Thai Journal of Mathematics
Special Issue: The 17^{th} IMT-GT ICMSA 2021
Pages 62–74 (2022)

http://thaijmath.in.cmu.ac.th

ISSN 1686-0209

Algorithms for the Test of Independence of Two Categorical Variables over Uncertain Data

Monchai Kooakachai[1]

[1] *Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, Thailand 10330*
e-mail : Monchai.K@chula.ac.th

Abstract It is known that the independence of two categorical variables can be analyzed by using the Chi-squared test. However, when some information in the original data was incomplete, it is not clear how to adjust it in the calculation. Rather than throwing away those uncertain data, it is recommended to include such observations. This issue has been classically addressed by using the Expectation and Maximization (EM) algorithm. In this paper, we proposed two alternative ways to tackle this problem. The first method, adapting the E-step in the EM algorithm, was derived from relaxing some restrictions in the original approach. The second method was obtained from the fact that in some cases the true level of uncertain data should be completely random and not depend on available data, therefore uncertain observations are taken into the calculation by the discrete uniform distribution. Finally, we investigated the performance of all methods through simulations. The simulations for the case of 2 × 2 contingency table were performed. The real data examples, such as general social survey and victimization status, were also presented to illustrates pros and cons of each method.

MSC: 62H17; 62G05
Keywords: test of independence; contingency tables; uncertain data; incomplete data; EM algorithm

Submission date: 15.03.2022 / Acceptance date: 31.03.2022

1. Introduction

The two-way contingency table is the most basic way to summarize results from two categorical variables obtained from a survey. The summarized data in the table is called complete or certain if there is no missing answer in the survey, i.e., respondents answer all questions without skipping them. If the data is complete, the independence of those two categorical variables can be checked by using the Chi-squared test. However, it is possible that people who answered the survey could leave some questions blank due to the uncomfortableness of answering some questions or it is the case that the unanswered questions are not applicable. This leads to the incomplete or uncertain contingency table where information obtained on one or more of the categorical variables is missing. The

Published by The Mathematical Association of Thailand.

Chi-squared test of independence cannot be applied until the missing data problem is solved.

Including missing or uncertain observations in the analysis is quite challenging. In general, researchers can throw away incomplete data but it is not recommended on some types of data since it may cause the losing of information. If the ratio between missing and complete data is high, deleting all uncertain observations will not be appropriate. It was mentioned in [1] that if both categorical variables are missing 30% of their observations and missingness happens independently for variables, then about half (51%) of the sample is likely to be lost. Rather than throwing away such incomplete data, statisticians have developed tools to deal with incomplete data, including how to keep them in the analysis. In 2005, the maximum likelihood estimation procedures for an incomplete two-way contingency table were revised and written formally in Ehlers dissertation [2]. Three years later, Takai and Kano extended original results to the case of a 2 × 2 contingency table with nonignorable responses [3]. Later in 2010, Petitrenaud introduced an idea to adjust uncertain data in a 2 × 2 contingency table by using a frequentist method and his belief function [4]. These pieces of evidence indicate that there were many attempts for improving the imputation of incomplete two-way contingency tables.

A classical way to get around the uncertain data problem in the contingency table is using the Expectation and Maximization (EM) algorithm. One drawback of this classical EM algorithm is that, for the process of replacing missing data, it only uses the observed data that is restricted to one level that the observation belongs to. For example, suppose we have two questions on the survey. The first one is whether the respondent has a cat and the second question asks if he or she favors the death penalty. For those people who answered that they have a cat but didn't respond to their decisions on the death penalty, their decisions on the death penalty will be imputed back in the calculation by using a ratio from a group of respondents that have a cat. Such a restriction does not make sense if we believe that opinions on the death penalty do not depend on having a cat.

Based on the simple idea on the previous paragraph, we propose two different ways to impute the incomplete data in the contingency table and compare it to the classical EM method. In section 2, we will first discuss the EM algorithm and how it applies to the context of independence tests of two categorical variables. Then section 3 defines the two proposed methods. The simulations for the case of 2 × 2 contingency table were performed and the result will be presented in section 4. Real data examples will be discussed in section 5 and the conclusion and discussion will be mentioned in section 6.

2. The EM Algorithm on the Chi-Squared Test of Independence

The EM (Expectation and Maximization) algorithm was introduced in a classic 1977 paper by Arthur Dempster, Nan Laird and Donald Rubin [5]. Basically, it is the algorithm that is used to estimate parameters when the data is incomplete or has missing values due to limitations of the observation process. One motivation of the algorithm is to fix an issue of intractable equations when using the maximum-likelihood estimation approach.

