THE EXPECTED VALUE OF GUTTMAN'S COEFFICIENT OF RELIABILITY
BASED ON THE MULTINOMIAL DISTRIBUTION

Janice L. Ween, MAC and HSR Units, Boston University Medical School*
John A. McElhiney, Department of Psychology, Southern Illinois University

ABSTRACT

Although Guttman's Coefficient of Reproducibility (CR) is a widely-used statistic for assessing scale reliability, its statistical basis is not well established in common use. A minimum value of 0.9 for the CR has become the generally accepted standard for a reliable scale, even though the distribution of the CR has been shown to heavily depend, both on the number of items in the scale, and their marginal distributions. The Minimum Marginal Reproducibility (MMR) statistic, generally used as an expected CR, assuming a random distribution of responses to scale items, actually underestimates this value.

This paper gives the expected value and standard error of the CR based on the multinomial distribution, as well as the Green's B statistic, an approximation to it. Expected values for the CR are compared to Green's B and the MMR for a variety of scale item distributions and by the total number of scale items. Additionally, a modification to the current SAS GUTTMAN procedure, which calculates the expected CR and its standard error, is described. These additions to GUTTMAN will increase the usefulness of the CR for evaluating scale reliability by providing it with a statistical basis.

INTRODUCTION

Guttman scale analysis has been widely used to examine the properties of composite scales in the fields of Psychology and Sociology, and more recently, Health. The analysis is useful for additive as well as true Guttman scales. An additive scale with good Guttman properties is useful over a wider range of the trait being measured than a scale constructed only to maximize reliability.

The coefficient of reproducibility (CR) is the central statistic of Guttman scalability analysis. It is a measure of scale inter-item consistency, and ranges from some minimum value greater than 0 to 1, with 1 representing perfect consistency. There is some confusion in the literature as to how the statistic is calculated. The method used by SAS and SPSS is not one of those prominent in the social science literature. A brief description of the various calculation methods is given in the following section.

A widely used criterion for determining whether a group of responses constitutes a scale is that the CR exceed 0.9. This paper shows this criterion to be inadequate because the distribution of the CR changes with a number of factors including (a) the number of responses in the scale, and (b) the distributions of the individual responses. Specifically, the expected value of the CR declines with the number of responses but increases with increasing skewness, e.g., when most subjects respond either positively or negatively to all questions. In this paper an expected CR assuming the null hypothesis of independence between responses developed by Goodman (1959) is described, along with its standard error, and several indices derived from it. These are included in the authors' modifications to the SAS GUTTMAN procedure, and may be used to test the hypothesis that the observed CR is greater than the CR due to chance.

COEFFICIENT OF REPRODUCIBILITY

The coefficient of reproducibility (CR) is defined for a set of k items, where each item has only two outcomes (+ or -). These k items are assumed to be ordered by increasing popularity of a positive response. If the order is not determined a priori, then the assumption is made that the ordering for the N respondents sampled is almost certainly the same as the ordering if the total population were studied. There are then 2^k possible types of responses to the k items.

An additional assumption made regarding the k items is that there are only k + 1 "correct" response combinations, and that all other responses are "erroneous". For k = 4, these "correct", or ideal types are (- - - -), (- - +), (- + +), and (+ + +). A positive response to any item implies a positive response to all more popular items, with any deviation from this rule being considered erroneous, e.g., (- + + -).

Several versions of the CR have been developed as an index for erroneous responses, incorporating differing methods of error measurement. The CR is defined as CR = 1 - e/N, where e is the total number of errors appearing in response to the k items where N individuals are sampled. It is an index to the amount of deviation from the k + 1 correct responses, with perfect adherence to these responses resulting in e = 0, and CR = 1. Likewise, as e increases, the CR decreases.

ENUMERATION OF ERRORS

There have been several methods devised for the enumeration of errors, with the value for e and CR differing markedly by the choice of method. They fall into two groups, with intra-group variation much smaller than that found between groups.

