ANOVA

PURPOSE

Carry out an Analysis of Variance.

DESCRIPTION

Analysis of Variance is a data analysis technique for examining the significance of the factors (= independent variables) in a multi-factor model. The number of factors must be between 1 and 5 inclusive. Each factor then has a certain number of values it can have (referred to as the levels of a factor). The number of levels does not have to be the same for each factor. Each factor and level combination is a cell (the number of cells is the product of the number of levels in each factor). Balanced designs are those in which each cell has an equal number of observations and unbalanced designs are those in which the number of observations can vary between cells. The DATAPLOT ANOVA command only works with balanced designs (an error message is printed if an unbalanced design is detected). The number of arguments specifies whether a 1-factor, a 2-factor, or higher ANOVA is carried out.

SYNTAX

ANOVA <y> <x1> ... <x5> <SUBSET/EXCEPT/FOR qualification>

where <y> is the response (= dependent) variable;

<x1> ... <x5> is a list of at least 1 and no more than 5 independent variables;

and where the <SUBSET/EXCEPT/FOR qualification> is optional.

EXAMPLES

ANOVA Y X1

ANOVA Y X1 X2 X3

NOTE 1

There are two main approaches to analysis of variance in the statistics literature: the cell means model and the factor effects model. These 2 models are mathematically equivalent. For simplicity, these are demonstrated for the 2-factor case.

The cell means model takes the following form:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \quad \textbf{(EQ 3-38)}$$

where *i* represents the level of factor 1, *j* represents the level of factor 2, and *k* represents the observation number in the *ij*th cell. That is, the response variable is modeled as a cell mean plus an error term. The errors are assumed to be independent, normally distributed, and with a constant variance. The sample output for program example 1 below shows a column labeled MEAN. This is the sample cell mean and it is an estimate of the μ_{ij} terms in the above model.

The factor effects model takes the following form:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad \textbf{(EQ 3-39)}$$

where *i*, *j*, and *k* are the same as for the cell means model. The μ term is the overall mean. This form models the response as a grand mean plus a factor 1 effect plus a factor 2 effect plus an error term. Again, the errors are assumed to be independent, normally distributed, and with a constant variance. The sample output shows a column labeled EFFECT. These are the sample estimates of the α and β factor effects. More complicated versions of this model also include an interaction term for the factor effects. However, the DATAPLOT ANOVA does not calculate interaction terms.

Comparable models can be written for the 1-factor or the higher factor cases.

NOTE 2

ANOVA saves the residuals in the variable RES and the predicted values in the variable PRED. These can be used to generate various diagnostic plots. Specifically, the residuals can be plotted to test the normality, independence, and constant variance assumptions. Typically, some type of transformation (such as taking the log or reciprocal of the response variable) is required if the assumptions are not met. The Neter, Wasserman, and Kunter text in the REFERENCE gives guidance as to what type of transformations are often required. The LET command can be used to generate any desired transformations.

For the cell means model, the predicted values are the sample cell means. For the factor effects model, the predicted values are the grand mean plus the factor 1 effect plus the factor 2 effect. These generate the same predicted values. In either case, the residuals are the observed values minus the corresponding predicted value.

NOTE 3

ANOVA problems can be formulated as regression problems. This allows the DATAPLOT FIT command to be used to analyze the ANOVA problem. Specifically, unbalanced designs (i.e., unequal number of observations in the cells) can be analyzed. In addition, most designed experiments do not include all possible factor and level combinations. These types of designed experiments can be analyzed with the FIT command (this is demonstrated with several macro files in the DATAPLOT reference catalog). Using FIT is also required if you want to include interaction terms in your model.

The texts listed in the REFERENCE section give detailed examples of formulating ANOVA problems as regression problems.

