To Quantify Measurement Error
for Binary Data

Roshanak Alimohammadi

Department of Mathematics
Alzahra University
Tehran, Iran
r_alimohammadi@alzahra.ac.ir

Abstract

To estimate data accuracy, survey errors should be determined. Measurement error is an unavoidable and essential part of survey errors. Therefore, modeling of measurement error has many applications. There are some essential differences between modeling of measurement error for continues data and discrete data. In the previous works, modeling of measurement error for continues data is assessed.

In this paper, the probability distribution and the model of measurement error for binary data in surveys is studied. In the other words, agreement concept is applied to assess consent between observations and true values in order to quantify measurement error of binary data in surveys.

Mathematics Subject Classification: 62D05, 62P99

Keywords: Survey Error, Measurement Error, Binary Data, Agreement, Cohen’s Kappa

1 Introduction

To determine data accuracy, survey error should be estimated. To design a model of measurement error for continues data, sources of measurement error may be assessed.

Mahalanobis [7] was the first statistician that applied ANOVA for modeling of measurement error. Kish [6] considered an ANOVA model in which the response from the jth unit to the ith interviewer is expressed as:

\[ y_{ij} = \mu_{ij} + A_i + e_{ij} \]
Where $\mu_{ij}$ is the true value and $y_{ij}$ is the observed value of $ij$th unit, $A_i$ is the effect of $i$th interviewer on any interview and $e_{ij}$ is the effect of $ij$th respondent that may be considered as containing other sources of error.


Hansen, Hurwitz and Pritzker [5] have proposed a model of measurement error for binary data by considering two stage cluster sampling design. Primary stage is related to selecting sampling units and second stage is applied for repeated responds of each sample unit.

Biemer and Trewin [4] have considered two general models for measurement error, one for continuous data and one for binary data. Here the first type is described. Based on their notation, $U = \{1,2,\ldots,N\}$ is a label set for the target population containing $N$ units and $S = \{1,2,\ldots,n\}$ denotes the label of sample units so that $n=mI$. It is assumed that $S$ is partitioned into $I$ assignments of $m=n/I$ units. $S_i$ is the set of units assigned to the $i$th operator and $d_{ij}$ is the error of $j$th unit in $S_i$ for $j = 1,\ldots,m$ and $i = 1,\ldots,I$. For $j \in S_i$, it is assumed that $d_{ij}$ is sum of two error terms, $A_i$ and $B_{ij}$, where $A_i$ is an operator error which is assumed to be same for all units in the $i$th operator assignment and $B_{ij}$ is the elementary error due to respondent as well as other sources of error including the operator. Thus, the model for $ij$th observation is:

$$y_{ij} = \mu_{ij} + d_{ij} = \mu_{ij} + A_i + B_{ij}$$

(1)

Where $A_i$ can be considered as fixed or random and $B_{ij}$ as random variables.

Biemer and Trewin [4] have presented a model of measurement error for binary data too. The assumptions of continues data model are not appropriate yet. As for continues data, they assume that the population is partitioned into $I$ operator assignments: $S_1, S_2,\ldots, S_i$. Let $y_{ij}$ denote the observation for the $j$th unit in $S_i$, ($i = \{1,2,\ldots,n\}$ where $y_{ij} = 1$ if the response is ”yes” and $y_{ij} = 0$ if ”no”. Let $\mu_{ij}$ denote corresponding true value which is 1 if unit (i,j) is a true ”yes” and 0 if a true ”no”.

For $j \in S_i$, the following misclassification probabilities are defied:

$$\theta_i = P(y_{ij} = 0|\mu_{ij} = 1)$$

$$\phi_i = P(y_{ij} = 1|\mu_{ij} = 0)$$

where $\theta_i$ and $\phi_i$ are referred as the probability of false negative and the probability of false positive, respectively. Measurement error is defined as $d_{ij} = y_{ij} - \mu_{ij}$ and distribution of the error, $d_{ij}$ is given as:

Table 1 represents probability model of measurement error.
To Assess Measurement Error for Binary Data

In this section, agreement indices are applied to assess measurement error. Furthermore, a model of measurement error is proposed by justifying Cohen’s Kappa.

Measurement error is defined as difference between observed and true value. Suppose that \( \pi(i, j) \) is probability of observed value equals to \( i \) and true value is \( j \). Also, \( j = 1 \) denotes corresponding true value is a true positive and \( j = 0 \) if a true negative.

