ABSTRACT


A general purpose macro which computes the Kruskal-Wallis H-test, including the ties correction factor, and Scheffe type contrasts for ordered categories in a one-way layout is presented. An example of its use on a set of ground beef taste panel data is given.

INTRODUCTION

Most procedures for categorical data are designed to test hypotheses about cell probabilities. However, when the categories are ordered the need to test hypotheses concerning central tendency is common. Tests designed for cell probabilities have low power for these problems, and so, an alternative testing procedure is needed.

Klotz and Tong (1977) have proposed that the Kruskal-Wallis H-test be used. A Scheffe type method of contrasts was also presented. They derived an algorithm, which can require large amounts of cpu time, to compute the exact null distribution of their test statistics.

A macro, KRWALC, which performs this analysis is presented. The only difference is that the chi-square approximation is used to compute p-values rather than the exact distribution. An example of the macro’s use is presented.

STATISTICS

The H-Statistic

The H-statistic is computed in the usual manner. Each individual observation is replaced by its rank, breaking ties by using average ranks. Letting \( n \) denote the total number of observations, \( k \) the number of treatments, \( c \) the number of categories, \( n_i \) the number of observations from treatment \( i \), and \( r_j \) the number of observations in ordered category \( j \) we have

\[
H = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i^2 - 3n(n+1) \quad (1)
\]

\[
H \text{ is asymptotically distributed as chi-square with } k-1 \text{ degrees of freedom.}
\]

Contrasts

We define contrasts in the usual way,

\[
b = \sum_{i=1}^{k} a_i r_i \quad (2)
\]

where \( \sum_{i=1}^{k} a_i = 0 \).

Tests of the hypothesis \( b=0 \) can be obtained from simultaneous confidence intervals, given by Klotz and Tong (1977), for all contrasts described by (2) above. The chi-square distribution with \( k-1 \) degrees of freedom is used to compute p-values.

MACRO KRWALC DOCUMENTATION

Usage

KRWALC performs the analysis described. A listing of the macro is given in Figure 2. It can be used both on batch and interactively. If no contrasts are desired, another macro may be defined by deleting those statements with a vertical line in the margin of Figure 2.

Treatment of Missing Values

Observations are deleted only for the variables for which they are missing. Missing values are allowed for the classification variables and are treated as a distinct level. A terminal error occurs if any of the dependent variables is missing on every observation for any level of the classification variables.

SAS Statements

The following statements are necessary to execute KRWALC:

- MACRO IND A 8
- MACRO JIND B 8
- MACRO INVAR C 8
- MACRO CLASSES D 8
- MACRO TITLES E 8
- KRWALC

271
A is the name of the data set containing the raw data.

B is the name of the data set containing the contrasts. The number of observations should equal the number of levels of the classification variables. The variables define the contrasts and their names are used as contrast identifiers. The classification variables must be in this set.

C is the list of dependent variables (numeric only).

D is the list of classification variables (numeric/character). Each unique combination of values defines a level.

E is the listing information corresponding to title lines one through nine.

**Job Control Language**

Either a JCL or an allocate statement must be provided for the dname Proc001 to store the printout from PROC FREQ. The amount of space must be large enough to contain the printout of the frequency distributions for each of the dependent variables.

**NOTE:** no ODS parameters should be specified. Failure to observe this requirement may cause incorrect results.

**Error Messages**

'***ALL OBS MISSING FOR AT LEAST 1 VARIABLE. CELL FREQUENCIES PRINTED.' For each problem level, the values of the classification variables and the cell frequencies are printed. Cell frequencies are printed as "variable frequency."

'***DIVERSE IN CLASS VARIABLE VALUES' indicates that the values of the classification variables differ in the raw data and the contrasts. Each combination of values (level) that appears in one but not in the other is printed along with the values of the special variables INRAW and INCONT. A value of one for the variables indicates that the specified level is in the data set of raw data or contrasts respectively. A zero indicates that it is not in the specified data set.