2.1. Maximum-Likelihood

The method of maximum likelihood is, by far, the most popular technique for deriving estimators [6]. Suppose we have an iid sample $X_1, X_2, \ldots, X_n$ from a population with pdf or pmf $f(\mathbf{x};\theta)$, the likelihood function $L(\theta;\mathbf{x})$ is defined by

$$L(\theta;\mathbf{x}) = f(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

It can be thought of as a function of parameters θ where the data is fixed. The parameter that maximizes $L(\theta;\mathbf{x})$ is called the maximum likelihood estimator (MLE), denoted by $\widehat{\theta}_{\mathrm{MLE}}$. Mathematically,

$$\widehat{\theta}_{\mathrm{MLE}} = \underset{\theta}{\operatorname{argmax}}\ L(\theta;\mathbf{x}).$$

In most cases, especially when differentiation is to be used for optimization, it is easier to work with the logarithm of likelihood than it is to work with the likelihood. Using log-likelihood function is valid because the logarithmic function is increasing on $(0,\infty)$. Depending on the form of the log-likelihood, this optimization problem can be easy or difficult. For example, if there is a sum of terms inside the logarithm function, then deriving the explicit form of MLE would be intractable. The EM algorithm was developed to deal with this issue under some certain assumptions.

2.2. Mathematical Framework for the EM Algorithm

Assume the data $\mathbf{X} = (X_1,\ldots,X_n)$ is observed and generated by some distribution. Our goal is estimating parameter θ given the data. Suppose there is a missing or hidden data $\mathbf{Y}$ so that we call $\mathbf{X}$ incomplete data. Assume $\mathbf{Z} = (\mathbf{X},\mathbf{Y})$ is a complete data set. The EM-algorithm is useful when the incomplete likelihood function $L(\theta;\mathbf{x})$ is intractable and working with the complete likelihood function $L(\theta;\mathbf{z})$ is much easier. By the conditional probability, the complete data density function is

$$f(z;\theta) = f(x,y;\theta) = f(x;\theta)\cdot f(y \mid x;\theta).$$

Taking logarithm on both sides yields

$$\log f(z;\theta) = \log\left[f(x;\theta)\cdot f(y \mid x;\theta)\right] = \log f(x;\theta) + \log f(y \mid x;\theta)$$

which means

$$\log L(\theta;z) = \log L(\theta;x) + \log f(y \mid x;\theta)$$

and therefore

$$\log L(\theta;x) = \log L(\theta;z) - \log f(y \mid x;\theta).$$

Now suppose that maximizing $\log L(\theta;x)$ is difficult or does not have a closed form. Instead, we work with $\log L(\theta;z)$. Since $\mathbf{Y}$ is unknown or hidden, the complete log-likelihood function $\log L(\theta;\mathbf{z})$ can be thought as a random variable where the random part is $\mathbf{Y}$ and constant terms are $\mathbf{X}$ and θ. Taking expectation over $\mathbf{Y}$ given the observed data $\mathbf{X}$ and the current parameter estimates $\theta^{(t)}$ yields

$$E\left[\log L(\theta;x) \mid x,\theta^{(t)}\right] = E\left[\log L(\theta;z) \mid x,\theta^{(t)}\right] - E\left[\log f(y \mid x;\theta) \mid x,\theta^{(t)}\right].$$

The left handed side is the incomplete log-likelihood function which does not depend on $\mathbf{Y}$ so we have

$$\log L(\theta;x) = E\left[\log L(\theta;z) \mid x,\theta^{(t)}\right] - E\left[\log f(y \mid x;\theta) \mid x,\theta^{(t)}\right].$$

The previous equation holds for any value of θ so

$$\log L(\theta^{(t)};x) = E\left[\log L(\theta^{(t)};z) \mid x,\theta^{(t)}\right] - E\left[\log f(y \mid x;\theta^{(t)}) \mid x,\theta^{(t)}\right].$$

Gibb's inequality [7] guarantees that

$$E\left[\log f(y \mid x; \theta) \mid x, \theta^{(t)}\right] \geq E\left[\log f(y \mid x; \theta^{(t)}) \mid x, \theta^{(t)}\right]$$

which means

$$\log L(\theta; x) - \log L(\theta^{(t)}; x) \geq E\left[\log L(\theta; z) \mid x, \theta^{(t)}\right] - E\left[\log L(\theta^{(t)}; z) \mid x, \theta^{(t)}\right]$$