By definition, measurement error equals \( i - j \), and its possible values are in the set \( \{-1, 0, 1\} \). Obviously, \( \pi(0, 0) \) and \( \pi(1, 1) \) indicate complete agreement between observed and true values. Therefore, measurement error equals to zero.

Furthermore, \( \pi(0, 1) \) and \( \pi(1, 0) \) indicate complete disagreement between observed and true values. In this case, measurement error is occurred. Therefore, table 2 represents probability distribution of \( (i, j) \):

<table>
<thead>
<tr>
<th>True value</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>( \pi(0,0) )</td>
<td>( \pi(0,1) )</td>
</tr>
<tr>
<td>1</td>
<td>( \phi(1,0) )</td>
<td>( \phi(1,1) )</td>
</tr>
</tbody>
</table>

Table 2: Distribution of \( (i,j) \)

Table 3 represents probability distribution of measurement error:

<table>
<thead>
<tr>
<th>Measurement error</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>( \pi(1,0) )</td>
</tr>
<tr>
<td>0</td>
<td>( \theta_0 )</td>
</tr>
<tr>
<td>1</td>
<td>( \phi(0,1) )</td>
</tr>
</tbody>
</table>

Table 3: Distribution of measurement error

Where, \( \theta_0 = \pi(1,1) + \pi(0,0) = 0 \).
Now, Cohen’s kappa is applied to determine measurement error as:

\[ \alpha = \frac{\theta_0 - \theta_e}{1 - \theta_e} \]

where \( \theta_0 \) is probability of agreement and \( \theta_e = \pi(.,0)\pi(0,.) + \pi(.,1)\pi(1,.) \) indicates the proportion of units for which agreement is expected by chance.

It can be shown that, \( \alpha = 1 \) if and only if \( \theta_0 = 1 \). Therefore, there are complete agreement between observed and true values. In the other words, the amount of measurement error equals to zero.

On the other hand, \( \alpha = 0 \) if and only if \( \theta_0 = \theta_e \). This equality is justified when \( \pi(1,1) = \pi(.,1)\pi(1,. \) and consequently, \( \pi(0,0) = \pi(.,0)\pi(0,. \). That is by considering table 2, probability of agreement equals to the probability of agreement to be expected by chance. In the other words, observed and true values be independent.

Also, \( \alpha = -1 \) if and only if \( \pi(0,0) = \pi(1,1) = 0 \) and \( \pi(0,1) = \pi(1,0) = 0.5 \). Therefore, probability of measurement error is large and can be obtained as:

\[ \theta_0 = \pi(1,1) + \pi(0,0) = 0 \] (2)

and

\[ \theta_e = \pi(.,0)\pi(0,.) + \pi(.,1)\pi(1,.) = 2\pi(0,1)\pi(1,0) = 0.5 \] (3)

Relations (2) and (3) resulted to \( \pi(0,1) = \pi(1,0) = 0.5 \). In this case, probability of measurement error equal -1 or 1 is the same and for both of them true response are obtained by chance.

Now, we consider absolute value of measurement error or absolute of difference between observed and true values, \(|i - j|\). It’s possible amounts are in the set \{0, 1\}. Therefore, \( P(|e| = 0) = \theta_0 \) and \( \theta_0 = 1 \) when \( \alpha = 1 \). In the other words, absence measurement error and complete agreement between observes and true values. Furthermore, \( \alpha = 0 \) if and only if, the chance of presence and absence of measurement error be equal, that is \( P(|e| = 0) = P(|e| = 1) = 0.5 \). On the other hand, \( P(|e| = 1) = \pi(0,1) + \pi(1,0) \) if \( \alpha = -1 \).

Then, application of the proposed criterion can be applied for modeling measurement error and interpreting the amount of the error of binary data. In the other words, if the criterion equals 1, observed and true values is the same and if it equals 0, the probability of measurement error is 0.5. If it be -1, measurement error is maximum.

3 Conclusion

Estimation of data quality of survey results is important because of their applications.
In this paper, measurement error of binary data in surveys is assessed and a criterion to quantify measurement error is proposed. Besides, the assessed method can be applied for modeling of measurement error of binary data. Furthermore, the method may be applied to compare data gathering modes and to compute the agreement between modes of data gathering.

**ACKNOWLEDGEMENTS.** This research is supported financially by Alzahra University, tehran, Iran.

**References**


**Received: July, 2011**