Both of the above are terminal errors. The messages are printed in the SAS log. If the first error is encountered, no check is made for the second. As soon as either error is encountered SAS enters syntax checking mode.

'SUM OF COEFFICIENTS NOT ZERO FOR AT LEAST ONE CONTRAST. EXECUTION CONTINUES.' This message will be printed in the procedure output pages along with the sum of the coefficients for all contrasts if any contrast has a non-zero sum. This is a warning message only.

**EXAMPLE**

The data (courtesy of R. Cross and J. Nichols, ARS, AMRI) is consumer ratings of seven different formulations of ground beef. The ratings are on a scale of one to nine. The constituents of the formulations are mechanically deboned meat (MMD), soy protein feed (SPF), and peanut meal (PM). The formulations are given in Table 1.

<table>
<thead>
<tr>
<th>Formulation</th>
<th>% MMD</th>
<th>% SPF</th>
<th>% PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>.05</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>.1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>.05</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>.1</td>
</tr>
</tbody>
</table>

The dependent variables are tenderness and juicyness. We are interested in the contrasts corresponding to the linear and quadratic effects of MMD, SPF, and PM.

The statements used and the resultant output are given in Figure 1.

**REFERENCES**


### Sample Run of Kruskal-Wallis on Taste Panel Data

**Rank Means**

<table>
<thead>
<tr>
<th>OBS</th>
<th>TRT</th>
<th>TENDER</th>
<th>JUICY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>154.486</td>
<td>178.176</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>173.787</td>
<td>164.120</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>204.152</td>
<td>185.811</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>136.888</td>
<td>144.874</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>78.141</td>
<td>105.487</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>150.614</td>
<td>102.200</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>208.474</td>
<td>161.103</td>
</tr>
</tbody>
</table>

**Listing of Contrast Coefficients**

<table>
<thead>
<tr>
<th>OBS</th>
<th>MDMLIN</th>
<th>MDQUAD</th>
<th>SPFLIN</th>
<th>SPFQUAD</th>
<th>PMLIN</th>
<th>PMQUAD</th>
<th>TRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

**Sums of Coefficients** (Blank page denotes all sums = 0)

**Kruskal-Wallis H and Asymptotic P-values**

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>H</th>
<th>P_VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TENDER</td>
<td>71.0912</td>
<td>0.0001</td>
</tr>
<tr>
<td>JUICY</td>
<td>25.3568</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

**Standardized Contrast Values and Asymptotic P-values**

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>CONTRAST</th>
<th>STDVALUE</th>
<th>P_VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TENDER</td>
<td>MDMLIN</td>
<td>4.0703</td>
<td>0.0108</td>
</tr>
<tr>
<td>TENDER</td>
<td>MDQUAD</td>
<td>1.8775</td>
<td>0.0927</td>
</tr>
<tr>
<td>TENDER</td>
<td>SPFQUAD</td>
<td>-4.2299</td>
<td>0.0005</td>
</tr>
<tr>
<td>TENDER</td>
<td>SPFQUAD</td>
<td>-0.8703</td>
<td>0.9532</td>
</tr>
<tr>
<td>TENDER</td>
<td>PMLIN</td>
<td>9.8842</td>
<td>0.1741</td>
</tr>
<tr>
<td>TENDER</td>
<td>PMQUAD</td>
<td>1.3085</td>
<td>0.9242</td>
</tr>
<tr>
<td>JUICY</td>
<td>MDMLIN</td>
<td>1.9021</td>
<td>0.9662</td>
</tr>
<tr>
<td>JUICY</td>
<td>MDQUAD</td>
<td>1.2227</td>
<td>0.9175</td>
</tr>
<tr>
<td>JUICY</td>
<td>SPFQUAD</td>
<td>-4.0859</td>
<td>0.0012</td>
</tr>
<tr>
<td>JUICY</td>
<td>SPFQUAD</td>
<td>-0.1738</td>
<td>1.0030</td>
</tr>
<tr>
<td>JUICY</td>
<td>PMLIN</td>
<td>-0.3547</td>
<td>0.8887</td>
</tr>
<tr>
<td>JUICY</td>
<td>PMQUAD</td>
<td>0.4154</td>
<td>0.9909</td>
</tr>
</tbody>
</table>