The last inequality indicates that choosing θ to improve $E\left[\log L(\theta; z) \mid x, \theta^{(t)}\right]$ beyond $E\left[\log L(\theta^{(t)}; z) \mid x, \theta^{(t)}\right]$ will result in increasing the incomplete log-likelihood function as well and that is why the EM algorithm works. Note that the main part of the EM algorithm is evaluating and maximizing $E\left[\log L(\theta; z) \mid x, \theta^{(t)}\right]$. Define

$$Q(\theta, \theta^{(t)}) := E\left[\log L(\theta; z) \mid x, \theta^{(t)}\right].$$

The first θ in $Q(\theta, \theta^{(t)})$ refers to the parameters that will be optimized to maximize the log-likelihood while the second $\theta^{(t)}$ refers to the parameters that we use to evaluate the expectation. The EM algorithm consists of an Expectation step (E-step) followed by a Maximization step (M-step) as follows:

E-step : Compute $Q(\theta, \theta^{(t)})$ where

$$Q(\theta, \theta^{(t)}) = E\left[\log L(\theta; z) \mid x, \theta^{(t)}\right]$$

M-step : Find $\theta^{(t+1)}$ such that

$$\theta^{(t+1)} = \underset{\theta}{\text{argmax}}\ Q(\theta, \theta^{(t)})$$

The two steps are repeated as necessary. Each iteration is guarantee to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function.

Note that, in the case of the exponential family, The E-step reduces to computing the conditional expectations of the complete data sufficient statistics given the observed data. After that, the conditional expectations of the sufficient statistics computed in the E-step can be directly substituted in the M-step to obtain the next iteration [8].

2.3. Applying the EM Algorithm to the Chi-Squared Test

Consider two categorical variables, A and B, where A consists of m levels and B has n levels. For each $i \in \{1, 2, 3, \ldots, m\}$ and $j \in \{1, 2, 3, \ldots, n\}$, we define the parameter of interest π_{ij}, the probability that an observation falls in cell (i, j) in the corresponding $m \times n$ contingency table. Observe that $\sum\sum \pi_{ij} = 1$.

In the classical test of independence with complete data, we draw a sample of size N and count the number of observations falling in each cell. Let Y_{ij} be the count in the cell (i, j). Then the vector $Y = (Y_{11}, Y_{12}, \ldots, Y_{mn})$ is multinomially distributed with parameter n and $\pi = (\pi_{11}, \pi_{12}, \ldots, \pi_{mn})$. Our target parameter π_{ij} can be estimated by using the ratio of the number of observed objects in the cell (i, j) and the total count. That is, if the observed count is $(y_{11}, y_{12}, \ldots, y_{mn})$ where $y_{11} + y_{12} + \ldots + y_{mn} = N$ then

$$\widehat{\pi}_{ij} = \frac{y_{ij}}{N}.$$

For the complete data, this estimator is a MLE of π_{ij}. Without any incomplete value, a test of independence of two variables A and B is utilized in the standard procedure, i.e., calculating the chi-squared statistic from the $m \times n$ contingency table and making a conclusion based on the corresponding p-value.

For the analysis with the existence of incomplete data, we partition the sample into three parts denoted by M_0, M_A and M_B respectively. The M_0 part stands for observations having both A and B fully observed, i.e., the data is complete on part M_0. We define

$$Y^{M_0} := (Y_{11}, Y_{12}, \ldots, Y_{mn}).$$

On the other hand, the M_A part includes those having only B observed but not A so the level of B is known and the level of A is unknown. Similarly, M_B part refers to those having only A observed but not B. For the M_A part, we define the marginal total $Y_{+j} = \sum_{i=1}^{m} Y_{ij}$ and, similarly for the M_B part, we define the marginal totals $Y_{i+} = \sum_{j=1}^{n} Y_{ij}$. The random vectors representing the counts for parts M_A and M_B are

$$Y^{M_A} := (Y_{+1}, Y_{+2}, \ldots, Y_{+n}) \qquad \text{and} \qquad Y^{M_B} := (Y_{1+}, Y_{2+}, \ldots, Y_{m+}).$$

The uncertain part in the data is Y^{M_A} and Y^{M_B}. Note that the elements having both A and B unobserved are excluded in the analysis.

