# Non-parametric Analysis of the Variance in SAS®

Hend Aljobaily; University of Northern Colorado

2018

## ABSTRACT

Comparing more than two groups has important applications within many fields. To analyze such an event, an analysis of the variance can be performed using ANOVA or MANOVA. However, these procedures have to assume normality of residuals. When researchers cannot determine the distribution of the response or cannot determine the parameters of the distribution, non-parametric methods would be used to perform necessary analyses. The Kruskal-Wallis test is a non-parametric test that can be used, when the normality of residuals cannot be assumed, to perform a one-way analysis of variance. This study is discussing methods of modeling non-parametric one-way analysis of variance (ANOVA and MANOVA) using different options in SAS®.

## INRODUCTION

Analysis of Variance (ANOVA) is a statistical method used to test differences among three or more group means by analyzing the variation among groups. The formula below represents the one-factor ANOVA model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Where $Y_{ij}$ represents the sampled observations, $\mu$ represents the grand mean, $\alpha_i$ represents the effect of group $i$, and $\varepsilon_{ij}$ represents the random error.

Similarly, Multivariate Analysis of Variance (MANOVA) is used to test the difference between three or more group means when several dependent variables exist. In other words, it is an ANOVA with multiple dependent variables where group mean vectors would be compared instead of group means. The formula below represents the one-factor MANOVA model:

$$\underline{Y_{ij}} = \underline{\mu} + \underline{\alpha_i} + \underline{\varepsilon_{ij}}$$

Where $\underline{Y_{ij}}$ represents the sampled observations vector of size $p \times 1$, $\underline{\mu}$ represents the grand means vector of size $p \times 1$, $\underline{\alpha_i}$ represents the effect of group $i$ vector of size $p \times 1$, and $\underline{\varepsilon_{ij}}$ represents the random error vector of size $p \times 1$.

To perform an ANOVA test, the following assumptions need to be satisfied first:

1- Normality of the residuals
2- Homogeneity of the group variances
3- Independence of the sampled observations

Additionally, to perform a MANOVA test, the following assumptions need to be satisfied (in addition to the assumptions of the ANOVA test) (French et al.,2018):

1- Absence of multivariate outliers

2- Absence of multicollinearity
3- Linearity among all pairs of dependent variables, all pairs of covariates, and all dependent variable-covariate pairs in each cell
4- Homogeneity of covariance matrices

In some situations, when the assumptions on ANOVA/MANOVA are violated, the regular parametric ANOVA method becomes invalid. Therefore, the nonparametric approach needs to be taken.

## NONPARAMETRIC ANALYSIS OF THE VARIANCE MODELS

### *ONE-WAY ANOVA*

The Kruskal-Wallis (KW) test is a non-parametric method for comparing the mean or the median of k-independent samples. It is roughly equivalent to a parametric one way ANOVA with the data replaced by their ranks. The Kruskal-Wallis test assumes that the samples drawn from the population are random, the observations are independent of each other, and the measurement scale for the dependent variable should be at least ordinal.

To perform the KW test, the following steps have to be taken:

1- Rank all sampled observation in all groups from 1 to N. Assign the average of a rank to any tied values.
2- Calculate the test statistic.
   a- When ties are present the formula is given by:

$$T = (N-1)\frac{\sum_{i=1}^{k} n_i(\bar{r}_{i.} - \bar{r})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} n_i\left(\bar{r}_{ij} - \bar{r}\right)^2}$$

   b- When no ties are present the formula is given by:

$$T = \frac{12}{N(N+1)}\sum_{i=1}^{k} n_i\,\bar{r}_{i.}^2 - 3(N+1)$$

Where $N$ represents the total number of observations, $n_i$ represents the number of observations in group $i$, $\bar{r}_{i.}$ represents the average rank of all observations in group $i$, $r_{ij}$ represents the rank of observation $j$ in group $i$, and $\bar{r}$ represents the average of all $r_{ij}$.

### *ONE-WAY MANOVA*

The Kruskal-Wallis test (KW) can be used to perform a one-way MANOVA as well. The same steps of KW applied to perform the non-parametric ANOVA can be applied to each of the $p$ variables in the design to perform a Multivariate Kruskal-Wallis test (MKW). In other words, multiple KW tests will be performed (He et al., 2017).