NOTE 4

Although DATAPLOT does not generate the traditional ANOVA table, the relevant quantities are easy to derive. After an ANOVA command, the following code calculates the error sum of squares (SSE), the total sum of squares (SSTO), the treatment sum of squares (SSTR), and the corresponding mean squares (MSE, MSTR) for the cell means model. The variables PRED and RES contain the fitted values and the residuals from the ANOVA, respectively. The parameters RESSD and REPDF are the residual standard deviation and the residual degrees of freedom, respectively. The PRED, RES, RESSD, and RESDF quantities are automatically created and saved whenever a DATAPLOT ANOVA command is performed. The computed parameters GMEAN and DF are the overall mean (of the dependent variable) and the treatment sum of squares degrees of freedom, respectively.

```
LET SSE = RES**2
LET MSE = RESSD**2
LET GMEAN = MEAN Y
LET SSTO = (PRED - GMEAN)**2
LET SSTR = SSTO - SSE
LET N = SIZE Y
LET DF = N - REPDF
LET MSTR = SSTR/DF
```

In addition, the treatment sum of squares are usually broken down into sums of squares for each factor. Although DATAPLOT does not save these values in a form that can be retrieved automatically, it can be obtained from the values in the printout. Specifically, the residual standard deviation printed in the “MODEL” section and the number of levels for each factor in the “TESTING” section can be used. For example, manually enter the values from the output for the first program example below as follows:

```
LET RSD = DATA 10.1321 10.8009 10.7281 10.5422 10.1166
LET FACTDF = DATA 2 3 2 2
```

Then the following commands can be used to compute the treatment sum of squares and the treatment mean squares for the individual factors:

```
LET FACTDF = FACTDF - 1
LET MSFACT = RSD**2
LET FACTSS = FACTDF*MSFACT
```

The F tests in the testing section are derived from these mean squares (specifically, it is MSFACT/MSE).

NOTE 5

Various formal intervals and tests can be computed for the ANOVA problem. For example, confidence levels for the various factor level means are often desired. In addition, Tukey, Scheffe, or Bonferroni methods for computing comparisons of factor level means are often desired. Although DATAPLOT does not generate any of these directly, they can be derived in a straightforward manner from the ANOVA printout. See the Neter, Wasserman, and Kutner text for guidance (chapter 15 for one factor, chapter 19.2 for two factor, and chapter 22 for more than 2 factors).

NOTE 6

The second program example demonstrates how to perform a graphical analysis of variance using DATAPLOT. The block plot and the various DEX plots can also be useful in ANOVA problems. The box plot is a useful method for displaying 1-factor ANOVA problems.

NOTE 7

Median polish is a robust method (based on medians rather than means) for analyzing ANOVA problems. See the documentation for MEDIAN POLISH for details.

The YATES ANALYSIS command can be used to analyze Yates designs.

The Kruskal-Wallis test is a non-parametric test for a one-way ANOVA problem. This method does not require the normality assumptions of the standard ANOVA model. Although DATAPLOT does not perform this test directly, it is straightforward to do and is demonstrated in the third program example.

DEFAULT

None

SYNONYMS

ANALYSIS OF VARIANCE is a synonym for ANOVA.

RELATED COMMANDS

MEDIAN POLISH	=	Carries out a robust ANOVA.
YATES ANALYSIS	=	Analyze a Yates design.
BLOCK PLOT	=	Generate a block plot.
BOX PLOT	=	Generate a box plot.
DEX SCATTER PLOT	=	Generates a dex scatter plot.
DEX ... PLOT	=	Generates a dex plot for a statistic.
DEX ... EFFECTS PLOT	=	Generates a dex effects plot for a statistic.
T TEST	=	Carries out a t test.
PLOT	=	Plots (e.g., residuals and GANOVA).

REFERENCE

"Applied Linear Statistical Models," 3rd ed., Neter, Wasserman, and Kunter, 1990, Irwin.

"Applied Regression Analysis," 2nd ed., Draper and Smith, John Wiley, 1981.

