PROC ELPRINT: a Print Procedure that Supports Footnotes and Outlier detection

S. Stanley Young, George F. Fraction, and Richard E. Skrzynecki, Eli Lilly and Co.
Jeffrey B. Shotts, Purdue University

ABSTRACT

PROC ELPRINT is similar to PROC PRINT. Individual values can be footnoted with user defined footnotes. Two data sets are read - a print data set and a footnote dictionary. The symbol of a marked variable value is matched with its definition and the footnote is printed on the bottom of the print page. If requested, the procedure will mark statistical outliers - row or column outliers or interaction outliers (Bradu-Hawkins, 1982). The outlier detection methods are robust to presence of multiple outliers.

0.0 INTRODUCTION

PROC ELPRINT is a print procedure designed to produce report ready tables of individual observations. It is similar to PROC PRINT, but it has two major enhancements. First, if variable values are marked, they will be matched to a footnote dictionary and footnotes will be printed on the bottom of the page. Second, statistical methods can be invoked to detect unusual numbers, termed outliers, in the data.

The remainder of this paper is divided into four sections: footnotes, outliers, specifications for PROC ELPRINT, and examples.

1.0 FOOTNOTES

Often during the collection of data special circumstances are noted that should be included in reports. Footnotes are a common method used to include such information. Footnotes may be handled with customized software; oftentimes computer printouts are typed over by hand. Both procedures are expensive, time consuming and prone to error. It would be useful to have the original data marked and, during the printing, footnotes included by matching the marked values to a dictionary.

PROC ELPRINT requires two data sets. The "data" data set is in cbaracter format and a variable value to be footnoted has the footnote code, maximum of 2 characters, appended to the end of the variable value. The "dictionary" data set contains the footnote codes and their corresponding definitions.

An individual data value may be marked with one footnote; the footnote symbol may be any one or two alphanumeric characters. These are defined in the footnote dictionary. The footnote dictionary may contain sub-dictionaries. There is a limit of a total of 100 dictionary entries.

Stacked labels are supported; SUM, SUMBY and DOUBLESPACE are not supported at this time.

2.0 OUTLIERS

In addition to marking data values during collection for footnotes, it is often useful to assess values against the rest of the data and mark them if they are unusual. These unusual values have been termed outliers, mavericks, aberrant values, etc. Determination of outliers is an old and important topic in statistics. It has been the subject of recent books, Barnett and Lewis (1978) and Hawkins (1980) and a review paper, Beckman and Cook (1983). Outliers are important for two reasons. First, inclusion of truly aberrant numbers often greatly and unfairly influence subsequent statistical analysis. Second, unusual numbers are often a clue to important aspects of the data. In any case it is valuable to be able to quickly and efficiently find unusual numbers in a data set.

Detection of outliers is not a trivial problem. Their detection is dependent upon knowing the underlying model of the data and also, unfortunatly, on the presence of other outliers. Detection of multiple outliers has proven to be a particularly difficult task. The technique of Bradu and Hawkins (1982) appears to be an improvement over previous methods for detection of outliers in two-way tables. Their technique is modified slightly here so that it is non-graphical. A similar technique is developed for univariate outlier detection - detection of outliers in rows or columns of the data.

Outliers can be identified in one of three ways: (1) within each row, (2) within each column or (3) values that differ from an additive model of the table, i.e., interaction outliers, Bradu-Hawkins. Outlier identification is controlled by specifying the type, the α-level, and in the case of interaction outliers the "window" - area of the table around a value - over which the value is judged.

2.1 BRADU-HAWKINS TECHNIQUE WITH MODIFICATIONS

Consider a two dimensional array of numbers Yij,
i = 1,2,3,...n; the rows,
j = 1,2,3,...m; the columns, where

$$Yij \text{ is distributed } N(\mu ij, \sigma^2) \text{ and}$$

$$\mu ij = \mu + \alpha i + \beta j + \delta ij.$$

Except for a subset of the numbers, the δij are zero. The task is to find the Yij with $\delta ij \neq 0$. These are outliers from additivity. For each Yij in the table, all possible "tetrads"

