Behavior Research Methods & Instrumentation
1976, Vol. 8 (3), 302-305

Program Abstracts/Algorithms

JACKK: A general jackknifing routine

CRAIG E. SPITZER
University of Illinois, Champaign, Illinois 61820

A FORTRAN subroutine subprogram is presented which will perform the jackknife statistical procedures on varied statistics and data sets of the user's choice. The procedures will provide a best single estimate of the population value of the user's parameter and will provide a confidence interval around that estimate. The procedures do not make any assumptions about the sampling distribution of the statistics and thus are useful for making estimates for statistics with unknown sampling distributions. Input to and output from the routine are discussed with examples.

Function. JACKK is a subroutine subprogram written for the WATFIV compiler. It will also run on standard FORTRAN compilers (IBM FORTRAN Level G).

Using the procedures described by Mosteller and Tukey (1969), subroutine JACKK jackknifes any statistic of the user's choice. Subroutine JACKK is designed to accept data sets of variable sizes and to perform the jackknife procedures systematically excluding subsets of data from the calculations when the size of the excluded subset may be fixed at the user's request. JACKK is designed to accept input data in any format selected by the user.

Jackknifing. Jackknifing is a procedure which will generate the best estimate of a population value and the confidence limits around that value for a sample statistic when the sampling distribution of that statistic is unknown. This procedure is thus very useful for research in the social sciences, as the measures used in this research are often developed around a particular problem and a particular data set. Use of the jackknife in these cases is worthwhile, then, to test hypotheses about population parameters when the sampling distributions of the estimators are unknown.

The procedures for jackknifing (see Quenouille, 1956; Tukey, 1958; Durbin, 1959) estimate confidence limits for a measure by dividing the data into groups (subsets), and seeing the effect on the measure that would be produced by systematically omitting each of the groups. This is accomplished by calculating the set of values Y^*_j, where

$$Y^*_j = k\,Y_{all} - (k-1)\,Y_{(j)} \qquad j = 1, 2, 3, \ldots, k \qquad (1)$$

In Equation 1, Y_{all} is the measure calculated with all k data points and $Y_{(j)}$ is the measure calculated with all the data points except the jth group. The Y^*_j (called "pseudovalues") are averaged to yield the best single estimate of the true population value, and the variance of the distribution (s_*^2) is estimated by

$$s^2 = \frac{\Sigma Y^{2*}_j - 1/k(\Sigma Y^*_j)^2}{k-1} \qquad (2)$$

$$s_*^2 = s^2/k. \qquad (3)$$

(see Mosteller & Tukey, 1969). The confidence limits around the best single estimate are simply

$$\pm |t| \times s^* \qquad (4)$$

where the T value is calculated at the desired degree of confidence, and it has k − 1 degrees of freedom. If a measure like a median is jackknifed where, due to the type of the measure, only a few discrete Y^*_j values result, Mosteller and Tukey (1969) argue that the degrees of freedom for the T statistic should be the number of distinct Y^*_j values generated minus 1.

Mosteller and Tukey (1969) also note that, due to rounding in the calculations, some lack of precision occurs. They then suggest a correction factor. This correction factor is not present in this version of JACKK. The correction factor was omitted because, for the precision offered by the present routine, it would amount to an increase in s_*^2 of only 10^{-13}, approximately.

Input to JACKK. JACKK is a subroutine subprogram, and may thus be called many times in the course of a user's program. For each call of JACKK, the following data cards must be entered.

Card 1 consists of two parameters: NS and NG. These values represent the number of subjects and the number of groups, respectively. This data card is read by the user's calling program (discussed subsequently) and can be formatted by the user as integers of any length.

Card 2 contains two pieces of data. In the first three columns of the card are entered the limits of the confidence interval (e.g., ".95" or ".66"). The next 16 columns are used to input a label for the statistic being jackknifed (e.g., "ABSOL. DEV. ").