APPLICATIONS

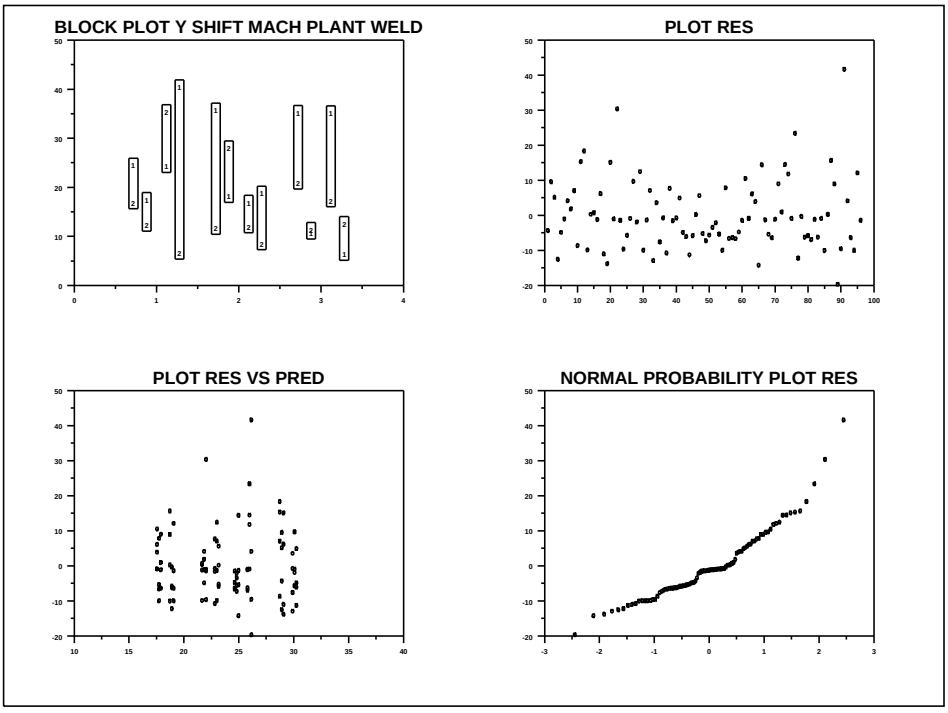
Analysis of Variance

IMPLEMENTATION DATE

Pre-1987

PROGRAM 1

```
. THIS IS DATAPLOT DATA FILE SHEESLEY.DAT (RAW DATA)
. LIGHT BULB LEAD WIRE WELD PROCESS COMPARISON
. JOHN SHEESLEY (GE) ARTICLE IN
. EXPERIMENTS IN INDUSTRY (ED. BY SNEE, HARE, TROUT), PAGES 54-57
. NUMBER OF OBSERVATIONS = 96
. NUMBER OF VARIABLES PER LINE IMAGE = 5
. ORDER OF VARIABLES ON A LINE IMAGE--
. RESPONSE = AVERAGE NUMBER OF WELDED LEAD WIRES MISSED PER HOUR
. FACTOR 1 = WELDING PROCESS (2 LEVELS) (PRIMARY)
. FACTOR 2 = SHIFT (3 LEVELS)
. FACTOR 3 = MACHINE (2 LEVELS)
. FACTOR 4 = PLANT (2 LEVELS)
. FACTOR 5 = REPLICATION (4 LEVELS) (A RANDOM FACTOR)
SKIP 25
READ SHEESLEY.DAT Y WELD SHIFT MACH PLANT REP
ANOVA Y WELD SHIFT MACH PLANT
MULTIPLOT 2 2; MULTIPLOT CORNER COORDINATES 0 0 100 100
TITLE AUTOMATIC
CHARACTER 1 2
LINES BLANK BLANK
BLOCK PLOT Y SHIFT MACH PLANT WELD
CHARACTER CIRCLE
CHARACTER SIZE 1.0
LINES BLANK
PLOT RES
PLOT RES VS PRED
NORMAL PROBABILITY PLOT RES
END OF MULTIPLOT
```