$$Tij.eg = Yij - Yej - Yig + Yeg$$

are computed. There are (n-1)*(m-1) possible tetrads. These are ranked and the median tetrad, Tij, is saved for each Yij. At this point, the technique of PROC ELPRINT deviates from Bradu-Hawkins. A robust estimate of the standard deviation, Huber (1981), is computed:

$$s = \frac{\text{med} \mid Tij - \bar{T}ij \mid}{0.6745}$$

where $\bar{T}ij$ is the median of all median tetrads. If the Yij are normally distributed, s is an unbiased estimate of the standard deviation of the Tij. Bradu-Hawkins construct a half-normal plot of the Tij and visually inspect for outliers. Our procedure is non-graphical and proceeds as follows. First we standardize: $STij = (Tij - \bar{T}ij)/s$. Next the standard tetrads are ranked. If there are no outliers, these should be distributed as order statistics from a N(0,1). Approximations to the expected values and variances of the order statististics are computed using formulae of David and Johnson (1954) given as (80) and (81) in Biometrika Tables for Statisticians, Volumn 2 (1976). These were used to compute

$$(r)Zij = \frac{(r)Tij - Exp}{\sqrt{Var}}$$

where (r) indicates that the standardized median tetrads are ranked. There are nxm resulting (r)Zij which should be distributed as N(0,1). Various rules were considered for marking the Yij with large (r)Zij. As a rough and ready rule, which partially corrects for multiple comparisons, any Yij with |Zij| greater than Zcrit, where Zcrit is the normal deviate that cuts off $\alpha/\text{sqrt}(n*m)$ of the area under the standard normal curve, is marked as a statistical outlier. This rule partially corrects for multiple comparisons; a strict application of the Bonferroni correction would cut off $\alpha/n*m$ of the area. The user can specify the per row, column, or table α so this procedure can be tuned. Row and column outlier detection is discussed next.

2.2 ROW OR COLUMN OUTLIERS

In a number of cases the row or column may represent a more natural grouping than the "complete table". The user may specify that outliers be determined within the row or column. In these cases the values will be standardized, subtract median, divide by robust estimate s, ranked, and tested against the expected values of the normal order statistics. In the row or column case the error rate will be controlled within the row or column.

2.3 OUTLIER EXAMPLE: WORKED

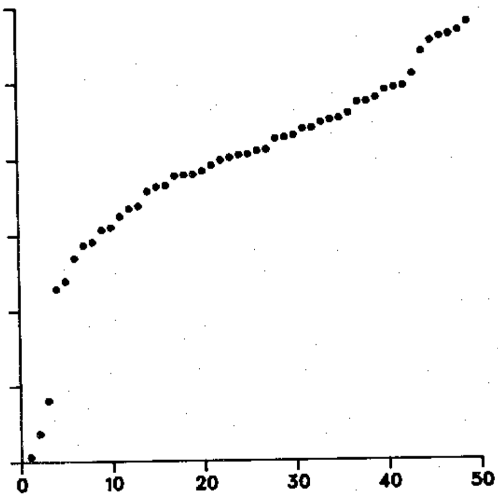
	col1	col2	col3	col4	col5	col6	col7
row1	21	68	84	14	146	79	48
row2	17	81	84	14	120	37	45
row3	144	148#	270	309	366	364	314
row4	574	624	374#	633	656	656	598
row5	652	817	533#	807	797	808	824
row6	752	940	743	879	933	878	805
row7	41	102	107	55	181	114	61

The Bradu-Hawkins data, x10, is given above. We now work through the computations for the detection of outliers in a two-way table.

For each element in the matrix, tetrads are computed. For example, for the 1,1 element, 21, the first tetrad is

$$21 - 68 - 17 + 81 = 17.$$

The ranked tetrads for the 1,1 element are
-263, -192, -72, -43, -43, -38,
-22, -7, -3, 1, 3, 3,
4, 4, 10, 14, 15, 15,
17, 21, 24, 26, 56, 63,
66, 68, 88, 97, 108, 134,
135, 141, 143, 152, 162, 172,
and the median tetrad is 16. At this point Bradu-Hawkins plot the ranked median tetrads against the intergers



to identify outliers. We proceed by dividing each median tetrad by a robust estimate of the standard deviation - for this data set, 59.3, to give the standardized median tetrads:

	col1	col2	col3	col4	col5	col6	col7
row1	0.27	-0.22	1.66	-0.90	0.38	-0.23	0.05
row2	0.58	0.25	1.77	-0.43	-0.35	-0.94	0.04
row3	-1.71	-3.19	1.61	0.46	0.30	1.10	0.73
row4	0.94	0.21	-3.93	0.73	-0.62	0.39	-0.09
row5	-1.14	0.78	-3.63	0.89	1.10	0.11	1.38
row6	0.76	1.59	-1.61	0.91	0.51	-0.17	-1.31
row7	0.09	-0.03	1.53	-0.66	0.50	0.01	-0.37