Card 3 contains the format that the user will use to input the sample data. Card 3 begins and ends with a left and right parenthesis, and the format is to be in standard FORTRAN. The format may be up to 72 characters long.

Card 4 to Card N + 3 follow and are the N data cards, formatted as the user described in Card 3.

Output from JACKK. Each time JACKK is called the following will be outputted.

(1) A list of all the pseudovalues (Y^*_j) generated. This list will print the number of the group omitted in calculating the particular Y^*_j for j = 1,999, and for cases where there are more than 1,000 groups involved in the calculation, JACKK will print asterisks in place of the j values for those values greater than 1,000. Nevertheless, JACKK will compute the estimate and interval using all the data.

This work was supported by NSF Grant SOC-7305697 awarded to James H. Davis, principal investigator. The author would like to thank James H. Davis and Ronald Hinkel for suggestions and criticisms.

(2) JACKK will then print the statistic it has just jackknifed and the estimate and interval associated with that statistic and with the particular set of data used.

The Calling Program. JACKK was written as a subroutine subprogram rather than a program to enable the user: (1) to store JACKK in object form and save the cost of recompilation at each use and (2) to allow several different statistics and/or data sets to be jackknifed from one main program. Therefore, to use JACKK, the user has to have a main program to call JACKK. The main program can be any program the user wishes and is only bound by the following conditions.

(1) The user must write a function subprogram for each statistic he wishes to jackknife. The function can have any name acceptable in standard FORTRAN and must operate from three parameters: DATA, NS, and NPG. DATA is the array of data from which the statistic will be calculated, NS is the number of data points in the data set, and NPG is the number of points in each data group. The function should contain a DIMENSION statement dimensioning the array DATA to have NS values [" DIMENSION DATA(NS)"]. Finally, the function must have a statement assigning to a variable with the name of the function, the value of the function, given the input data.

(2) The user's main program must contain a statement reading

EXTERNAL f1, f2, . . . , fN

where f1-fN are the N function names assigned by the user to compute the N statistics to be jackknifed in the program.

(3) The user's main program must contain a statement reading

DIMENSION X(v1), Y(v2), Z(v3)

where v1, v2, and v3 are the greatest number of data points in any one data set in the program to be jackknifed. In the example presented following this section, there are two statistics jackknifed. The first has a data set with 11 values, and the second has a data set with 55 values. For this example, in the main program, X, Y, and Z are dimensioned to 55.

(4) Each call of JACKK must have the following form. (a) Preceding the CALL statement must be a READ statement, where the user reads in the values of NS and NG. This can be formatted in any manner desired by the user. (b) Following the READ and before the CALL statement will be a statement reading

NET = NS – NS/NG.

(c) Following these statements will be the CALL statement using the following form:

CALL JACKK (NS, NET, NG, function, X, Y, Z)

where function is the user chosen name of the function that calculates the statistic to be jackknifed.

Example. To aid in using JACKK an example is provided. The program jackknifes two statistics using two separate data sets. The examples chosen are the ones used by Mosteller and Tukey (1969, pp. 133-144) and that text may be consulted for further details.

Note first that the two functions to be jackknifed were named EX1 and EX2 and that the first statement of the main program is the EXTERNAL statement listing these function names. Note also that in the main program arrays X, Y, and Z are dimensioned to be 55, the size of the largest data set.