The idea of the EM algorithm is to impute the incomplete data from parts M_A and M_B back into part M_0 and reestimate the parameter π_{ij} many times. We first set up the parameter $\pi_{ij}^{(r)}$, an estimate of π_{ij} in the r-step of the algorithm. To start the algorithm, we first give the initial value of π_{ij} called $\pi_{ij}^{(0)}$ and then, in the E-step, we calculate the expectation of the count of the cell (i, j) by

$$E(Y_{ij} \mid \pi_{ij}^{(r)}) = \underbrace{y_{ij}}_{\text{complete part}} + \underbrace{y_{+j}\left(\frac{\pi_{ij}^{(r)}}{\pi_{+j}^{(r)}}\right)}_{\text{incomplete part } (M_A)} + \underbrace{y_{i+}\left(\frac{\pi_{ij}^{(r)}}{\pi_{i+}^{(r)}}\right)}_{\text{incomplete part } (M_B)}$$

Observe that, in the formula above, the y_{ij} is a count from the complete part. The second piece comes from multiplying the total number of incomplete observations in the M_A part, called y_{+j}, with the proportion $\pi_{ij}^{(r)}/\pi_{+j}^{(r)}$ obtained from a current estimate restricted on the information that observations belong in the level j. The third piece is defined in the same fashion. In the M-step, $\pi_{ij}^{(r+1)}$ is computed by substituting the results from the E-step. With the help from MLE, we have

$$\pi_{ij}^{(r+1)} = \frac{1}{N}\left(E(Y_{ij} \mid \pi_{ij}^{(r)})\right)$$

Then we repeat the process of the E-step and M-step until the parameters π_{ij} converge. The chart of the algorithm is illustrated in Figure 1. If we multiply the final estimate of π_{ij} to the total number of observations, including complete and incomplete data, we will get the estimated number of observations in each cell. These values are carried to the Chi-squared test. The independence of two categorical is then concluded from the p-value of the Chi-squared test.

3. Proposed Methods

In this section, we propose two alternative ways to estimate the target parameter π_{ij}. Section 3.1 describes the method which is done by adapting a formula in the E-step in the EM algorithm while section 3.2 explains the method that avoids the iteration or

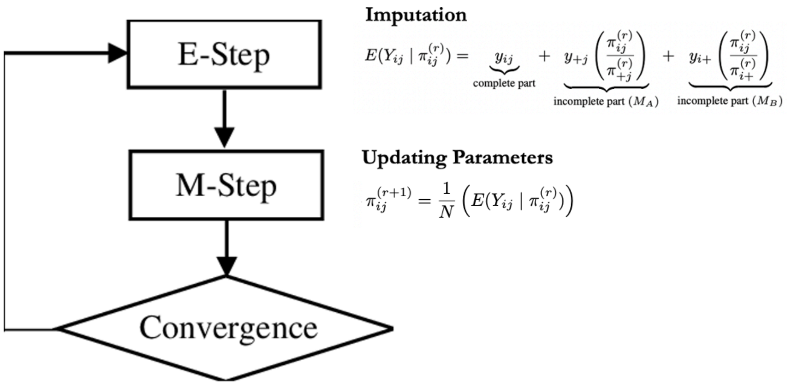
Imputation
$E(Y_{ij} \mid \pi_{ij}^{(r)}) = \underbrace{y_{ij}}_{\text{complete part}} + \underbrace{y_{+j}\left(\frac{\pi_{ij}^{(r)}}{\pi_{+j}^{(r)}}\right)}_{\text{incomplete part } (M_A)} + \underbrace{y_{i+}\left(\frac{\pi_{ij}^{(r)}}{\pi_{i+}^{(r)}}\right)}_{\text{incomplete part } (M_B)}$
E-Step
Updating Parameters
$\pi_{ij}^{(r+1)} = \frac{1}{N}\left(E(Y_{ij} \mid \pi_{ij}^{(r)})\right)$
M-Step
Convergence

Figure 1. Classical EM Algorithm for the Chi-squared Test of Independence

repeating algorithm, i.e., uncertain observations are taken into the calculation by the discrete uniform distribution.

