Paper 228-28

PROC FORMAT Supports PROC BOXPLOT to Handle Twofold Grouped Data

Barbara Schneider, Institute for Medical Statistics, University of Vienna, Austria
Kristan Schneider, University of Vienna, Austria

Abstract

Boxplots are an appropriate statistical method to compare the distributions of a continuous or ordinally scaled variable graphically between groups. Very often you have to visualize twofold grouped data using boxplots. For example think of a data set consisting of changes in blood pressure, drug and disease information where drug and disease represent the two grouping variables. Ready to use PROC BOXPLOT within SAS® Software is able to handle blocked data (nested grouping) but can't display twofold grouping in a general sense. PROC GCHART is able to handle twofold grouping to display means together with corresponding confidence intervals. The aim of the paper is to suggest a strategy to force PROC BOXPLOT to display twofold grouped data in a similar way as PROC GCHART proceeds with means. In the paper the concept of blocking and grouping will be explained in detail using examples and SAS code will be provided to demonstrate how PROC FORMAT in combination with PROC BOXPLOT facilitates to obtain the desired graphs. SAS Software refers to Version 8.2 for Microsoft ® Windows.

Objectives

Very often you have to compare twofold grouped data graphically using boxplots. For example lets assume we have performed a medical trial investigating changes in blood pressure between treatment groups across various diseases. We consider a sample data set that contains results of an experiment to evaluate drug effectiveness. Four drugs were tested against three diseases. In the following columns and some records of the data set drug are displayed:

drug	disease	chang_bp
1	1	42
1	1	44
1	1	36
.	. .	.
1	3	25
1	3	25
1	3	24
.	. .	.
.	. .	.
4	3	25
4	3	5
4	3	12
4	3	7

The adaquate statistical method to compare variable chang_bp visually is using boxplots.
PROC BOXPLOT can display grouped data that are organized in blocks. The values of a block variable must be the same for all observations with the same value of the group variable. In other words groups must be nested within blocks determined by block variables. As one can see from the data there are observations with drug=1 and disease ranging from 1 to 3. So the drug data set doesn't meet these assumptions. If the variable drug acts as the grouping and disease as the blocking variable each single drug value is associated with each value of the three diseases. A general twofold grouping that each value of one group is associated with any value of the other group cannot be handled with PROC BOXPLOT.
Visualizing only means with corresponding confidence intervals can be performed using PROC GCHART.

The following SAS code generates a graph (Tab.1) showing means together with confidence intervals for drugs grouped by disease.

```
goptions reset=global gunit=pct
border cback=white
        colors=(black blue green
red) ftitle=swissb
        ftext=swiss htitle=5
htext=3.5;

data drug;
set sasuser.drug;
 run;
title1 'Changes in Blood Pressure';
footnote;

   /* modify axes */
*axis1 label=('Changes in Blood
Pressure' j=c
             'Error Bar Confidence
Limits: 95%')
      minor=(number=1);
axis2 label=(a=90'change' j=c
)order=(0 to 50 by 10);

   /* define pattern characteristics
*/
pattern1 color=red;
pattern2 color=blue;
pattern3 color=green;

   /* produce vertical bar chart */
proc gchart data=drug;
   vbar3d drug / type=mean
   mean
             /* freqlabel='Number in
Group'
             meanlabel='Mean
'Changes in Blood Pressure' */
             sumvar=chang_bp
discrete
inside=mean
                 group=disease
             errorbar=top
             noframe
             clm=95
raxis=axis2
coutline=black
         /*   midpoints=(20  30  40
50)
             raxis=axis1
             maxis=axis2

*/
shape=block
patternid=group
gspace=3
```
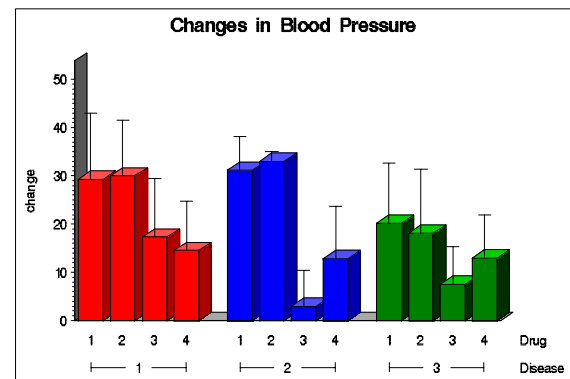
```
width=5
;
run;
quit;
title;
```

Tab.1

The weak point of this display is that means don't give as much information of the distribution as boxplots do. In the next section a strategy is introduced how to apply PROC BOXPLOT on the drug data set so that boxplots are displayed in a similar way as PROC GCHART displays means

PROC FORMAT and PROC BOXPLOT a powerfull duo to obtain the desired graph

The first step is to nest the groups artificially by coding combinations of the two group variables drug and disease in a data step. The generated codes are stored in a new colum named disdrug.

```
data f;
label disdrug="drug";
format
disdrug boxgr_.
;
set sasuser.drug;
if drug =1 and disease=1 then
disdrug=1;
if drug =1 and disease=2 then
disdrug=6;
if drug =1 and disease=3 then
disdrug=11;
if drug =2 and disease=1 then
disdrug=2;
if drug =2 and disease=2 then
disdrug=7;
```