To perform the MKW test the following steps have to be taken:

1- Rank all sampled observation in each of the variables separatly from 1 to $n_i$. Assign the average of a rank to any tied values.
2- Calculate the test statistic.
   a- When ties are present the formula is given by:

$$W^2 = \sum_{i=1}^{k} n_i U_i' V^{-1} U_i$$

Where $U_i = \left(\bar{R}_{i.1} - m, \ldots, \bar{R}_{i.p} - m\right)$ which measures the distance between the mean vector of ranks for the $i$th group, $\bar{R}_{i.1} = \sum_{j=1}^{n_i} \frac{R_{ijk}}{n_i}$, $m = \frac{n+1}{2}$, and $V = \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (R_{ij} - m1_p)(R_{ij} - m1_p)'$ which estimated the pooled within-group covariance matrix (He et al., 2017).

## SAS® PROCEDURE

To perform the above analyses in SAS®, different procedures can be used, such as PROC NPARA1WAY and the macro KWMULT. Assuming the existence of dependent variables called DV1 and DV2 and the independent variables IV1, IV2, and IV3 in a dataset called Data.

### *ONE-WAY ANOVA*

PROC NPAR1WAY can be used to perform a one-way ANOVA, as shown below:

```
PROC NPAR1WAY DATA=Data;
        CLASS IV1;
        VAR DV1;
RUN;
```

The CLASS statement is used to define the grouping variable. VAR statement is used to define the variable of interest. The PROC NPAR1WAY performs a various number of non-parametric tests including the Kruskal-Wallis test (KW).

### *ONE-WAY MANOVA*

The KWMULT macro shown in Appendix A can be used to perform a one-way MANOVA for multivariate data (May & Johnson, 1997) as shown below:

```
%include 'macro file path\KWMULT.sas';

%KWMULT ( DATA = Data,
                GROUP = IV1,
                VARIATE = DV1 DV2,
                PRNTVEC = 0
                RUNPERM= 1);
```

The GROUP statement is used to define the grouping variable. The VARIATE statement is used to define the variables of interest (Dependent Variables). The PNTVEC statement is used to print the distribution of the statistic, where 1 is used to display the printout and 0 is used to suppress the printout. The RUNPERM statement is used to perform the permutation test where 1 is used to display the permutation test and 0 is used to suppress the permutation test. The KWMULT macro performs the same test as PROC NPAR1WAY for multivariate data including Multivariate Kruskal-Wallis test (MKW).

## CONCLUSION

When the normality assumption is violated, regular or parametric analysis of the variance cannot be applied. In this case, non-parametric analysis of the variance can be used for univariate and multivariate data which performs ANOVA and MANOVA on the ranks of observations instead of the original observations. Procedures such as PROC NPAR1WAY can be used to perform a one-way ANOVA. When dealing with a multivariate data, KWMULT macro can be used to perform a one-way MANOVA.

## LIMITATIONS

The non-parametric procedures for analyzing the variance can have some limitations. For example, non-parametric ANOVA and MANOVA models described in this paper can only be applied to one-factor models such as one-way ANOVA and one-way MANOVA. Two or more factor models need different models and procedures such as Monte Carlo simulations. Also, another limitation that could cause difficulties is that there is not a built-in procedure in SAS® to perform a non-parametric MANOVA. The KWMULT macro used when performing a non-parametric one-way MANOVA can be complicated for someone without extensive experience in SAS®.

## REFERENCES

French, A., Macedo, M., Poulsen, J., Waterson, T. and Yu, A. (2018). Multivariate Analysis of Variance (MANOVA). [online] Usersfsu.edu. Available at: http://userwww.sfsu.edu/efc/classes/biol710/manova/MANOVAnewest.pdf

He, F., Mazumdar, S., Tang, G., Bhatia, T., Anderson, S. J., Dew, M. A., … Reynolds, C. F. (2017). NONPARAMETRIC MANOVA APPROACHES FOR NON-NORMAL MULTIVARIATE OUTCOMES WITH MISSING VALUES. *Communications in Statistics: Theory and Methods*, 46(14), 7188–7200.

Lane, D. M. (n.d.). Introduction to Analysis of Variance. Available at: http://onlinestatbook.com/2/analysis_of_variance/intro.html

May, Warren & Johnson, William. (1997). A SASR macro for the multivariate extension of the Kruskal-Wallis test including multiple comparisons: Randomization and $\varkappa 2$ criteria. *Computational Statistics & Data Analysis*. 26. 239-250. 10.1016/S0167-9473(97)82107-X.