*Supported in part by Grant #AM23613-NIH.
The first group contains the error counting method used in the SAS and SPSS Guttman subprograms, and referred to by Hofman (1979) as Guttman's own. The method is straightforward: first, the number of positive responses is enumerated; the corresponding correct response is assumed to be that ideal response corresponding to the number of positive responses obtained. Each deviation from this pattern is considered to be an error. Thus, for k = 5 and erroneous response pattern (+ - + +), the correct response is (- - + +), and responses to items 1 and 3 are in error. For the jth respondent, ej = 2, with e = n

\[ L e_j, \]

\[ j=1 \]

The other group (Group II) consists of variations on Guttman's method, but with the computed error roughly half that of the above methods. The objective of assignment of erroneous responses to the most appropriate ideal type is secondary to the objective of minimizing errors. Henry (1952) attempts to minimize the errors through selection of the comparison, ideal response. For the pattern (+ - + + +), the ideal response would be (- - + +), and e would then be 1. Sagi (1955) approximates e while simplifying hand calculation by counting the number of reversals, or (+-) patterns. Green (1956) counts not only (+-) patterns, but (-+) patterns as well. Both of these methods yield e = 1 for the pattern (+ - + + +) above as does Henry's method. There is even confusion in the literature as to Guttman's method (Sagi (1955), Goodman (1956) and Chilton (1969), where it is described as producing an e similar to that of Henry.

An example should help illustrate the differences. Table 1 gives all possible response patterns for a 4-item scale. A (-) indicates a negative response while a (+) indicates a positive response. It can be seen that the ideal type for the SPSS/SAS Guttman analysis is equal to the total number of positive responses. This differs from the Group II ideal type in every erroneous response pattern. However, the Group II procedures is chosen to minimize error, the number of errors is also smaller for every erroneous response pattern. The Green and Henry methods give identical error counts for the case where k = 4. This is also true for k = 5, but for k = 6, the Green method underestimates the number of errors for the pattern (+ + - - -); where Sagi, counting only (+-) patterns estimates e = 1; Green, counting additionally, (+ + - - ) patterns estimates e = 2; and where Henry, more precisely, estimates e as 3.

For Group I estimation, assignment of erroneous responses to an ideal pattern is straightforward. For Group II estimation, assignment to more than one type may be justified. Where there is more than one ideal type for which e is at a minimum, the ideal type is chosen to be the one shown by the data to be more likely. For example, if the ideal type 0 pattern is more common in the data than the type 2, then the error pattern (+ - + +) would be assigned to type 0. The assignment of ideal type to error responses is done for data interpretive purposes only and does not affect the number of errors, or the CR, in any way.

Because both error counting schemes are present in the literature and reference is made predominately to the Group II estimation, Henry's CR is included in the author's version of the SAS GUTMAN procedure. The e for Henry's CR is always approximately half that for the SAS/SPSS CR. Because it is considered to be the most accurate method for error estimation, all further reference to CR in this paper will be to Henry's CR.

**DISTRIBUTION OF THE CR ASSUMING INDEPENDENT RESPONSES**

Goodman (1959) has developed a general method for calculation of an expected value and standard error for all versions of the CR, irrespective of the error-counting method employed. The method is based on the assumption that all k responses are independent, i.e., the scale responses are uncorrelated and therefore do not contain properties requisite for a good scale. The k2 possible patterns are assumed to have a multinomial distribution with the probability of the tth pattern, where t = 0, ..., K-1, equal to

\[ P_t = \frac{\prod_j (1 - \delta_{jt} P_j + \delta_{jt} (1 - P_j))}{\sum_j P_j} \]

where \( P_j = \text{Prob}(\text{tth response is positive}) \), \( P_j = 1 - P_j = \text{Prob}(\text{tth response is negative}) \), and \( \delta_{jt} = 1 \) if the jth response in the tth pattern is positive, \( \delta_{jt} = 0 \) otherwise. Then, using Goodman's notation,