We now compute the expected values and variances of the order statistics, then compute a z-test of each standardized median tetrad:

N	Obs	(r)Tij	Exp	Var	Z
1	374	-3.93	-2.25	0.213	-3.65
2	533	-3.63	-1.85	0.116	-5.22
3	148	-3.19	-1.62	0.086	-5.36
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
47	270	1.61	1.62	0.086	-0.03
48	84	1.66	1.85	0.116	-0.54
49	84	1.77	2.25	0.213	-1.02

Finally, those numbers with |z| > Zcrit are marked, see # on original data.

2.4 DISCUSSION: OUTLIERS

Large quantities of data are routinely processed in research. It is commonly known that the actual frequency of gross errors in data sets ranges from 1 to 10 percent. False inferences can result with error rates as low as 1 percent. The discovery of gross errors in tabular data is important so that they can be corrected or removed. It is also important that correct, but unusual, numbers be identified so that scientific explanations and interpretations will address unusual features of experiments.

There are a number of advantages of these techniques over previous methods. The procedure finds multiple outliers in a single pass. It is not necessary to "find the first outlier, remove it, and try again". It has been demonstrated to be robust to multiple outliers in a data set. As a PROC, the method is easy to invoke. The method is also relatively fast when compared to humans. The scanning of a 50x10 table with a window size of 7x7 takes 6.83 seconds of cpu time on our mainframe computer and cost $1.40. In our laboratory, the identification of outliers by experts takes about 1-3 minutes per page and surprisingly costs about the same.

The Bradu-Hawkins use of half-normal plots has much to commend it. It allows the analyst to examine each point in the context of all others and see "where the real data ends and the outliers begin". This takes some skill, more time, and better trained analysts. Our procedure is a compromise: we sacrifice a graphicel look at all the data for more speed, and less paper. In our laboratory, the originators of the data rather than the statisticians usually evaluate outliers. For example, one of our laboratories generates millions of numbers per year; we have only two professional statisticians for the area. It is essential that routine data examination be done by laboratory personel, so PROC ELPRINT was developed.

The determination of median tetrads involves considerable sorting and this sorting increases dramatically with increases in the size of the table. Small windows compute rapidly; large windows give more stable estimates, but compute more slowly. To cut tbe number of objects to be sorted we allow the user to place a "window" around each Yij; the median tetrad is determined within the window. Choice of window size is a tradeoff between speed of computation - a small window - and stability of the resulting median tetrad - a large window. Our limited experience indicates a 7x7 window is a good compromise; the PROC will default to that size (if possible). The user may override the default. With a '7x7' table, there are 49 median tetrads, each of which is the median of 36 simple tetrads - 1764 objects. With a 50x10 table, about one page of numbers, there are 500 medians each determined from 441 simple tetrads - 220,500 objects.

The Bradu-Hawkins method appears to be a real breakthrough in the identification of multiple, unusual numbers in a two-way table and we consider the addition of this technique to the SAS system to be an important aid to quality control and interpretation of research data.

2.5 REFERENCES, OUTLIERS

Barnett,V. and T.Lewis. Outliers in Statistical Data. New York: John Wiley. (1978).

Beckman, R. J. and R. D. Cook. Outlier.......s. Technometrics 25:119-163. (1983).

Bradu,D. and D.M.Hawkins. Location of multiple outliers in two-way tables, using tetrads. Technometrics 24:103-108. (1982).

David,F.N. and N.L.Johnson. Statistical treatment of censored data. Part I. Fundamental formulae. Biometrika 41:228-240. (1954).

Hawkins,D.M. Identification of Outliers. London: Chapman and Hall. (1980).