Figure 1. Listing of SAS statements used (top) and the output from the sample run of KRUSCAL.
MACRO KRUWALC PERFORMS THE KRUSKAL-WALLIS H-TEST IN A ONE-WAY LAYOUT. FOR OBSERVATIONS WHICH ARE COUNTS CORRESPONDING TO DISTINCT CATEGORIES AND ARE TREATED AS A DISTINCT LEVEL, AN ERROR OCCURS IF ANY OF THE DEPENDENT VARIABLES IS MISSING ON EVERY OBSERVATION FOR ANY LEVEL OF THE CLASSIFICATION VARIABLES.

THE FOLLOWING STATEMENTS ARE NECESSARY TO INVOKE MACRO KRUWALC:

MACRO IND A X
MACRO INCD B X
MACRO INVAR C X
MACRO CLASSES D X
MACRO TITLES E X
MACRO KRUWALC

WHERE A IS THE NAME OF THE DATA SET CONTAINING THE RAW DATA;
B IS THE NAME OF THE DATA SET CONTAINING THE CONTRASTS;
C IS THE LIST OF DEPENDENT VARIABLES (NUMERIC ONLY);
D IS THE LIST OF CLASSIFICATION VARIABLES (NUM/ALPHA);
E IS THE TITLING INFORMATION FOR LINES 1 THRU 99;
A JCL (ALLOG) STATEMENT IS NECESSARY FOR FT20F88!
A CONSULT PRINTTO NEW UNIT=2; DATASETS ARE THROWN AWAY ON CONFLICTS.

MACRO KRUWALC

• COMPUTE TIES CORRECTION FACTOR, OPTIONS NODATE NONUMBER MISSING="", TITLE;
• PROC PRINTTO NEW UNIT=20; PROC FRE2 DATA=IND; TABLES INVAR;
• PROC PRINTTO, RUN;
• OPTIONS DATE NUMBER; TITLES;
• DATA GAMMA; INFILE FT20F88; LENGTH=L;
• INPUT CC & #'s;
• IF L=10 OR CC=-1 THEN DELETE, +ELIMINATE BLANK RECORDS;
• INPUT & ID ; F? OF P? CP ?;
• ERROR=0;
• IF ID=0;
• GAMMA(F=FF-F);
• IF DP<100 THEN RETURN.
• OUTPUT;
• GAMMA:=0;
• KEEP GAMMA;
• REPLACE DATA BY RANKS, COMPUTE RANK MEANS AND PRINT;
• PROC SORT DATA=IND; BY CLASSES;
• PROC RANK OUT=RANKS; VAR INVAR;
• PROC MEANS NOPRINT;BY CLASSES; VAR INVAR; OUTPUT OUT=RANKS MEAN=INVAR;
• PROC PRINT;
• TITLE1 RANK MEANS;
• PROC MEANS NOPRINT DATA=IND; BY CLASSES; VAR INVAR;
• OUTPUT OUT=N=N=INVAR;

• CHECK FOR ALL OBS MISSING FOR ANY LEVEL FOR ANY VARIABLE.
• DATA = NULL SET END=EOF;
• IF MINCOF INVAR D=* THEN ERROR
• *** ALL OBS MISSING FOR AT LEAST 1 VARIABLE; CELL FREQS PRINTED.
• CK=ERROR;
• DROP OK,
• IF EOF AND CK THEN ABORT;

• CHECK FOR CONFLICTS IN CLASS VARIABLE VALUES;
• PROC SORT DATA=IND; BY CLASSES;
• PROC PRINT;
• TITLE1 LISTING OF CONTRAST COEFFICIENTS;
• DATA = NULL ; MERGE RANKS=INVAR KEEP=CLASSES) INCD=IND INCONT
• KEEP=CLASSES) END=EOF; BY CLASSES;
• IF INVAR=INCONT<=2 THEN ERROR
• ***CONFLICT IN CLASS VARIABLE VALUES
• CK=ERROR;
• DROP OK,
• IF EOF AND CK THEN ABORT;