3.1. Adapted EM Algorithm

Recall that in the classical EM algorithm, we calculate $E(Y_{ij} \mid \pi_{ij}^{(r)})$ in the E-step by splitting into three parts. Our goal is to fix the derivation in last two parts M_A and M_B. When the observations level is unknown in A but known in another categorical variable B, the original EM algorithm is utilized by using the estimated ratio from complete data restricted to the level of variable B that the observation belongs. That is why the proportion is of the form $\pi_{ij}^{(r)}/\pi_{+j}^{(r)}$. Here, we remove such a restriction so the estimated ratio is basically computed from the entire complete data, not just the level of variable B that the observation belongs. Repeating this idea on the M^A and M^B parts yields the formula in the E-step as follows:

$$
\begin{aligned}
E(Y_{ij} \mid \pi_{ij}^{(r)}) &= \underbrace{y_{ij}}_{\text{complete part}} + \underbrace{y_{+j}\left(\frac{\sum_{j=1}^{n}\pi_{ij}^{(r)}}{\sum_{j=1}^{n}\pi_{+j}^{(r)}}\right)}_{\text{incomplete part } (M_A)} + \underbrace{y_{i+}\left(\frac{\sum_{i=1}^{m}\pi_{ij}^{(r)}}{\sum_{i=1}^{m}\pi_{i+}^{(r)}}\right)}_{\text{incomplete part } (M_B)} \\
&= y_{ij} + y_{+j}\sum_{j=1}^{n}\pi_{ij}^{(r)} + y_{i+}\sum_{i=1}^{m}\pi_{ij}^{(r)} \\
&= y_{ij} + y_{+j}\pi_{i+}^{(r)} + y_{i+}\pi_{+j}^{(r)}
\end{aligned}
$$

Now, in the M-step, $\pi_{ij}^{(r+1)}$ is computed by substituting the results from the E-step. With the help from MLE, we have

$$\pi_{ij}^{(r+1)} = \frac{1}{N}\left(E(Y_{ij} \mid \pi_{ij}^{(r)})\right)$$

Figure 2 shows the flowchart of the process. We repeat the process of the E-step and M-step multiple times until the parameters π_{ij} are convergent. The Chi-squared test of independence can be applied after the adapted EM algorithm is done.

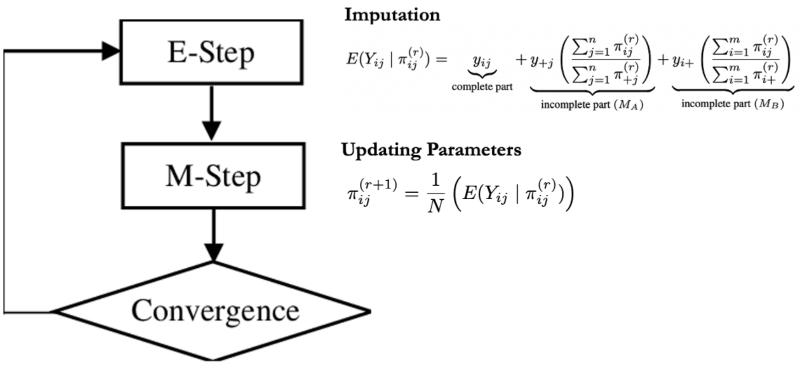
E-Step
M-Step
Convergence
Imputation
$E(Y_{ij} \mid \pi_{ij}^{(r)}) = \underbrace{y_{ij}}_{\text{complete part}} + \underbrace{y_{+j}\left(\frac{\sum_{j=1}^{n}\pi_{ij}^{(r)}}{\sum_{j=1}^{n}\pi_{+j}^{(r)}}\right)}_{\text{incomplete part }(M_A)} + \underbrace{y_{i+}\left(\frac{\sum_{i=1}^{m}\pi_{ij}^{(r)}}{\sum_{i=1}^{m}\pi_{i+}^{(r)}}\right)}_{\text{incomplete part }(M_B)}$
Updating Parameters
$\pi_{ij}^{(r+1)} = \frac{1}{N}\left(E(Y_{ij} \mid \pi_{ij}^{(r)})\right)$

Figure 2. Adapted EM Algorithm for the Chi-squared Test of Independence

3.2. Using the Discrete Uniform Distribution

The second proposed method was obtained from the fact that in some cases the true level of uncertain data should be completely random and not depend on available data, therefore uncertain observations are taken into the calculation by the discrete uniform distribution. For example, consider two questions on the survey. The first question asks if the answerer has a congenital disease. The second question asks on the religion that the respondent believes. Now suppose that there is an incomplete data because the respondent is not comfortable enough to answer what religion she belongs to but she answered that she doesn't have a congenital disease. If we use the EM algorithm, such an observation will be imputed back to the analysis by guessing the religion based on the proportion of congenital disease from the survey. This doesn't make sense in practice because the religion shouldn't be assumed to be proportional with other categorical variable. We then fix this issue by assigning such an observation back in the analysis by using the discrete uniform distribution. That is, we compute the updated total count by

$$Y_{ij} = \underbrace{y_{ij}}_{\text{complete part}} + \underbrace{y_{+j}\left(\frac{1}{m}\right)}_{\text{incomplete part }(M_A)} + \underbrace{y_{i+}\left(\frac{1}{n}\right)}_{\text{incomplete part }(M_B)}$$

Recall that m and n are numbers of levels so this is actually not the repeating algorithm. Once we obtain the updated Y_{ij}, the parameter of interest is calculated by $\widehat{\pi}_{ij} = y_{ij}/N$ as in the classical approach using the MLE.