Huber, P. J. Robust Statistics. New York: John Wiley. (1981).

Pearson, E. S. and H. O. Hartley. Biometrika Tables for Statisticians. Volumn 2. (1976).

3.0 SPECIFICATIONS

```
PROC ELPRINT options;
      VAR variables;
      BY  variables;
      ID  variables;
      PAGEBY variables;
```

PROC ELPRINT statement.

```
PROC ELPRINT options;
```

ELPRINT requires two input SAS data sets. The first contains the data to be printed, and is given by the "DATA =" parameter. The PROC refers to this data set as data_set_1. If the "DATA =" parameter is omitted, the PROC will default to the most recently created SAS data set. The second data set is the footnote dictionary, and is given by the "DICT=" parameter. The PROC refers to this data set as data_set_2. The "DICT=" parameter must be specified. See below for the form of the two data sets.

DATA = SASdataset	names the SAS data set to be printed. Variables that appear in the Variables statement must be CHARACTER with the right most 2 characters blank (no footnote) or containing the footnote code. Only variables in the VARIABLES list may contain footnotes.

Instructions will be provided on the preparation of the "DATA =" data set.

DICT = SASdataset	names the SAS data set footnote dictionary. It must contain the following variables.

VARIABLE	TYPE	LENGTH	EXPLANATION
FOOTNOTE DESIGNATOR	CHAR	2	Designates an internal code for the subject area. This allows multiple subject areas to "reuse" a footnote code, i.e. multiple dictionaries are contained in one global dictionary. Only one subject dictionary may be used in an invocation of PROC ELPRINT.
FOOTNOTE CODE	CHAR	2	Footnote code that is the cross-reference between the data and the dictionary, i.e. the last two characters of a variable value.
SEQUENCE NUMBER	NUM	1	provides a sort sequence number to order the lines of text in a footnote.
NUMBER OF LINES	NUM	1	provides the number of lines of text in a footnote, current limit is 2 lines per footnote.
FOOTNOTE TEXT	CHAR	60	provides text of footnote, current limit is 60 characters per line, or 120 characters for the complete footnote.

Any valid SAS name may be assigned to the above variables. We are considering the inclusion of examples that will utilize PROC FSEDIT to illustrate the creation of the dictionary data set. We are also considering the incorporation of a variable list statement for this data set. The variable list would be used to list the dictionary variables in the order expected by the procedure. See also BY variables.

OUTLIER=row (r) column (c) interaction (i)	specifies the type of outlier detection. ROW causes each row to be scanned independently. COLUMN causes each column to be scanned. INTERACTION invokes the Bradu-Hawkins method which finds non-additive elements in the table.
LEVEL=.xx L=.xx	specifies the false positive error rate of the outlier detection. The default value is 0.05.
WINDOW='XXxXX' W='XXxXX'	specifies the window size for interaction outlier detection. The window surrounds the particular Yij and tetrads are computed within the window. The default value is 7 by 7 (denoted as '7x7'), where possible.

The PROC allows column headings to be stacked up to five heading lines above a given variable. The heading list is supplied through the SAS LABEL statement. This list is split using the delimiter character given in the "SPLIT=" parameter of the PROC statement. If the "SPLIT=" parameter is omitted then the text of the LABEL is split using the first character of the LABEL as the delimiter. The LABEL option must be used if labels are to be stacked and the "SPLIT=" parameter is omitted. This code was taken from PROC TABLES, G. Fraction and S. Young, and modified. Since the processing of stacked headings is not consistent with either PROC TABLES or PROC PRINT, we are considering the establishment of uniformity in the processing of stacked headings with one of these procedures.

CODE = xx	specifies the subject dictionary within the global dictionary. Only codes with this footnote designator will be used. If this parameter is omitted then the first occurrence of the footnote code will be used.

VAR Statement

VAR variables;	specifies the numeric variables to be printed. They must be in character format in the form C....Cbb or C....CSb or C....CSS where the last two positions are blank(b) or S or SS where S is a symbol in the footnote dictionary.

NOTE: there is no variables statement for DICT, although one is being considered, as noted above.

ID Statement

ID variables;	specifies the alphabetic variables to be printed. They will be printed first. ID variables will not be scanned for footnotes nor tested for outliers. They may be character-coded numerics.