The following alphanumeric output is generated:

```
                 **********************************
                 **********************************
                 **  4-WAY ANALYSIS OF VARIANCE  **
                 **********************************
                 **********************************

      NUMBER OF OBSERVATIONS           =       96
      NUMBER OF FACTORS                =        4
      NUMBER OF LEVELS FOR FACTOR 1    =        2
      NUMBER OF LEVELS FOR FACTOR 2    =        3
      NUMBER OF LEVELS FOR FACTOR 3    =        2
      NUMBER OF LEVELS FOR FACTOR 4    =        2
      RESIDUAL    STANDARD DEVIATION   =    0.10116604805E+02
      RESIDUAL    DEGREES OF FREEDOM   =       90
      REPLICATION CASE
      REPLICATION STANDARD DEVIATION   =    0.10128160477E+02
      REPLICATION DEGREES OF FREEDOM   =       72
      NUMBER OF DISTINCT CELLS         =       24

                      ****************
                      *  ESTIMATION  *
                      ****************

      GRAND MEAN                       =    0.23897912979E+02
      GRAND STANDARD DEVIATION         =    0.10687572479E+02


             LEVEL-ID      NI      MEAN      EFFECT     SD(EFFECT)
-----------------------------------------------------------------------
FACTOR  1--    1.00000     48.   27.43542    3.53751    1.03252
         --    2.00000     48.   20.36042   -3.53749    1.03252

FACTOR  2--    1.00000     32.   23.90313    0.00521    1.46021
         --    2.00000     32.   23.72188   -0.17603    1.46021
         --    3.00000     32.   24.06875    0.17084    1.46021

FACTOR  3--    1.00000     48.   23.31458   -0.58333    1.03252
         --    2.00000     48.   24.48125    0.58334    1.03252

FACTOR  4--    1.00000     48.   25.95000    2.05209    1.03252
         --    2.00000     48.   21.84583   -2.05208    1.03252



        MODEL               RESIDUAL STANDARD DEVIATION
-----------------------------------------------------
CONSTANT             ONLY--       10.6875724792
CONSTANT & FACTOR  1 ONLY--       10.1320877075
CONSTANT & FACTOR  2 ONLY--       10.8009223938
CONSTANT & FACTOR  3 ONLY--       10.7280855179
CONSTANT & FACTOR  4 ONLY--       10.5422353745
CONSTANT & ALL 4 FACTORS --       10.1166048050
```

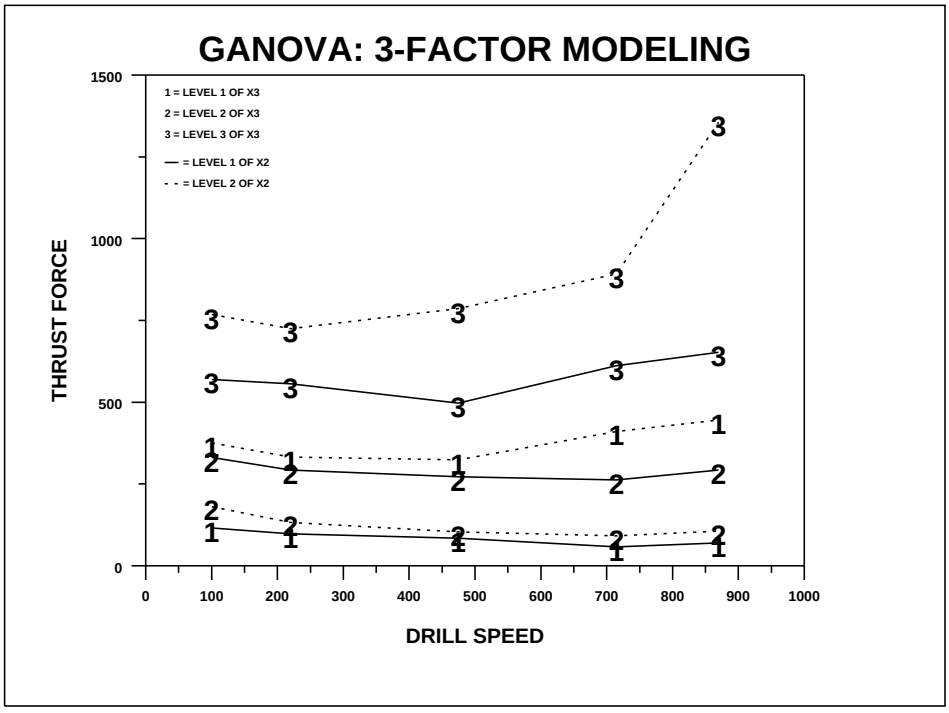
```
                          *************
                          *  TESTING  *
                          *************

             NUM. LEVELS    F STAT.       F CDF
-------------------------------------------------
FACTOR  1--       2      11.73801422119  99.908%
FACTOR  2--       3       0.00941137969   0.937%
FACTOR  3--       2       0.31917828321  42.649%
FACTOR  4--       2       3.94994902611  95.009%

      RESIDUAL    STANDARD DEVIATION =       10.11660480499
      RESIDUAL    DEGREES OF FREEDOM =            90
      REPLICATION STANDARD DEVIATION =       10.12816047668
      REPLICATION DEGREES OF FREEDOM =            72
      LACK OF FIT F RATIO =      0.9886 = THE   51.7679% POINT OF THE
      F DISTRIBUTION WITH     18 AND     72 DEGREES OF FREEDOM
```

