	21	250.13	11.91	--	--
B x Linear	7	257.33	12.25	--	--
B x Quadratic	7	103.16	4.91	--	--
B x Cubic	7	108.73	5.18	--	--
T x B x D:	42	219.09	5.21	--	--
B x D x Linear	14	118.72	8.48	--	--
B x D x Quadr	14	47.42	3.39	--	--
B x D x Cubic	14	52.95	3.78	--	--