BY Statement

BY variables;	specifies the variables that control the printing. Also any outlier detection will be done within the lowest level of the BY variables. Error control will be within the BY variables for interaction, within the column within the BY variables for column outlier detection and within the row for row outlier detection. For internal technical reasons the BY variables must be in both data sets with the same attributes. In the DICT data set they must be right-most variables.

DETAILS

Missing values	Missing values are supported and will be printed. Outlier detection will ignore missing values. Sample sizes in the outlier detection methods will be adjust.

4.0 EXAMPLE: CODE TO TEST ELPRINT

```
 DATA BHTEST;
   LENGTH IDENT $ 4;
   INPUT (V1-V15)($CHAR4.);
   IDENT=PUT(_N_,4.);
   CARDS;
42  50  38  25  66  64  66  48  56  51  33  46  55  63  35
46  53  71  53  60  21  67  61  44  46  60AA50  54  54  40
46  59  45  67  63  54  52C 59  41  56  51  44  53  48  70
38  62  47  62  61  66  48  58  47  62  60  36  54  36  53
54  50  61  44  44  49  55  66  39  61  50  34  49  63  42
51  42AA56  99OL45  55  52  56B 65  32  53  42  56  63  46
52  53  43  66EE76  55  55  54  51  48  47  38  39  71  52
50  54  37  49  60  43  51  51  57  60  45  56  39  40  39
60  24  48  51  45  55  57  42  42  59  37  48  45  63  42
52  55  36  56  47  57  46  39  56  52  60  42  54  37  54
56  47  45  56  50  54  54  55  48Y 55  65  46  36  59  46
40  54  56  55  50  56Z 32  69  50  47  58  37  37  56  50
50  39  47  71  52  50  46  48  49  58EE57  56  48  45  44
63  30  52  45  55  46  59  27  47  42  54  54  36  50  36
42  48  60  44  33  31  61  48  62  63  55  59  62  45  44
 ;

 PROC PRINT;
   ID IDENT;
   VAR V1-V15;

 PROC ELPRINT LABEL
      CODE=XX
      DATA=BHTEST
      DICT=SAVE.ELPDICT;
 ID IDENT;
 VAR V1-V15;
 LABEL V1=/FIRST/VARIABLE;
 TITLE1 'TEST OF ELPRINT, A SAS PROC';
 TITLE3 'WRITTEN BY';
 TITLE4 'JEFFREY D. SHOTTS';
 TITLE5 'RICHARD SKRZYNECKI';
 TITLE6 'S. STANLEY YOUNG';
 TITLE7 'GEORGE F. FRACTION';
```

4.1 EXAMPLE: PRINT OF DATA

SAS

IDENT	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1	42	50	38	25	66	64	66	48	56	51	33	46	55	63	35
2	46	53	71	53	60	21	67	61	44	46	60AA	50	54	54	40
3	46	59	45	67	63	54	52C	59	41	56	51	44	53	48	70
4	38	62	47	62	61	66	48	58	47	62	60	36	54	36	53
5	54	50	61	44	44	49	55	66	39	61	50	34	49	63	42
6	51	42AA	56	99OL	45	55	52	56B	65	32	53	42	56	63	46
7	52	53	43	66EE	76	55	55	54	51	48	47	38	39	71	52
8	50	54	37	49	60	43	51	51	57	60	45	56	39	40	39
9	60	24	48	51	45	55	57	42	42	59	37	48	45	63	42
10	52	55	36	56	47	57	46	39	56	52	60	42	54	37	54
11	56	47	45	56	50	54	54	55	48Y	55	65	46	36	59	46
12	40	54	56	55	50	56Z	32	69	50	47	58	37	37	56	50
13	50	39	47	71	52	50	46	48	49	58EE	57	56	48	45	44
14	63	30	52	45	55	46	59	27	47	42	54	54	36	50	36
15	42	48	60	44	33	31	61	48	62	63	55	59	62	45	44

