PROC ITEM - AN ITEM ANALYSIS PROCEDURE

Cynthia Deitz
Texas Instruments, Incorporated
Dallas, Texas

David A. Smith
Department of Statistics & Computer Science
West Virginia University
Morgantown, West Virginia

ABSTRACT

The Procedure ITEM analyzes item data from multiple-response tests which have a single correct response for each item. A KEY statement which indicates the correct response is used to compute a score for each individual. The output consists of four parts:

1) Total test statistics which includes the mean, median, standard deviation, homogeneity reliability KR-21, standard error of measurement, possible low and high score, obtained low and high score, and number of scores with counts of blank, valid, and invalid scores.

2) A test frequency distribution is presented which indicates the raw score, standard score, percentile rank, percentage of people in total group who received the given score, frequency, cumulative frequency, and a histogram portraying the frequency of the scores at each score value.

3) An item quintile table which compares the item response versus the total score distribution for each item. The percentage of responses to each alternative by fifths of the distribution is plotted.

4) Item statistics which can help determine which items are good and which need improvement or deletion from the examination. A matrix of responses by fifths is printed which shows the frequency of responses within each fifth. Also presented for each alternative is the proportion of responses and a point-biserial correlation.

GENERAL DESCRIPTION

This procedure analyzes item data from multiple response tests which have a single correct response for each item. A "key" is used to compute a score for each individual by counting the number of correct responses.

Information based on the computed test score includes total test statistics, test frequency distribution, percents, percentiles, and standard score. Each item is then presented as a graph of the percent or correct response by fifths and also as a matrix of the responses by fifths. In addition to the matrix, for each response is calculated the proportion of people responding and a point-biserial correlation.

FORMULAS USED

TEST STATISTICS

N = number of respondents in sample
X = sum of the correct responses for each respondent

MEAN = SUM X/N

MEDIAN = A = 1/2 + (CUMA - N/2)/FREQA
where
FREQA = frequency of scores at A
CUMA = cumulative frequency of scores starting with the highest and summing down to and including A

A = lowest score that one half or more of the distribution obtained

SIGMA (STANDARD DEVIATION) =
SQR((N*SUM(X**2) - (SUM(X))**2)/N(N-1))

KR-21 (homogeneity reliability) =
[k/(k-1)]*[1-(MEAN**2)/SIGMA**2]
where
K = number of items

STANDARD ERROR OF MEASUREMENT =
SIGMA*SQR(1-KR_21)

STANDARD SCORE =
100*(X-MEAN)/SIGMA + 500
PERCENTILE = \( \text{SUM(CUMB)}/N \)

where

\( \text{CUMB} \) = cumulative frequency of scores starting with the lowest score and summing up to and including the frequency of the score the percentile represents.

Adjustments have been made at upper and lower limits to prevent percentiles less than 1 or greater than 99.

ITEM STATISTICS

\[ \text{PROP} = \text{NP}/N \]

where

NP = number of respondents responding to an alternative

\[ \text{RPDI (point-biserial correlation)} = ((\text{NP} - \text{MEAN})/\text{SIGMA}) \times \sqrt{\text{NP}/\text{NG}} \]

where

MP = mean score of NP

NP = number responding to the alternative

NG = number not responding to the alternative

PROCEDURE STATEMENTS

REQUIRED:

PROC ITEM NR= NG;

where

NR = number of valid responses

NG = number of test items

RESPONSE string;

where

string = string of the valid responses(length=NR)

KEY string;

where

string = string of the correct responses(length=NG)

OPTIONAL:

VAR list of variables which are test items;

TITLE title of output;

OPTIONS list of options;

PLOT parameter on PROC statement produces item plots and matrices.

ERRORS ON SAS STATEMENTS:

1. Length of response string not equal to number of responses specified.
2. Response statement must precede key statement.
3. Length of key string not equal to number of questions specified.
4. Number of responses not specified on PROC statement.
5. Number of questions not specified on PROC statement.

INTERPRETATION OF THE OUTPUT

The output is divided into three parts: 1) Test Statistics, 2) Distribution Characteristics, and 3) Item Statistics.

The Test Statistics include the number of items, mean or average of all scores, median which is the point that divides the raw score distribution such that 50 of the scores are above and 50 are below the median. The standard deviation is a measure of the variability of the scores from the mean. The reliability is a measure of the relationship between test items for the group tested. The standard error indicates the accuracy of the measure in that it reflects how close the the observed score is to the true score. Also presented are possible low and high scores, obtained low and high scores, and the number of blank and invalid scores.

The frequency distribution orders the raw scores from high to low. Also presented are standard score (mean=500, standard deviation=100), percentile, percent, frequency, and cumulative frequency. A bar histogram indicates the distribution of the scores at each value.

The item statistics are useful for observing how well the items performed on a given sample and indicates which items could be improved. The difficulty of an item is the proportion (PROP) of subjects who passed it. Ebel (1965) and Lingquist (1951) recommend using items with difficulty ranging from .30 to .70 in order to obtain differential information. The discrimination power of an item is measured by the point-biserial correlation (RPDI).
The relationship between the item score and the total test statistic can be obtained from this score. The highest RPBI's would represent the most discriminating items. Since a dichotomous variable is compared to a continuous one and since teacher-made tests rarely achieve RPBI's above 50, it is rare for a high RPBI to be obtained. Therefore, good items which do discriminate should have RPBI's which fall in the range of .30 to .70.

