MEANS

Cases are cross-classified on the basis of multiple independent variables, and for each cell of the resulting cross-classification, basic statistics are calculated for a dependent variable.

Notation

The following notation is used throughout this chapter unless otherwise stated:

X_{ip}	Value for the *p*th independent variable for case *i*
Y_i	Value for the dependent variable for case *i*
w_i	Weight for case *i*
P	Number of independent variables
N	Number of cases

Statistics

For each value of the first independent variable $\left(X_1\right)$, for each value of the pair $\left(X_1, X_2\right)$, for the triple $\left(X_1, X_2, X_3\right)$, and similarly for the *P*-tuple $\left(X_1, X_2, \ldots, X_P\right)$, the following are computed:

Sum of Case Weights for the Cell

$$W = \sum_{i=1}^{N} w_i l_i$$

where $l_i = 1$ if the *i*th case is in the cell, $l_i = 0$ otherwise.

The Sum and Corrected Sum of Squares

$$SMY = \sum_{i=1}^{N} w_i l_i Y_i$$

$$SSY = \sum_{i=1}^{N} w_i l_i Y_i^2$$

$$CSS = SSY - SMY^2 / W$$

The Mean

$$\overline{Y} = \frac{\sum_{i=1}^{N} w_i l_i Y_i}{W}$$

Harmonic mean

$$\overline{Y}_h = \frac{\sum_{i=1}^{N} w_i}{\sum_{i=1}^{N} w_i y_i^{-1}}$$

Both summations are over cases with positive w_i values.

Geometric mean

$$\overline{Y}_g = \left(\prod_{i=1}^{N} y_i^{\,w_i} \right)^{1/W}$$

The product is taken over cases with positive w_i values.

Variance

$$S^2 = \frac{CSS}{W-1}$$

Standard Deviation

$$S = \sqrt{\text{variance}}$$

Standard Error of the Mean

$$SEM = \frac{S}{\sqrt{W}}$$

Skewness (computed if $W \geq 3$ and $S^2 > 0$), and its standard error

$$g_1 = \frac{WM_3}{(W-1)(W-2)S^3} \qquad se(g_1) = \sqrt{\frac{6W(W-1)}{(W-2)(W+1)(W+3)}}$$

Kurtosis (computed if $W \geq 4$ and $S^2 > 0$), and its standard error

$$g_2 = \frac{W(W+1)M_4 - 3(W-1)M_2^2}{(W-1)(W-2)(W-3)S^4} \qquad se(g_2) = \sqrt{\frac{4\left(W^2-1\right)se(g_1)^2}{(W-3)(W+5)}}$$

Minimum

$$\min_i X_i$$

Maximum

$$\max_i X_i$$

Range

Maximum – Minimum

Percent of Total N

For each category *j* of the independent variable,

$$\%TotN_j = \left(\frac{\sum_{i=1}^{N} w_i l_i}{W} \right) \times 100$$

where $l_i = 1$ if the *i*th case is in the *j*th category, $l_i = 0$ otherwise.

Percent of Total Sum

For each category *j* of the independent variable,

$$\%TotSum_j = \left(\frac{\sum_{i=1}^{N} w_i l_i Y_i}{W} \right) \times 100$$

where $l_i = 1$ if the *i*th case is in the *j*th category, $l_i = 0$ otherwise.

Median

Find the first score interval (*x2*) containing more than *t* cases.

$$\text{median} = \begin{cases} x_2 & \text{if } t - cp_1 \geq 100/W \\ \{1 - [(W+1)/2 - cc_1]\} x_1 & \text{if } t - cp_1 < 100/W \\ \quad + [(W+1)/2 - cc_1] x_2 & \end{cases}$$

where

$t = (W+1)/2$

$cp_1 < t < cp_2$

x_1 and x_2 are the values corresponding to cp_1 and cp_2, respectively

cc_1 is the cumulative frequency up to x_1

cp_1 is the cumulative percent up to x_1

Grouped Median

The formulas for the grouped median can be found in “Appendix 8: Grouped Percentiles” (*app08_grouped_percentiles.pdf*).

ANOVA and Test for Linearity

If the analysis of variance table or test for linearity are requested, only the first independent variable is used. Assume it takes on J distinct values (groups). The previously described statistics are calculated and printed for each group separately, as well as for all cases pooled. Symbols subscripted from 1 to J will denote group statistics, unsubscripted the total. Thus for group j,

- SMY_j is the sum of the dependent variable.

  and

- X_j the value of the independent variable. Note that the standard deviation and sum of squares printed in the last row of the summary table are pooled within group values.

Analysis of Variance

Source	Sum of Squares	df
Between Groups	Total-Within Groups	$J-1$
Regression	$\dfrac{\left(\sum_{j=1}^{J} X_j SMY_j - \left(\sum_{j=1}^{J} w_j X_j\right)\left(\sum_{j=1}^{J} SMY_j\right)\Big/ W\right)^2}{\sum_{j=1}^{J} w_j X_j^2 - \left(\sum_{j=1}^{J} w_j X_j\right)^2 \Big/ W}$	1
Deviation from Regression	Between-Regression	$J-2$
Within Groups	$\sum_{j=1}^{J} CSS_j$	$W-J$
Total	$\sum_{j=1}^{J} SSY_j - \left(\sum_{j=1}^{J} SMY_j\right)^2 \Big/ W$	$W-1$

The mean squares are calculated by dividing each sum of squares by its degrees of freedom. The F ratios are the mean squares for each source divided by the within groups mean square. The significance level for the F is from the F distribution with the degrees of freedom for the numerator and denominator mean squares. If there is only one group the ANOVA is not done; if there are fewer than three groups or the independent variable is a string variable, the test for linearity is not done.

Correlation Coefficient

$$r = \frac{\sum_{j=1}^{J} X_j SMY_j - \left(\sum_{j=1}^{J} W_j X_j\right) SMY \Big/ W}{\sqrt{\left(\sum_{j=1}^{J} W_j X_j^2 - \left(\sum_{j=1}^{J} W_j X_j\right)^2 \Big/ W\right)\left(SSY - SMY^2 / W\right)}}$$

Eta

$$(eta)^2 = \frac{\text{Sum of Squares Between Groups}}{\text{Total Sum of Squares}}$$

References

Blalock, H. M. 1972. *Social statistics*. New York: McGraw-Hill.

Bliss, C. I. 1967. *Statistics in biology*, Volume 1. New York: McGraw-Hill.

Hays, W. L. 1973. *Statistics for the social sciences*. New York: Holt, Rinehart and Winston.