2

```
if drug =2 and disease=3 then
disdrug=12;
if drug =3 and disease=1 then
disdrug=3;
if drug =3 and disease=2 then
disdrug=8;
if drug =3 and disease=3 then
disdrug=13;
if drug =4 and disease=1 then
disdrug=4;
if drug =4 and disease=2 then
disdrug=9;
if drug =4 and disease=3 then
disdrug=14;

run;
```

The second step is to provide an appropriate format for the new codes to hide them behind the original values. This part is covered by PROC FORMAT

```
proc format;
value boxgr_
1= '1'
2= '2'
3='3'
4= '4'
5= ' ' /* to insert spaces between
blocks */
6='1'
7= '2'
8= '3'
9= '4'
10=' '
11='1'
12='2'
13='3'
14='4'
 ;
run;
```

To compare drugs blocked by disease visually the following SAS code is submitted and displayed in Tab.2

```
goptions reset=all
;
data f;
format disdrug boxgr_.;
set f;
run;

proc sort data=f;
by   disdrug;
run;
```
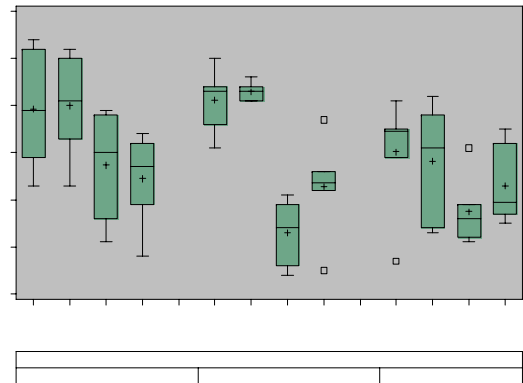
```
proc boxplot data=f;

      plot chang_bp*disdrug
(disease) /
cframe = liggr
cboxes=black
cboxfill=lig
      NPANELPOS=30 BLOCKPOS=4
/*BLOCKREP*/
CONTINUOUS
boxstyle=schematic
;
run;
quit;

title;
```

Tab.2

To compare diseases blocked by drugs (Tab.3) the following SAS code has to submitted

```
goptions reset=all
;
proc format;
value boxgr
1= '1'
2= '2'
3='3'
4= ' '
5= '1'
6='2'
7= '3'
8= ' '
9= '1'
10='2'
11='3'
12=' '
13='1'
14='2'
15='3'
 ;
run;
```

3

```
 data f;
  label disdrug="disease";

format
disdrug boxgr.
;
set sasuser.drug;
if drug =1 and disease=1 then
disdrug=1;
if drug =1 and disease=2 then
disdrug=2;
if drug =1 and disease=3 then
disdrug=3;
if drug =2 and disease=1 then
disdrug=5;
if drug =2 and disease=2 then
disdrug=6;
if drug =2 and disease=3 then
disdrug=7;
if drug =3 and disease=1 then
disdrug=9;
if drug =3 and disease=2 then
disdrug=10;
if drug =3 and disease=3 then
disdrug=11;
if drug =4 and disease=1 then
disdrug=13;
if drug =4 and disease=2 then
disdrug=14;
if drug =4 and disease=3 then
disdrug=15;
run;

proc sort data=f;
by   disdrug;
run;
proc boxplot data=f;

     plot chang_bp*disdrug (drug) /
cframe = liggr
cboxes=black
cboxfill=lig
       NPANELPOS=30 BLOCKPOS=4
/*BLOCKREP*/
CONTINUOUS
boxstyle=schematic
;

   run;
quit;

title;
```
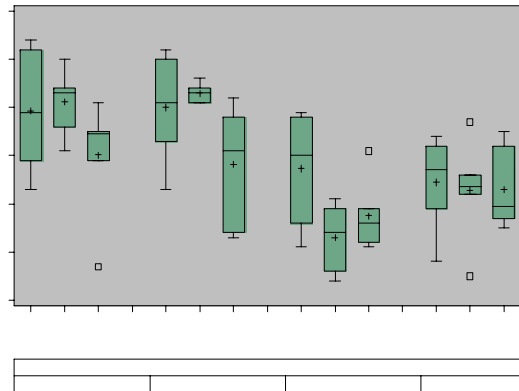
Tab.3

SAS, SAS Software, Base SAS, SAS/STAT, and SAS/GRAPH are registered trademarks of SAS Institute Inc., Cary, NC, USA

REFERENCES

SAS ® Language: Reference, Version 8
SAS/STAT® User's Guide, Version 8,
SAS/GRAPH® User's Guide, Version 8
SAS Procedures Guides

ADDRESS FOR CORRESPONDENCE

Ass. Prof. Dipl.Ing. Dr. Barbara Schneider
Institute for Medical Statistics
University of Vienna
Schwarzspanierstr. 17
A - 1090 Vienna
Austria
barbara.schneider@univie.ac.at