4. Simulations

In this section, we use a simulation study to compare the performance of proposed two methods and the classical EM Algorithm. We consider nine models with different parameters as shown in Table 1. For each model, the ratio between the complete and incomplete data (C:I) is given. Models 1, 2 and 3 represent the cases of having more complete observations while models 4, 5 and 6 represent the cases of having complete and incomplete data on a par. The cases of having more incomplete data are demonstrated in models 7, 8 and 9. Another input of the simulation is the target parameter π_{ij}. We run a simulation study when $m = 2$ and $n = 2$ so there are four parameters in total which are $\pi_{11}, \pi_{12}, \pi_{21}$ and π_{22}. Three variations of parameters are evaluated. First, models 1, 4 and 7 use $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.25, 0.25, 0.25, 0.25)$, indicating the cases of equal parameters. The slight and strong distinctions in parameters are illustrated in the remaining six models by using $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.10, 0.20, 0.30, 0.40)$ and $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.05, 0.40, 0.05, 0.50)$, respectively.

Table 1. Parameters Setting on Nine Models.

	Ratio C:I	Parameters
Model 1	75 : 25	$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.25, 0.25, 0.25, 0.25)$
Model 2	75 : 25	$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.10, 0.20, 0.30, 0.40)$
Model 3	75 : 25	$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.05, 0.40, 0.05, 0.50)$
Model 4	50 : 50	$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.25, 0.25, 0.25, 0.25)$
Model 5	50 : 50	$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.10, 0.20, 0.30, 0.40)$
Model 6	50 : 50	$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.05, 0.40, 0.05, 0.50)$
Model 7	25 : 75	$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.25, 0.25, 0.25, 0.25)$
Model 8	25 : 75	$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.10, 0.20, 0.30, 0.40)$
Model 9	25 : 75	$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.05, 0.40, 0.05, 0.50)$

For each model we first generate dataset of sample size 1,000. If the observation is complete, it will be taken into the cell (i, j) with probability π_{ij}. If the observation is not complete, it will be taken into either M_A or M_B parts with equal probability. Then it will be assigned to the cell by using the marginal probability distribution obtained from parameters π_{ij}. Once the dataset is created, we apply the classical EM algorithm, the adapted EM algorithm and the uniform distribution approach as described in section 3. We then repeat creating dataset and apply three methods 500 times and compare the performance of each method by using the mean square error (MSE). Table 2 shows MSE of the parameter estimates from all methods for nine models, respectively. The lowest MSE of each parameter is marked by the star sign.

5. Real Data Examples

In this section, two scenarios are provided to illustrate the implementation of the methodology proposed in this article.

5.1. the General Social Survey Data

The General Social Survey (GSS) data is a major survey that has tracked American demographics, characteristics and views on social and cultural issues since the 1970s. The dataset contains 2,765 observations on about a dozen variables. The full description of the dataset can be found in [9]. One of the questions asked whether the respondent believed that permit is required to buy a gun (GunLaw) and also one question asked to the participant was whether he or she favored or opposed the death penalty for murder (DeathPenalty).

There are a lot of incomplete data in this survey. Out of 2,765 observations, we found that 1,421 respondent did not answer both GunLaw and DeathPenalty questions so those observations are excluded from the analysis. Out of the remaining observations, 428 participant answered the DeathPenalty question but not on the GunLaw question and

Table 2. Mean Square Errors for Estimators across Nine Models.