\[ C = 1 - (\frac{\sum_t P_t Y_t}{N}) = 1 - (\frac{\sum_t P_t X_t/k}{N}) \]

where C is the expected CR, \( D_t/N = P_t \) is the estimated proportion of the population in category t, assuming independence between responses, and \( X_t = \text{number of errors for the tth response pattern} \). Then if \( Y_t = X_t/k \) is the proportion of the responses in error for the tth pattern,

\[ C = 1 - (\frac{\sum_t P_t Y_t}{N}) = 1 - (\frac{\sum_t P_t X_t}{N}) \]

the variance of C is then

\[ \sigma = \frac{(\sum_t P_t Y_t^2 - (\sum_t P_t Y_t)^2)/N}{C} \]

Analogously, if \( P_t \) = the observed proportion in the tth pattern is used instead of \( P_t \), then \( c \), the observed CR, is given by

\[ c = 1 - \frac{\sum_t P_t Y_t}{N} \]
and its variance is estimated by
\[ \sigma^2 = \frac{\sum (y_i - \bar{y})^2}{N-1} \]
and is the variance of the CR for the non-null case. This variance can also be derived using the Delta Method developed by Bishop, Flomberg, and Holland (1975). The \( \sigma^2 \), along with \( c \), can be used inferentially to test the hypothesis that the CR differs from the expected CR assuming independence, and \( \sigma^2 \) is useful for constructing a confidence interval for the observed CR. The variance for the observed CR has not been widely used in practice, but is seen as potentially very useful. It is not included in the authors' present version of ROC GUTTMAN but will appear in subsequent versions.

A rule of thumb widely used regarding the CR is that a good scale must have a CR greater than 0.9. Like all rules of thumb, this one has its limitations. A well designed scale differentiates uniformly across the range of observations, with each response contributing equally to the scale. This is reflected in the uniform distributions of the \( p_j \) = \( \text{Prob}(j^{th} \text{ response is positive}) \). For \( k = 4 \), the ideal distribution of the \( p_j \) = \( (0.2, 0.4, 0.6, 0.8) \). The 0.9 rule may be valid for this distribution, but deviations from this pattern are common, and result in changes to the Expected CR(C), which is a function of those marginal probabilities. Table 2 illustrates several patterns for \( k = 4 \).

The \( C \) generally increases as the \( p_j \) converge on the uniform \( p_j = 0.5 \) pattern. For symmetric distributions, \( C \) increases as all \( p_j \) approach 0 or 1. It should be noted that the Minimum Marginal Reproducibility (MMR), generally used as the CR due to chance, does not correspond with \( C \), except for response pattern (b).

Because the MMR is calculated based on the probability of the most common response pattern, MMR's for (b) and (e), (c) and (f), and (d) and (g) are the same, and does not adjust for skewness.

The value of \( C \) also decreases generally with increasing \( k \). This is true because the proportion of true types decreases with increasing \( k \).

To the extent that the observed CR is random, it is also affected by these factors, and a measure of the improvement made over chance as measured by \( C \) becomes more important. The authors' modifications to the GUTTMAN procedure calculate the following statistics necessary to make this comparison.

1. Henry's coefficient of reproducibility (CR);
2. The expected value of Henry's CR (C), assuming independent responses;
3. The standard error of C based on Green's B approximation;
4. A modification of Green's I (Green, 1956), using Henry's CR. This statistic has a range of \((-1, 1)\) and has an interpretation analogous to that of a correlation coefficient.
5. A z statistic
\[ z = \frac{(CR - E(C))}{\text{s.e.} \cdot C} \]
whose distribution is asymptotically normal \((0, 1)\) under the hypothesis of independent responses (Goodman 1959).

The above statistics should prove useful in determining the scalability of a group of responses, especially for deviations of the marginal probabilities from the ideal.