The program is straightforward and simply twice goes through the steps necessary to call JACKK. Following the program listing are the pseudovalue lists and confidence intervals for the two statistics and data sets input into JACKK.

```
      $JOB
      C        DEMONSTRATION USE OF SUBROUTINE JACKK.
      C             THIS MAIN PROGRAM USES JACKK TO JACKKNIFE TWO
      C                  DIFFERENT STATISTICS, USING TWO DIFFERENT SETS
      C                  OF DATA.  EXAMPLES WERE CHOSEN FROM THE
      C                  HANDBOOK OF SOCIAL PSYCHOLOGY, CHAPTER 10.
 1          EXTERNAL EX1,EX2
 2          DIMENSION X(55),Y(55),Z(55)
 3          READ(5,91)NS,NG
 4          NET=NS-NS/NG
 5          CALL JACKK(NS,NET,NG,EX1,X,Y,Z)
 6          READ(5,91)NS,NG
 7       91 FORMAT(2I3)
 8          NET=NS-NS/NG
 9          CALL JACKK(NS,NET,NG,EX2,X,Y,Z)
10          STOP
11          END

12          SUBROUTINE JACKK(NS,NET,NG,FUNCT,DATA,DATA1,YSTAR)
13          DIMENSION YSTAR(NS),STAT(4),DATA(NS),DATA1(NET),FORM(18)
14          WRITE(6,900)
15      900 FORMAT('1',25X,'PSEUDOVALUES'/)
16          READ(5,901)       CI,(STAT(I),I=1,4),(FORM(J),J=1,18)
17      901 FORMAT(    F3.2,4A4/18A4)
18          READ(5,FORM)(DATA(I),I=1,NS)
19          NPG=NS/NG
20          FALL=FUNCT(DATA,NS,NPG)
21          CI=1.-CI
22          SUM=0.
23          SUM2=0.
24          DO 1 J=1,NG
25          N=1
26          DO 2 I=1,NS,NPG
27          IF(((I+NPG-1)/NPG).EQ.J)GO TO 2
28          IND=I+NPG-1
29          DO 3 L=I,IND
30          DATA1(N)=DATA(L)
```

```
      3 N=N+1
      2 CONTINUE
        YSTAR(J)=(NG*PALL)-((NG-1)*FUNCT(DATA1,NET,NPG))
      1 WRITE(6,902)J,YSTAR(J)
    902 FORMAT(' ',20X,'Y*(',I3,') = ',F10.5)
        DO 4 K=1,NG
      4 SUM=SUM+YSTAR(K)
        YASTER=SUM/FLOAT(NG)
        DO 5 I=1,NG
      5 SUM2=SUM2+YSTAR(I)**2
        SSTAR=SQRT(((SUM2-(SUM**2)/NG)/(NG-1))/NG)
        WRITE(6,903)(STAT(I),I=1,4),YASTER,SSTAR,YASTER,SSTAR,CI,NG
    903 FORMAT('1',///10X,'THE TWO SIDED CONFIDENCE LIMITS ON ',4A4,' ARE:'
     1,///F14.6,' + ABS VALUE(T)  X ',F14.6,//20X,'AND',//F14.6,' - ABS
     2VALUE(T)  X ',F14.6,///10X,'WHERE "ABS VALUE(T)" IS THE ABSOLUTE V
     3ALUE OF THE T STATISTIC AT ',F4.2,' WITH ',I3,'-1 DEGREES OF FREED
     4OM,',////20X,'IF THE JACKKNIFED STATISTIC IS A MEDIAN, OR ONE THAT
     5 WOULD PRODUCE'/'      ONLY A FEW PSEUDOVALUES (Y*J), USE THE
     6NUMBER OF DISTINCT VALUES MINUS ONE AS THE DEGREES OF FREEDOM,',//
     7/)
        RETURN
        END

        FUNCTION EX1(DATA,NS,NPG)
        DIMENSION DATA(NS)
C           EXAMPLE: STANDARD DEVIATION - SEE HANDBOOK OF SOC PSY, VOL 2,
C                P. 139
        XBAR=0.
        DO 1 I=1,NS
      1 XBAR=XBAR+DATA(I)
        XBAR=XBAR/FLOAT(NS)
        TOP=0.
        DO 2 I=1,NS
      2 TOP=TOP+(DATA(I)-XBAR)**2
          EX1=SQRT(TOP/(NS-1))
        RETURN
        END

        FUNCTION EX2(DATA,NS,NPG)
        DIMENSION DATA(NS)
        DIMENSION IND(200)
C   EXAMPLE 2: ESTIMATING TOP 10% OF A POP OF MEASURES
C   SEE THE HANDBOOK OF SOC PSY, VOL 2, PP 141-144.
        DO 11 I=1,NS
     11 IND(I)=I
        NLES1=NS-1
        DO 16 I=1,NLES1
        IPLUS1=I+1
        B=DATA(IND(I))
        DO 12 J=IPLUS1,NS
        IF(B.GE.DATA(IND(J)))GO TO 12
        MTEMP=IND(J)
        IND(J)=IND(I)
        IND(I)=MTEMP
        B=DATA(MTEMP)
     12 CONTINUE
     16 CONTINUE
        SPOT=FLOAT(NS+1)/10.
        ISPOT=(NS+1)/10
        EXTRA=SPOT-FLOAT(ISPOT)
          EX2=DATA(IND(ISPOT))-(EXTRA*(DATA(IND(ISPOT))-DATA(IND(ISPOT+1
     1)))))
        RETURN
        END
```