Model	Method	π_{11}	π_{12}	π_{21}	π_{22}
1	EM Algorithm	0.00025	0.00023	0.00021	0.00022
	Adapted EM	0.00020	0.00020	0.00018	0.00019
	Uniform Dist	0.00017*	0.00016*	0.00015*	0.00016*
2	EM Algorithm	0.00011	0.00018*	0.00022*	0.00025*
	Adapted EM	0.00010*	0.00018	0.00023	0.00026
	Uniform Dist	0.00043	0.00016	0.00020	0.00049
3	EM Algorithm	0.00006	0.00025	0.00006	0.00026
	Adapted EM	0.00004*	0.00025*	0.00005*	0.00025*
	Uniform Dist	0.00053	0.00044	0.00086	0.00135
4	EM Algorithm	0.00030	0.00031	0.00029	0.00029
	Adapted EM	0.00020	0.00021	0.00020	0.00021
	Uniform Dist	0.00013*	0.00013*	0.00012*	0.00013*
5	EM Algorithm	0.00016	0.00025	0.00030	0.00036*
	Adapted EM	0.00019	0.00029	0.00030	0.00037
	Uniform Dist	0.00015*	0.00024*	0.00027*	0.00157
6	EM Algorithm	0.00009	0.00031	0.00008	0.00030
	Adapted EM	0.00005*	0.00028*	0.00005*	0.00029*
	Uniform Dist	0.00194	0.00106	0.00313	0.00484
7	EM Algorithm	0.00045	0.00044	0.00044	0.00043
	Adapted EM	0.00022	0.00022	0.00022	0.00022
	Uniform Dist	0.00010*	0.00010*	0.00009*	0.00010*
8	EM Algorithm	0.00028*	0.00034*	0.00044*	0.00052*
	Adapted EM	0.00032	0.00038	0.00048	0.00053
	Uniform Dist	0.00324	0.00043	0.00045	0.00326
9	EM Algorithm	0.00013	0.00044	0.00013	0.00047
	Adapted EM	0.00005*	0.00036*	0.00007*	0.00040*
	Uniform Dist	0.00443	0.00227	0.00723	0.01099

36 respondent answered the GunLaw question but left the DeathPenalty question empty. Our goal is to use the classical method and the two proposed methods to this dataset. First, we tabulate the 2 × 2 contingency table as shown in Table 3. The fully classified data were used to determine an initial value for the algorithm. That is,

$$\pi_{ij}^{(0)} = \left(\frac{494}{880}, \frac{130}{880}, \frac{212}{880}, \frac{44}{880}\right) = (0.5614, 0.1477, 0.2409, 0.0500)$$

Table 3. Contingency Table for GunLaw and DeathPenalty Variables

	Favor GunLaw	Oppose GunLaw	Missing
Favor DeathPenalty	494	130	275
Oppose DeathPenalty	212	44	153
Missing	31	5	

We then perform the classical EM algorithm, adapted EM algorithm and discrete uniform distribution approach. Table 4 shows the estimated parameters and the p-value from the Chi-squared test after applying each method.

Table 4. Estimated Parameters on GSS Example

Method	π_{11}	π_{12}	π_{21}	π_{22}	p-value
Ignore Uncertain Data	0.5614	0.1477	0.2409	0.0500	0.2175
EM Algorithm	0.5456	0.1416	0.2597	0.0530	0.1186
Adapted EM	0.5480	0.1393	0.2565	0.0562	0.3228
Uniform Dist	0.4814	0.2009	0.2262	0.0915	0.8107

5.2. Victimization Status Data

We consider the data obtained through the National Crime Survey conducted by the United States Bureau of the Census. This is a classic example provided in [10]. In this study, surveys about victimization status were done twice. At the first time stamp, housing unit occupants were interviewed to determine if they had been victimized by crime. Then, six months after, they were asked again on the same question. The result is shown in Table 5.

Table 5. Contingency Table on Victimization Status

	Crime-free on 2^{nd} visit	Victims on 2^{nd} visit	Missing
Crime-free on 1^{nd} visit	392	55	33
Victims on 1^{nd} visit	76	38	9
Missing	31	7	

As described in the previous subsection, the fully classified data were used to determine an initial value for the algorithm. That is,

$$\pi_{ij}^{(0)} = \left(\frac{392}{561}, \frac{55}{561}, \frac{76}{561}, \frac{38}{561}\right) = (0.6988, 0.0980, 0.1355, 0.0677)$$

We then perform the classical EM algorithm, adapted EM algorithm and discrete uniform distribution approach. Table 6 shows the estimated parameters and the p-value from the Chi-squared test after applying each method.

Table 6. Estimated Parameters on Victimization Status Example

Method	π_{11}	π_{12}	π_{21}	π_{22}	p-value
Ignore Uncertain Data	0.6988	0.0980	0.1355	0.0677	8.36×10^{-9}
EM Algorithm	0.6971	0.0986	0.1358	0.0685	7.47×10^{-9}
Adapted EM	0.6929	0.1031	0.1401	0.0639	5.94×10^{-7}
Uniform Dist	0.6615	0.1170	0.1498	0.0718	3.08×10^{-6}

6. Conclusion and Discussion

In this study, we proposed two new methods to impute or return the incomplete data in the analysis of contingency table. Although these two proposed methods are just minor changes in the formula of the E-step in the EM algorithm but they give few interesting insights. Based on our simulation in section 4, the algorithm with the best performance, in terms of MSE, can be summarized in Table 7.