## APPENDIX A

```
%MACRO KWMULT(DATA=_last_,
              GROUP= ,
              VARIATE= ,
              PRNTVEC= ,
              RUNPERM= );

PROC GLM DATA = &DATA;
CLASS &GROUP;
MODEL &VARIATE = &GROUP;
MANOVA H = &GROUP;
MEANS &GROUP;
DATA NEW; SET &DATA;
KEEP &GROUP &VARIATE;
PROC RANK OUT = RANK;
PROC SORT DATA = RANK; BY &GROUP;
PROC PRINT;
PROC SUMMARY DATA = RANK;
CLASS &GROUP;
VAR &VARIATE;
OUTPUT OUT = NUM N = NUMS;
DATA NUMS2 (KEEP=NUMS);
SET NUM;
IF _TYPE_ = 0 THEN DELETE;

PROC IML;
RESET NONAME;
RESET NOLOG;
RESET NOCENTER;
USE NUMS2 VAR {NUMS};
READ ALL INTO NS;
NPERGRP = NS`;
CLOSE NUMS2;
USE RANK; SETIN RANK;
READ ALL INTO RANKS;
DATA = {&DATA};
GROUP = {&GROUP};
VARIATE = {&VARIATE};
VARIATS = NCOL(VARIATE);
RUNPERM={&RUNPERM};
GROUPS = NCOL(NPERGRP);
TOTLOBS = NROW(RANKS);
N = NPERGRP(|+|);
MEAN = (TOTLOBS + 1) / 2.0;
PRANKS = RANKS;
```

```
PRINT 'DATA FILE USED FOR ANALYSIS = ' DATA;
PRINT 'VARIATES USED              = ' VARIATE;
PRINT 'TOTAL NUMBER OF GROUPS = ' GROUPS
(|FORMAT = 5.0|);
PRINT 'TOTAL NUMBER OF VARIATES = ' VARIATS
(|FORMAT = 5.0|);
PRINT 'TOTAL NUMBER IN EACH GROUP = ' NPERGRP
(|FORMAT = 5.0|);
PRINT 'TOTAL NUMBER OF OBSERVATIONS = ' TOTLOBS
(|FORMAT = 5.0|);
PRINT 'MEAN RANK FOR EACH VARIATE = ' MEAN
(|FORMAT = 6.2|);

START GETSTAT;
        RBAR=J(GROUPS,VARIATS,0);
        U = J(GROUPS,VARIATS,0);
        V = J(VARIATS,VARIATS,0);
        INDEX = 0;
        DO I = 1 TO GROUPS;
                BEGINS = INDEX + 1;
                ENDS = BEGINS + NPERGRP[,I] - 1;
                INDEX = ENDS;
RBAR(|I,|)=(PRANKS[BEGINS:ENDS,][+,])/NPERGRP[,I];
                U(|I,|) = RBAR(|I,|)-MEAN;
                DO J = BEGINS TO ENDS;
                V = V + (((PRANKS[J,] - MEAN)`)*(PRANKS[J,]-MEAN))/(TOTLOBS-1);
                END;
        END;
UVU = U*INV(V)*U`;
                KWSTAT = TRACE(NS#(UVU));
        FINISH;

START PERMUTE;
                A = J(GROUPS+1,TOTLOBS,0);
                PERM = J(1,TOTLOBS,0);
                PRANKS = J(TOTLOBS,VARIATS,0);
N=TOTLOBS;
                ELL = TOTLOBS;
                K = GROUPS;

DO I = 1 TO (K-l);
                A[I,l:NPERGRP[,I]] = I;
                A[I,(NPERGRP[,I]+1):ELL] = I + 1;
        ELL = ELL - NPERGRP[,I];
END;
KOUNTl = 0;
```

```
OFLAG=0;
DO UNTIL (OFLAG=1);
        I = K -1;
        T = NPERGRP[,K] + NPERGRP[,I];
        IF KOUNTl ^= 0 THEN DO;
                L7 = 0;
                DO WHILE(L7=0);
                        DO J = 1 TO T - 1;
                                IF ((A[I,J] = I) & (A[I,J+1] = (I + 1))) THEN DO;
                                        LIMIT = J - 2;
                                        A[I,J] = A[I,J+1];
                                        A[I,J+1] = A[I,J] - 1;
                                        IF (LIMIT <= 0) THEN GOTO L14;
                                        ELSE GOTO L12;
                                END;
L8: ;           END;
L9:                     IF (T ^= 1) THEN DO;
                                DO J = 1 TO (T/2);
                                        ITEMP = A[I,J];
                                        A[I,J] = A[I,T-J+1];
                                        A[I,T-J+1] = ITEMP;
                                END;
                        END;

L10:    I= I - 1;
                  IF (I <= 0) THEN GOTO L100;
                  T = T + NPERGRP[,I];
                END;

L12:    IFLAG = 0;
                        DO J = 1 TO LIMIT;
                                IF ((A[I,J] = (I + 1)) & (A[I,J+1] = I)) THEN DO;
                                        A[I,J] = A[I,J+1];