PSEUDOVALUES

```
Y*( 1) =    1.13997
Y*( 2) =    1.13997
Y*( 3) =    1.13997
Y*( 4) =    0.88932
Y*( 5) =    0.82427
Y*( 6) =    0.63248
Y*( 7) =    0.62033
Y*( 8) =    0.62189
Y*( 9) =    0.83542
Y*(10) =    0.83542
Y*(11) =    7.70394
```

```
   THE TWO SIDED CONFIDENCE LIMITS ON STAND. DEV.       ARE:

1.489360 + ABS VALUE(T)  X       0.624405

              AND

1.489360 - ABS VALUE(T)  X       0.624405

WHERE "ABS VALUE(T)" IS THE ABSOLUTE VALUE OF THE T STATISTIC AT 0.05 WITH  11-1 DEGREES OF FREEDOM,

      IF THE JACKKNIFED STATISTIC IS A MEDIAN, OR ONE THAT WOULD PRODUCE
ONLY A FEW PSEUDOVALUES (Y*J), USE THE NUMBER OF DISTINCT VALUES MINUS ONE AS THE DEGREES OF FREEDOM,
```

PSEUDOVALUES	
Y*(1) =	9.19699
Y*(2) =	4.97600
Y*(3) =	5.67200
Y*(4) =	4.97600
Y*(5) =	4.97600
Y*(6) =	4.97600
Y*(7) =	4.97600
Y*(8) =	5.67200
Y*(9) =	9.19699
Y*(10) =	4.97600
Y*(11) =	4.97600

THE TWO SIDED CONFIDENCE LIMITS ON UPPER 10% ARE:

5.869996 + ABS VALUE(T) X 0.502815

AND

5.869996 - ABS VALUE(T) X 0.502815

WHERE "ABS VALUE(T)" IS THE ABSOLUTE VALUE OF THE T STATISTIC AT 0.05 WITH 11-1 DEGREES OF FREEDOM.

IF THE JACKKNIFED STATISTIC IS A MEDIAN, OR ONE THAT WOULD PRODUCE ONLY A FEW PSEUDOVALUES (Y*J), USE THE NUMBER OF DISTINCT VALUES MINUS ONE AS THE DEGREES OF FREEDOM.

REFERENCES

Durbin, J. A note on the application of Quenouille's method of bias reduction to the estimation of ratios. Biometrika, 1959, 46, 477-480.

Mosteller, F., & Tukey, J. W. Data analysis, including statistics. In E. Aronson and G. Lindzey (Eds.), *The handbook of social psychology*. Reading, MA: Addison-Wesley, 1969.

Quenouille, M. H. Notes on bias in estimation. Biometrika, 1956, 43, 353-360.

Tukey, J. W. Bias and confidence in not-quite large samples. Annals of Mathematical Statistics, 1958, 29, 614. (Abstract)