PROGRAM 2 (Graphical ANOVA)

```
. THRUST FORCE OF A DRILL
. REFERENCE--HAMAKER, JISI, 1971, PAGE 354-358, 1971
. ORDER OF VARIABLES ON A CARD--
. 1. THRUST FORCE (RESPONSE VARIABLE)
. 2. DRILL SPEED (5 LEVELS)
. 3. MATERIAL (2 LEVELS)
. 4. FEEDS (3 LEVELS)
. 5. MATERIAL X FEED SUBSET DEFINITION VARIABLE (6 LEVELS)
SKIP 25
READ HAMAKER.DAT Y X1 X2 X3 TAG
.
TITLE GANOVA: 3-FACTOR MODELING
Y1LABEL THRUST FORCE
X1LABEL DRILL SPEED
LEGEND SIZE 1.5
LEGEND 1 1 = LEVEL 1 OF X3; LEGEND 1 COORDINATES 17 87
LEGEND 2 2 = LEVEL 2 OF X3; LEGEND 2 COORDINATES 17 84
LEGEND 3 3 = LEVEL 3 OF X3; LEGEND 3 COORDINATES 17 81
SEGMENT 1 COORDINATES 17 77.5 18.5 77.5
LEGEND 4 = LEVEL 1 OF X2; LEGEND 4 COORDINATES 19 77
SEGMENT 2 COORDINATES 17 74.5 18.5 74.5; SEGMENT 2 PATTERN DOTTED
LEGEND 5 = LEVEL 2 OF X2; LEGEND 5 COORDINATES 19 74
.
CHARACTERS 1 2 3 2 1 3
CHARACTER SIZE 4 ALL
LINES SOLID SOLID SOLID DOT DOT DOT
YMAXIMUM 1500
XLIMITS 0 1000
PLOT Y X1 TAG
```



PROGRAM 3

```
. Perform a Kruskal-Wallis non-parametric 1-way ANOVA
. Data from "Probability and Statistics for Engineers and Scientists" by Walpole and Myers.
LET P = 3
LET X1 = DATA 24.0 16.7 22.8 19.8 18.9
LET X2 = DATA 23.2 19.8 18.1 17.6 20.2 17.8
LET X3 = DATA 18.4 19.1 17.3 17.3 19.7 18.9 18.8 19.3
.
LOOP FOR K = 1 1 P
    LET N^K = SIZE X^K
    LET TAG^K = K FOR I = 1 1 N^K
END OF LOOP
LOOP FOR L = 2 1 P
    LET TEMP = X^L
    EXTEND X1 TEMP
    LET TEMP = TAG^L
    EXTEND TAG1 TEMP
    DELETE TEMP
END OF LOOP
.
LET N = SIZE X1
LET R = RANK X1
LET SUM1 = 0
LOOP FOR K = 1 1 P
    LET R^K = SUM R SUBSET TAG1 = K
    LET SUM1 = SUM1 + (R^K)**2/N^K
END OF LOOP
.
LET H = SUM1*(12/(N*(N+1))) - 3*(N+1)
LET ALPHA = 0.95
LET DF = P - 1
LET CRITICAL = CHSPPF(ALPHA,DF)
.
PRINT "H0: ^P INDEPENDENT SAMPLES ARE FROM IDENTICAL POPULATIONS"
PRINT "HA: ^P INDEPENDNET SAMPLES ARE FROM DIFFERENT POPULATIONS"
PRINT "KRUSKAL-WALLIS H STATISTIC = ^H"
PRINT "CHI-SQUARE CRITICAL VALUE = ^CRITICAL"
IF H <= CRITICAL
    PRINT "ACCEPT NULL HYPOTHESIS"
END OF IF
IF H > CRITICAL
    PRINT "REJECT NULL HYPOTHESIS"
END OF IF
```

The following output is generated:

```
H0: 3 INDEPENDENT SAMPLES ARE FROM IDENTICAL POPULATIONS
HA: 3 INDEPENDNET SAMPLES ARE FROM DIFFERENT POPULATIONS
KRUSKAL-WALLIS H STATISTIC = 1.658619
CHI-SQUARE CRITICAL VALUE = 5.991465
ACCEPT NULL HYPOTHESIS
```