                                        A[I,J+1] = A[I,J] + 1;
                                        IFLAG = 1;
                                END;
                        END;
                        IF (IFLAG = 1) THEN GOTO L12;
                END;

L14:    DO J = 1 TO N;
                                A[K,J] = A[1,J];
                        END;

L15:    IF (K ^= 2) THEN DO;
                                DO ELL = 2 TO K-1;
```

```
                                M = 1;
                                DO J = 1 TO N;
                                        IF A[K,J] = ELL THEN DO;
                                        A[K,J] = A[ELL,M];
                                        M = M+1;
                                  END;
                                END;
                        END;
                END;

Ll8:    LL = 1;
                  DO II = 1 TO K;
                        DO JJ = 1 TO N;
                                IF A[K,JJ]=II THEN DO;
                                        PERM[,LL]=JJ;
                                        LL = LL + 1;
                                END;
                        END;
                END;
                DO II = 1 TO VARIATS;
                 DO JJ = 1 TO TOTLOBS;
                        INDEX = PERM[,JJ];
                        PRANKS[JJ,II] = RANKS[INDEX,II];
                  END;
                END;
                KOUNTl = KOUNTl + 1;
                RUN GETSTAT;
                IF KWSTAT >= (FSTAT-0.0000001) THEN PVAL = PVAL+1;
                IF KOUNTl = 1 THEN DO;
                        CREATE OUTVEC VAR {KWSTAT};
                        APPEND VAR {KWSTAT};
                END;
                ELSE DO;
                        APPEND VAR {KWSTAT};
                END;
        END;


L100:   PRINT 'TOTAL NUMBER OF PERMUTATIONS = ' KOUNT1;
                                FINISH;
                        START;
                                PVAL = 0;
                                RUN GETSTAT;
                                DF = VARIATS*(GROUPS - 1);
                                PVALCHI = 1 - PROBCHI(KWSTAT,DF);
```

```
                                CREATE MULTOUT VAR {RBAR V GROUPS VARIATS};
                                APPEND VAR {RBAR V GROUPS      VARIATS};
                                PRINT,,;
                                PRINT 'VECTOR OF MEAN RANKS - RBAR';
                                PRINT RBAR (|FORMAT=5.4|);
                                PRINT 'COVARIANCE MATRIX V';
                                PRINT V (|FORMAT = 10.6|);
                                FSTAT = KWSTAT;
                                PRINT 'VALUE OF THE KRUSKAL-WALLIS TEST
STATISTIC = ' FSTAT;
                                PRINT 'DEGREES OF FREEDOM FOR OVERALL TEST       =
' DF;
                                PRINT 'P-VALUE BASED ON CHI-SQUARE
APPROXIMATION  = ' PVALCHI;
                                IF RUNPERM=1 THEN DO;
                                RUN PERMUTE;
                                PVAL = PVAL/KOUNT1 ;
                                PRINT 'P-VALUE BASED ON THE EXACT DISTRIBUTION =
' PVAL;

                                END;
                                FINISH;
                        RUN;
                        QUIT;
%IF &RUNPERM = 1 %THEN %DO;
%IF &PRNTVEC = 1 %THEN %DO;
        PROC SORT DATA=OUTVEC; BY KWSTAT;
        PROC FREQ DATA=OUTVEC;
. TABLES KWSTAT;
%END;
%END;
%MEND;
```

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Hend Aljobaily
University of Northern Colorado
hend.aljobaily@unco.edu