Figure 2. Listing of KRUWALC. Vertical line in left margin indicates statements to be deleted if no contrasts are desired.
PROC MATRIX ERRMAX=0;
FETCH RBAR DATA=RANKS<KEEP=INVAR> COLNAME=VID; VARID=VID'; FREE VID;
FETCH GAMMA DATA=SIGMA;
FETCH N DATA=(KEEP=INVAR);
NVAR=NROW(RBAR);
NTRT=NROW(RBAR);
NTOT=(NTRT)*N;
*COMPUTE H-TEST AND P-VALUES;
CONSTANT=(NVAR, 1); TCF=(NVAR, GAMMA, NTRT, NTOT));
P_VAL=(NVAR, 1, 0);
DO I=1 TO NVAR;
   PVAL(I, 1)=1-PROBCHI((1, I, NTRT-1);
END;
HP=P_VAL;
*OUTPUT RESULTS TO SAS DATA SET;
OUTPUT HP OUT=HPRENAME<KEEP=VARNAME(COL1=H COL2=P_VALUE)>
   ROWNAME=VARID;
FREE H P_VAL HP;
*GET CONTRAST COEFFICIENTS;
FETCH A DATA=INCD(DROP-CLASSES) COLNAME=CID; CONTID=CID'; FREE CID;
FREE A;
*CHECK FOR BAD CONTRASTS;
CONTRAST(ROWNAME(CID));
IF MAX(Abs(CONTRAST)>=0 THEN DO;
   NOTE SUM OF THE COEFFICIENTS NOT ZERO FOR AT LEAST ONE CONTRAST.;
   NOTE EXECUTION CONTINUING;
   D='COEF SUM';
   PRINT CONTRAST ROWNAME=CONTID COLNAME=A;
FREE D;
END;
FREE CONTRAST;
*COMPUTE CONTRAST VALUES, VARIANCES, STANDARDIZED VALUES AND P-VALUES;
NCONT=NROW(COEFF);
PHI=COEF*RBAR;
VAR(COEFF, COEFF)=RECIP(N)*JC(NCONT, 1)*TCF/C(CONSTANT);
FREE RBAR COEF TCF CONSTANT N;
RESULTS=JC(NCONT, NVAR, 2, 0);
DO I=1 TO NVAR;
   DO X=1 TO NCONT;
      X=(PHI(I, 1))/(VAR(I, X);  #STANDARDIZE CONTRAST;
      II=(X=1); NCONT=K;
      RESULTS((I, K)=1-X;
   RESULTS(I, 2)=1-PROBCHI(X, NTRT-1);
END;
FREE PHI VAR;
*SET UP IDENTIFIERS;
ID=VARIDJC(NCONT,11J(C(NVAR, 1))#CONTID;
*OUTPUT RESULTS TO SAS DATA SET;
OUTPUT RESULTS OUT=RESULTS<RENAME=(COL1=STDVALUE COL2=P_VALUE)>
   ROWNAME=VARID;
FREE ROWNAME ID RESULTS;
*PRINT H-TEST AND P-VALUES;
PROC PRINT DATA=HP; ID VARIABLE; FORMAT P_VALUE 6.4;
TITLE 'KRUSKAL-WALLIS H AND ASYMPTOTIC P-VALUES';
*PRINT STANDARDIZED CONTRAST VALUES AND P-VALUES;
DATA FINAL; MERGE ID RESULTS DROP ROW; FORMAT P_VALUE 5.4;
PROC PRINT; ID VARIABLE;
TITLE 'STANDARDIZED CONTRAST VALUES AND ASYMPTOTIC P-VALUES';

Figure 2 (Continued). Listing of NEWM.{