Also presented is a plot (scattergram) of the correct response by fifth and a matrix of the tabulation of the same. This tabulation along with the proportion of the group responding to each alternative provides necessary information for determining which items are good and which are weak. A good item would be one with a high proportion passing it and a high RPBI. An item which has a negative RPBI would indicate that the students with lower scores are passing it and the ones with higher scores are not.

ADDITIONAL OUTPUT INFORMATION

If the sample size is large, the scattergram in the frequency distribution section will represent a multiple of scores with one %. This is calculated with respect to the largest frequency.

The nocenter option is accounted for as are the linesize and pagesize options. The linesize option will split the frequency and item statistics into 2 pages if the linesize is less than 62.

The mode of the variables is alphanumeric; therefore, if no variable list is specified the standard SAS defaults will be in effect.

REFERENCES


SAMPLE SAS STATEMENTS
PROC ITEM NR=5 NG=10 DATA=ALL;
RESPONSE 01234;
KEY 0123423210;

PROC ITEM NR=5 NG=10 DATA=ALL;
RESPONSE 01234;
KEY 0123423110;
OPTIONS LS=81;

PROC ITEM NR=5 NG=10;
RESPONSE 01234;
KEY 0123423123;
VAR G1-G10.

SAMPLE OUTPUT

ITEM ANALYSIS
SUMMARY OF TEST STATISTICS

<table>
<thead>
<tr>
<th>NUMBER OF QUESTIONS</th>
<th>MEAN SCORE</th>
<th>MEDIAN SCORE</th>
<th>STANDARD DEVIATION</th>
<th>RELIABILITY (KR-21)</th>
<th>S.E. OF MEASUREMENT</th>
<th>POSSIBLE LOW SCORE</th>
<th>POSSIBLE HIGH SCORE</th>
<th>OBTAINED LOW SCORE</th>
<th>OBTAINED HIGH SCORE</th>
<th>TOTAL NUMBER OF QUESTIONS</th>
<th>BLANK</th>
<th>INVALID</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>5.000</td>
<td>5.000</td>
<td>1.449</td>
<td>-0.211</td>
<td>1.595</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>1000</td>
<td>11</td>
<td>70</td>
</tr>
</tbody>
</table>

ITEM ANALYSIS
TEST FREQUENCY DISTRIBUTION

<table>
<thead>
<tr>
<th>RAW SCORE</th>
<th>STANDARD SCORE</th>
<th>PERCENTILE</th>
<th>PERCENT FREQ</th>
<th>CUM FREQ</th>
<th>EACH # REPRESENTS UP TO 1 SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>844</td>
<td>99</td>
<td>1.0</td>
<td>1</td>
<td>100 #</td>
</tr>
<tr>
<td>9</td>
<td>759</td>
<td>98</td>
<td>0.0</td>
<td>0</td>
<td>99 **</td>
</tr>
<tr>
<td>8</td>
<td>706</td>
<td>97</td>
<td>2.0</td>
<td>2</td>
<td>99 **</td>
</tr>
<tr>
<td>7</td>
<td>637</td>
<td>92</td>
<td>8.0</td>
<td>8</td>
<td>97 **</td>
</tr>
<tr>
<td>6</td>
<td>568</td>
<td>76</td>
<td>24.0</td>
<td>24</td>
<td>89 **</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>46</td>
<td>34.0</td>
<td>34</td>
<td>65 **</td>
</tr>
<tr>
<td>4</td>
<td>431</td>
<td>21</td>
<td>18.0</td>
<td>18</td>
<td>31 **</td>
</tr>
<tr>
<td>3</td>
<td>362</td>
<td>8</td>
<td>8.0</td>
<td>8</td>
<td>13 **</td>
</tr>
<tr>
<td>2</td>
<td>293</td>
<td>3</td>
<td>3.0</td>
<td>3</td>
<td>5 **</td>
</tr>
<tr>
<td>1</td>
<td>224</td>
<td>1</td>
<td>2.0</td>
<td>2</td>
<td>2 **</td>
</tr>
</tbody>
</table>
ITEM ANALYSIS
ITEM STATISTICS

ITEM 1 PERCENT OF CORRECT RESPONSE BY FIFTHS

1ST + 
2ND + 
3RD + 
4TH + 
5TH + 

---------
0 10 20 30 40 50 60 70 80 90 100

ITEM 1 MATRIX OF RESPONSES BY FIFTHS

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>UNIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ST</td>
<td>5</td>
<td>4</td>
<td>21</td>
<td>8</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2ND</td>
<td>15</td>
<td>1</td>
<td>8</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3RD</td>
<td>13</td>
<td>3</td>
<td>6</td>
<td>15</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>4TH</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>16</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>5TH</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

PROP 0.30 0.07 0.23 0.31 0.09 0.01
RPB1 -0.17 -0.01 0.49 0.020 -0.09 -0.05