Asymptotic efficiency of M.L.E. using prior survey in multinomial distributions

Sheena Yo\textsuperscript{a,b}

\textsuperscript{a}Faculty of Economics, Shinshu University, Matsumoto, Japan; \textsuperscript{b}Faculty of Data Science, Shiga University, Shiga, Japan

ABSTRACT
Incorporating information from a prior survey is generally supposed to decrease the estimation risk of the present survey. This paper aims to show how the risk changes by incorporating the information of a prior survey through watching the first and the second-order terms of the asymptotic expansion of the risk. We recognize that the prior information is of some help for risk reduction when we can acquire samples of a sufficient size for both surveys. Interestingly, when the sample size of the present survey is small, the use of the prior survey can increase the risk. In other words, blending information from both surveys can have a negative effect on the risk. Based on these observations, we give some suggestions on whether or not to use the results of the prior survey and the sample size to use in the surveys for a reliable estimation.

ARTICLE HISTORY
Received 7 September 2019
Accepted 3 April 2020

KEYWORDS
Kullback-Leibler divergence; estimation risk; asymptotic expansion

MSC(2010) SUBJECT CLASSIFICATION
Primary 62F10; Secondary 62F12

1. Introduction
Combining data has been an important topic in statistics for a long time. In 1950s, some influential results on the topic like Cochran (1954), Cochran and Cox (1957), Good (1950) were published. Later on, the scope and depth of the research over the topic has developed and combining data becomes indispensable method in various fields. Meta-analysis is now an important issue in many statistical fields not to mention its usefulness in medical research. In the field of survey sampling, the use of the auxiliary data, i.e., the combination of principle and other data, is a basic tool for efficient sampling (e.g., see Chapter 14 of Valliant, Dever, and Kreuter (2013)). We also notice that combining data from existing multiple surveys has been a challenging task for econometrics (e.g., see Ridder and Moffitt (2007)).

In combining data, we have to pay attention to “homogeneity” between the data sets, which is defined in various ways according to the purpose of the research. Generally, it indicates there is some similarity in their distributions. It is noted that even if the two datasets are sampled from the same population, different sampling scheme could produce the different sample distributions. Anyway the more homogeneity exists, the more easily we can synthesize information.

In this paper, we suppose the following homogeneity between two datasets.
(1) Each dataset has a finite-dimensional discrete distribution (i.e. multinomial distribution).

(2) The categories (bins) of the two distributions have inclusion relations. Namely each category of one multinomial distribution is composed by merging some categories of another multinomial distribution. In this case, we say that the former distribution is “coarse” to the latter.

Because the homogeneity above is rather strong, the method of combining two datasets are simple; aggregation of the numbers of the two datasets for the estimation of the “large” categories (the large categories mean those of the coarse distribution). We suppose that this aggregation enables us to make more efficient estimation not only for the large categories but also small ones (see the two-stage multinomial distribution explained later).

For more practical situations, assume we have just completed one survey (“present survey”). A sampled individual in the present survey is classified into certain categories and we want to estimate the probability of each category. Besides we have another survey (“prior survey”) which is coarse to the present survey. Namely the prior survey is conducted with the same population and sampling method, but its categories are coarse to the present survey. For example, the present survey categorize people according to the age, gender and occupation, while the prior survey just checks the gender. We want to estimate the probability over small categories (each of which is specified by three attributes) by using both surveys. (Note that the words “present” and “prior” do not necessarily mean the chronological order.)

Since both the prior and present surveys give us information on the distributions over the gender (the large category), if we aggregate both numbers there, it is supposed to enhance the estimation efficiency compared to the estimation based only on the present survey. How much more estimation efficiency is gained by using the information from prior survey? We try to answer this question in this paper.

Note that the concept “coarse” can be defined to the general distributions including continuous ones. The similar argument to the present paper might be possible for the continuous distributions, but we focus on the discrete case here.

Now we state our estimation problem in a more formal way. Suppose the multinomial distribution model over the categories (bins) \( C_i, i = 1, \ldots, p + 1 \), where the parameter (the probability for each category) is given by

\[
m \triangleq (m_1, \ldots, m_{p+1})
\]

If there are no restrictions other than

\[
\sum_{i=1}^{p+1} m_i = 1, \quad m_i > 0, \quad i = 1, \ldots, p + 1
\]

we call the distribution model a “full model” (the dimension of the model equals \( p \)), while if some restrictions on the parameter reduce the model dimension to less than \( p \), we call the model a “submodel.”

Let \( \hat{m} \triangleq (\hat{m}_1, \ldots, \hat{m}_{p+1}) \) be maximum likelihood estimator (MLE) of \( m \triangleq (m_1, \ldots, m_{p+1}) \). We evaluate the discrepancy between the predictive distribution given by MLE and the true distribution by using Kullback–Leibler divergence:
Because Kullback–Leibler divergence is an “α-divergence” with \( \alpha = -1 \), we will use the notation \( D^1 \). (E.g., see Amari (2016) for the \( \alpha \)-divergence.) We evaluate the performance of MLE through its risk, that is,

\[
E[D^{1} | \hat{m} : m] = \sum_{i=1}^{p+1} \hat{m}_i \log \frac{\hat{m}_i}{m_i}
\]

It is difficult to obtain the explicit form of the risk. In this paper, alternatively, we derive the asymptotic expansion of the risk with respect to the sample size \( n \) up to the second-order term. The first-order term \( (n^{-1}-\text{order term}) \) and the second-order term \( (n^{-2}-\text{order term}) \) give us important information on the asymptotic performance of MLE.