4.2 EXAMPLE: PRINT OF FOOTNOTE DICTIONARY

FOOTNOTE DICTIONARY

OBS	DESIGNATOR	CODE	SEQ NUM	LINES	TEXT
1	XX	AA	1	2	WEIGHT APPEARS TO BE OUT OF THE APPROPRIATE LIMIT. EITHER E
2	XX	AA	2	2	XCEEDED BY 10 GMS. OR GAIN IS SUSPECT.
3	XX	B	1	1	NEGATIVE WEIGHT GAIN ... THIS ANIMAL NEEDS TO EAT MORE .....
4	XX	C	1	2	THIS ANIMAL WEIGHS 75 GMS. AND SHOULD WIN AN AWARD BECAUSE T
5	XX	C	2	2	ODAY, THE SPECIAL WEIGHT IS 75 !!!
6	XX	DD	1	1	FOOD CONSUMPTION IS SUSPECT. KEEP A CLOSE EYE ON THIS BEAST
7	XX	EE	1	1	FOOD CONSUMPTION MIGHT BE EXCEEDED.
8	YY	.1	1	2	TREATMENT NUMBER IS NOT CATALOGED. CHECK PREVIOUS RECORDS.
9	XX	W	1	1	SEX MAY BE CHANGED SINCE LAST TREATMENT ... CHECK THIS OUT.
10	XX	X	1	1	TREATMENT NUMBER APPEARS FIXED. PLEASE NOTE THIS.
11	XX	Y	1	2	THIS STUDY NUMBER LOOKS LIKE ALL THE OTHERS. PLEASE CHANGE
12	XX	Y	2	2	FOR ORIGINALITY PURPOSES.
13	XX	Z	1	1	SEX CHANGED FROM MALE TO FEMALE SINCE LAST STUDY.

4.3 EXAMPLE: TEST OF ELPRINT
FOOTNOTES
OUTLIER DETECTION

TEST OF ELPRINT, A SAS PROC

WRITTEN BY
JEFFREY D. SHOTTS
RICHARD SKRZYNECKI
S. STANLEY YOUNG
GEORGE F. FRACTION

IDENT	FIRST VARIABLE	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1	42	50	38	25-	66	64	66	48	56	51	33-	46	55	63	35
2	46	53	71	53	60	21	67	61	44	46	60 AA	50	54	54	40
3	46	59	45	67	63	54	52 C	59	41	56	51	44	53	48	70
4	38	62	47	62	61	66	48	58	47	62	60	36	54	36	53
5	54	50	61	44	44	49	55	66	39	61	50	34	49	63	42
6	51	42 AA	56	99+OL	45	55	52	56 B	65	32	53	42	56	63	46
7	52	53	43	66 EE	76	55	55	54	51	48	47	38	39	71	52
8	50	54	37	49	60	43	51	51	57	60	45	56	39	40	39
9	60	24	48	51	45	55	57	42	42	59	37	48	45	63	42
10	52	55	36	56	47	57	46	39	56	52	60	42	54	37	54
11	56	47	45	56	50	54	54	55	48 Y	55	65	46	36	59	46
12	40	54	56	55	50	56 Z	32	69	50	47	58	37	37	56	50
13	50	39	47	71	52	50	46	48	49	58 EE	57	56	48	45	44
14	63	30	52	45	55	46	59	27	47	42	54	54	36	50	36
15	42	48	60	44	33	31	61	48	62	63	55	59	62	45	44

+/- : DATA SUSPECT, BRADU/HAWKINS OUTLIER TEST.
AA=WEIGHT APPEARS TO BE OUT OF THE APPROPRIATE LIMIT. EITHER EXCEEDED BY 10 GMS. OR GAIN IS SUSPECT.
C=THIS ANIMAL WEIGHS 75 GMS. AND SHOULD WIN AN AWARD BECAUSE TODAY, THE SPECIAL WEIGHT IS 75 !!!
OL=OUTLIER? SEEMS OUT OF PLACE HERE
B=NEGATIVE WEIGHT GAIN ... THIS ANIMAL NEEDS TO EAT MORE .....
EE=FOOD CONSUMPTION MIGHT BE EXCEEDED.
Y=THIS STUDY NUMBER LOOKS LIKE ALL THE OTHERS. PLEASE CHANGE FOR ORIGINALITY PURPOSES.
Z=SEX CHANGED FROM MALE TO FEMALE SINCE LAST STUDY.