**DISCUSSION**

The above statistics supplementing the SAS GUTTMAN procedure are necessary adjuncts to the coefficient of reproducibility for evaluation of the scalability of a particular set of responses. They provide, among others, an unbiased estimate of the C under the assumption of independence, which the observed CR can be compared against, and also a standard error, so that statistical inference may be performed. They should replace, in scale analysis, the comparison to the MMR, which has been shown to be an inaccurate estimate of the chance CR, as well as the old .9 "rule of thumb".

Examination of the statistics described above is an important step in the assessment of scaling properties, but others have been developed to examine the particular patterns of error responses, and are briefly enumerated below. The following approaches test additional assumptions concerning the relationship of the responses. The Proctor (1970) goodness of fit test, also available in the SAS GUTTMAN procedure, examines whether all error responses, given the distribution of true types, are equally likely. A chi-square statistic is used for this purpose, but the available software makes no provision for grouping of response patterns when expected values are very small. Care must therefore be taken when interpreting results that subsequent adjustments to the statistic are made before statistical conclusions are drawn. Hofman (1979) performs a similar analysis assuming a binomial distribution for the error responses. Goodman (1975) uses general contingency table loglinear analysis techniques to determine scalability. These analyses all assume the existence of ideal types as evidenced by a CR significantly improved over chance and so would be performed subsequently to the GUTTMAN analyses.
**TABLE 1:** Comparison of Error Counting Procedures

<table>
<thead>
<tr>
<th>Response Number</th>
<th>Response</th>
<th>Ideal Type Errors</th>
<th>SAS/SPSS Ideal Type Errors</th>
<th>Sagi Errors</th>
<th>Green Errors</th>
<th>Henry Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(- - -)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>(- - +)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>(- + -)</td>
<td>1</td>
<td>2</td>
<td>0/2*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>(- + +)</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>(- + -)</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>(- + +)</td>
<td>2</td>
<td>2</td>
<td>1/3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>(- + +)</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>(- + +)</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>( + - -)</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>( + - +)</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>( + - +)</td>
<td>2</td>
<td>2</td>
<td>0/4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>( + - +)</td>
<td>3</td>
<td>2</td>
<td>2/4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>( + - +)</td>
<td>2</td>
<td>4</td>
<td>0/4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>( + - +)</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>( + + -)</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>( + + +)</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*More than 1 ideal type

**TABLE 2:** Expected CR, MMR for k = 4 Responses, Selected Patterns of Marginal Response Probabilities

<table>
<thead>
<tr>
<th>Distribution Shape</th>
<th>Response Probability</th>
<th>Expected CR</th>
<th>MMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Symmetric .5 .5 .5 .5</td>
<td>0.7969</td>
<td>0.5000</td>
<td></td>
</tr>
<tr>
<td>(b) Symmetric .2 .4 .6 .8</td>
<td>0.6924</td>
<td>0.7000</td>
<td></td>
</tr>
<tr>
<td>(c) Symmetric .1 .4 .6 .9</td>
<td>0.6571</td>
<td>0.7500</td>
<td></td>
</tr>
<tr>
<td>(d) Symmetric .1 .2 .8 .9</td>
<td>0.6204</td>
<td>0.8500</td>
<td></td>
</tr>
<tr>
<td>(e) Asymmetric .2 .2 .4 .4</td>
<td>0.7944</td>
<td>0.7000</td>
<td></td>
</tr>
<tr>
<td>(f) Asymmetric .1 .1 .4 .4</td>
<td>0.7951</td>
<td>0.7500</td>
<td></td>
</tr>
<tr>
<td>(g) Asymmetric .1 .1 .2 .2</td>
<td>0.8844</td>
<td>0.8500</td>
<td></td>
</tr>
<tr>
<td>(h) Asymmetric .1 .2 .3 .4</td>
<td>0.8034</td>
<td>0.7500</td>
<td></td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


Hofman, R.J. On testing a Guttman scale for significance. Educational and Psychological Measurement, 1979, 39, 297-301.