Table 7. Summary of Best Method for Nine Models

	Ratio C:I	Variation in Parameters	Best Method
Model 1	75 : 25	None	Uniform Dist
Model 2	75 : 25	Weak	Classical EM
Model 3	75 : 25	Strong	Adapted EM
Model 4	50 : 50	None	Uniform Dist
Model 5	50 : 50	Weak	Uniform Dist
Model 6	50 : 50	Strong	Adapted EM
Model 7	25 : 75	None	Uniform Dist
Model 8	25 : 75	Weak	Classical EM
Model 9	25 : 75	Strong	Adapted EM

We can see that the adapted EM algorithm, proposed in section 3.1, performs best when dealing with the strong variation in parameters, no matter what the ratio of the complete and incomplete data is. This method should be recommended when working with 2×2 contingency tables with an expectation of very distinct parameters. On the other hand, the uniform distribution approach performs better than others when there is no variation in parameters. This gives no surprise as it should be from the definition of the discrete uniform distribution. When there is a weak sign of variation in parameters, the classical EM algorithm works best, except the case that the ratio of the complete and incomplete data is 1:1. If there is an approximately equal portion of complete and incomplete data and weak sign of variation in parameters, the uniform distribution approach is preferred as shown in the result from model 5.

For the real data examples in section 5, the pattern of the variation in parameters is strong and the ratio of complete and incomplete data is close to 75 : 25 so model 3 would be the best fit to both datasets. We expect the adapted EM to work better than other approaches. The p-value obtained from adapted EM is more reliable compared to classical EM and uniform distribution approaches. For GSS example, the p-value is 0.3228 so we fail to reject the null hypothesis and conclude that there is no association between GunLaw and DeathPenalty variables. For the victimization status example, the p-value is 5.94×10^{-7} which leads to a conclusion of rejecting the null hypothesis. This means that we conclude the independence between the victimization status on the first visit and on the second visit.

From these examples, we can conclude clearly that different methods in imputing missing data could end up with very distinct p-values. This could potentially lead to the different conclusions, either reject the null hypothesis or fail to reject the null hypothesis. Thus, researchers should be aware and careful to pick the most appropriate method. Table 7 from the simulation is an initial guideline. Note that there is no limitation in using each method.

Lastly, we wrap up with an observation from both real data examples. From three methods, ignoring the missing data, classical EM algorithm and adapted EM algorithm, a slight difference in the estimate of parameters π_{ij} was detected. However, the results for the uniform distribution approach are very different from others. The uniform distribution approach seems to be off on these examples which is no surprise as we have an indication from Table 3 and Table 5 that the population parameters should have at least about moderate to strong variation.

Acknowledgements

The author is grateful to the editor, the associate editor, and the referees for their insightful comments and suggestions, which greatly improved this article.

References

[1] R.J. Little, D.B. Rubin, Statistical analysis with missing data, Vol. 793, John Wiley & Sons, 2019,

[2] R. Ehlers, Maximum likelihood estimation procedures for categorical data, Diss. University of Pretoria, 2005.

[3] K. Takai, Y. Yano, Test of independence in a 2 × 2 contingency table with nonignorable nonresponse via constrained EM algorithm. Computational Statistics & Data Analysis Vol.52 (2008) 5229-5241.

[4] S. Petitrenaud, Independence tests for uncertain data with a frequentist method, Combining Soft Computing and Statistical Methods in Data Analysis, Springer, Berlin, Heidelberg (2010), 519-526.

[5] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society: Series B (Methodological) 39.1 (1977) 1-22.

[6] G. Casella, R.L. Berger, Statistical Inference, Cengage Learning, 2021.

[7] P. Brmaud, An Introduction to probabilistic modeling, Springer Science & Business Media, 2012.

[8] G.J. McLachlan, T. Krishnan, The EM algorithm and extensions, John Wiley & Sons, 2007.

[9] L.M. Chihara, T.C. Hesterberg, Mathematical statistics with resampling and R, John Wiley & Sons, 2018.

[10] J.L. Schafer, Analysis of incomplete multivariate data, CRC press, 1997.