Behavior Research Methods & Instrumentation
1979, Vol. 11 (6), 605-606

APL programs for interactive data analysis: Basic statistics and histograms

SELBY EVANS and FRED H. GAGE
Texas Christian University, Fort Worth, Texas 76129

Most APL systems have libraries that offer functions to execute commonly needed statistical analyses. These are usually written to meet the need of users with minimal APL experience. More experienced APL users are likely to find them cumbersome and inflexible compared with the power available in the APL language. The function may also omit features that psychologists, at least, would find valuable.

The purpose of this report is to present a small group of functions designed to meet the needs of users who have enough experience with APL to work in the desk calculator mode and perhaps to use file operations, but who do not have time to develop analytic programs for themselves. These functions, and others being planned, are designed to be readily used in the desk calculator mode and equally suitable for use in functions written by users to meet their own needs. The functions have been developed in a local APL system and incorporate features found desirable in exploratory analysis of psychological data.

Functions

Of the three functions presented (DST, HST, and SUMR), DST and HST are basic; SUMR is a second-level function that calls DST and HST and integrates their output. In the present context, SUMR conveniently illustrates the way the others are used. Figure 1 presents a data generating function and an illustrative example of the use of SUMR. Figure 2 presents the code for the three functions. Documentation of each function is provided in the following sections.

```
    ∇ SUMRDA;A;B
[1]   A←,B∘.+B←ι7
[2]   A←(,B∘.≤B)/A
[3]   B←(A,A+10),[1.5],⍉(⍪2,⍴A)⍴1 2
[4]   ''SUMR B
    ∇

       N        M       SD      SEM       95       CI
   56.00    13.00     5.83     0.79    11.43    14.57
TUKEY (1977) 5-POINTS:
    2.00     8.00    13.00    18.00    24.00
CLASS INT:  1.00  N PER SYMBOL:  1
    |       X          2
    |    XXXXX X X 22222
    |  XXXXXXXXXX222222222
    |XXXXXXXXXX2222222222222
     |....|H...|M...|H...|..
     2.        12.       22.
```

Figure 1. Illustration of use and results of the function SUMR. At the top of the figure, the code for the function SUMRDA is given; this function generates the data used in the illustration and can be used to provide test data.

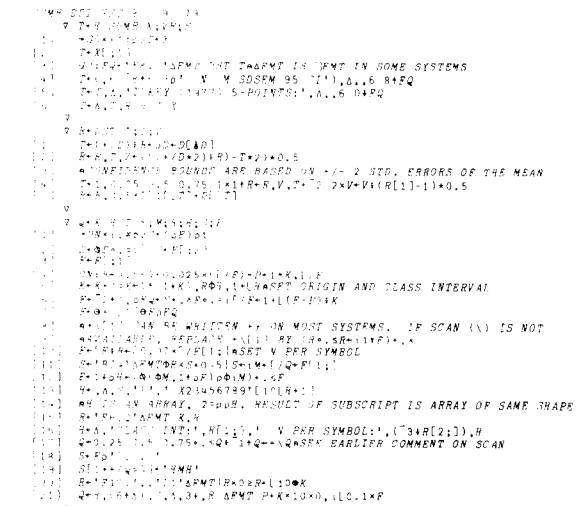

Figure 2. APL code for three functions described in the text.

R←DST D. The right argument of DST, D, is a vector of data points. DST delivers as output a numerical vector in which the elements present in order: (1) number of cases; (2) mean; (3) sample standard deviation; (4) standard error of the mean; (5, 6) approximate 95% confidence interval for the mean (strictly speaking, the interval is set as ±2 standard errors of the mean around the observed value); (7-11) Tukey's five-number summary of a distribution (Tukey, 1977, p. 33).

Q←K HST F. HST operates on a data vector, F, to produce a histogram. In the simplest mode, the left argument, K, is an empty vector and the right argument, F, is a vector of data. The result is a character vector that produces the histogram. Tukey's midpoint and the hinges are marked in the axis.

When K is an empty vector, HST determines an appropriate origin and class interval to fit conveniently on a page or CRT display. The resulting width is about 40 characters. The vertical axis will also be adjusted for convenient display, limiting column heights to about 15.

Since the maximum column height is not under control by the parameter K, users may want to be able to revise it. Column height is determined at Line 10 by the constant .07; this constant is approximately the reciprocal of maximum column height and can be changed by the user without other effects on the function.

These features permit the user to deliver any uninspected data set to HST and expect to get back a reasonably satisfactory result. To permit the user to make histograms that are easily compared, HST will accept, in K, one value specifying the origin or two values specifying the origin and the class interval, in that order.

HST offers another useful option in examining data composed of several subsets. The right argument, F, may

0005-7878/79/060605-02$00.45/0

be an N by 2 array in which the first column carries the N data points and the second column indicates the set associated with each data point. Set-designating numbers are positive integers, preferably less than 10. In this case, the resulting histogram displays a set designated by 1s as Xs; the other sets are displayed by the set number (if less than 10). The result readily shows the relationships among the distributions of the sets.

HST also leaves a global variable FQ containing the frequency distribution in numerical form. If separate sets are used in the histogram, FQ has rows corresponding to the sets, ordered by increasing magnitude of the set-designating numbers.

T←H SUMR X. SUMR accepts arguments identical to those specified for HST. It calls both DST and HST, organizes the result with appropriate labels, and delivers it in a character vector that may be output on the terminal, filed for later reproduction, or both.

The usefulness of the results produced by SUMR should be familiar. The subset frequencies provided by HST are particularly valuable in comparing distributions of treatment and control groups or in examining the discriminating power afforded by a discriminant function analysis.

Operating Considerations and Use

The number of data points processed is not limited by the functions. In practice, limits on the number of data points are imposed by the efficiency of the interpreter in use of space and by the size of available work space. The functions should be able to process at least a few hundred data points under normal APL working conditions. Execution time is generally unimportant in an APL context. From the user's viewpoint, execution produces no appreciable delay at the terminal beyond that imposed by the timesharing operations of the system.

The functions are operating on a Honeywell Sigma 9, but they were written to use only functions and conventions generally accepted in the APL community. Potential problems are noted in the program comments.

These functions may suggest the need for other functions to support the entry and correction of data or to meet needs in correlational as well as univariate studies. Reports are presently being prepared on functions compatible with those described here and designed to meet such needs.

Availability

Additional information is available from Selby Evans. Psychology Department, Texas Christian University, Fort Worth, Texas 76129.

REFERENCE

TUKEY, J. W. *Exploratory data analysis.* Reading, Mass: Addison-Wesley, 1977.

(Accepted for publication November 5, 1979.)