For the full model, the asymptotic expansion with respect to the sample size \( n \) is given by

\[
E[D^{1} | \hat{m} : m] \approx E[D^{1} | \hat{m} : m] - \frac{p}{2n} + \frac{1}{12n^2}(M - 1) + o(n^{-2})
\]

where

\[
M \triangleq \sum_{i=1}^{p+1} m_i^{-1}
\]

(see (42) of Sheena (2018)). The first-order term is determined by the ratio of the model’s dimension to the sample size (“\( p - n \) ratio”). Because this holds true for any parametric model, the first-order term for a submodel of \( p' \)-dimensions equals \( p'/(2n) \). However, the second-order term of most submodels is too complicated to be derived explicitly. (See (25) of Sheena (2018). The asymptotic expansion of the risk with respect to \( \alpha \)-divergence for a general parametric distribution is given in Theorem 1 of Sheena (2018).)

Suppose that random sampling is carried out from the multinomial distribution given by (1). We will call this sampling the “present” survey. Additionally, suppose that there is another random sampling from the same population (“prior” survey) and that the inclusion relation exists between the present and prior surveys as follows. The present and prior surveys are conducted over the categories \( C_i, i = 1, \ldots, p \) and \( C^*_j, j = 1, \ldots, I \), respectively, and each \( C^*_j \) is a union of some \( C_i \)'s. Namely the prior survey is coarse compared with the present survey.

To clarify the relationship of the categories between the two surveys, let us consider the present survey as a two-stage model. We use the two-stage notation \( C_{ij}, i = 1, \ldots, I, j = 1, \ldots, J_i \) for the categories of the present survey, where for \( i = 1, \ldots, I \),

\[
C_i \triangleq \bigcup_{j=1, \ldots, J_i} C_{ij}, \quad C_i = C^*_i
\]

Consequently, the first index \( i \) of \( C_{ij} \) tells to which \( C_i = C^*_i \) an observation belongs. The second index \( j \) runs over \( J_i \) categories that belong to \( C_i \). Note that when \( J_i = 1 \), the category \( C_i \) consists of a single \( C_{i1} \).

The definition of the two-stage multinomial distribution model is described as follows. Suppose that random variable \( X \) takes values that belong to one of the categories \( C_{ij} \) with the probabilities
\[ m_{ij} \triangleq P(X \in C_{ij}) \]  
for \( i = 1, \ldots, I, j = 1, \ldots, J_i \).

We define the related probabilities as

\[ m_{i} \triangleq \sum_{j=1}^{J_i} m_{ij}, \quad p_{ij} \triangleq \frac{m_{ij}}{m_i} \]  
for \( i = 1, \ldots, I, j = 1, \ldots, J_i \). The first-stage model is given by focusing on which \( C_i \) the value of \( X \) belongs to. Its parameters (probabilities for each category) are given by

\[ m_f \triangleq (m_1, \ldots, m_I) \]

In this paper, we only consider the case where the first-stage model is a full model. This illustrates well how the information on the prior survey changes the estimation risk of the present survey.

The \( i \)-th model \((i = 1, \ldots, I)\) on the second stage is given by the categories \( C_{ij} \), \( j = 1, \ldots, J_i \) and the corresponding probabilities

\[ p_i \triangleq (p_{i1}, \ldots, p_{ij}) \]

under the condition \( X \in C_i \). We use \( s_i \) for the dimension of the \( i \)-th model, hence if \( s_i = J_i - 1 \), it is a full model. Note that if the \( i \)-th second-stage model consists of a single category, i.e., \( I_i = 1, p_{i1} = 1 \), there are no unknown parameters for the model.

Correspondingly to the two-stage model, the two-stage estimator \( \hat{m}_{ij} \) of \( m_{ij} \) is given by

\[ \hat{m}_{ij} = \hat{m}_i \hat{p}_{ij}, \quad i = 1, \ldots, I, j = 1, \ldots, J_i \]  
where \( \hat{m}_i \) is the estimator for the first-stage model and \( \hat{p}_{ij} \) is the estimator for the \( i \)-th \((i = 1, \ldots, I)\) second-stage model. Note that for the \( i \)-th second-stage model with \( I_i = 1 \), the estimator \( \hat{p}_{i1} = 1 \).

Let \( x_{ij} \) and \( x_i \) \((i = 1, \ldots, I, j = 1, \ldots, J_i)\) denote the number of observations in the present survey which belong to \( C_{ij} \) and \( C_i \), respectively. Hence,

\[ x_i = \sum_{j=1}^{J_i} x_{ij}, \quad \sum_{i=1}^{I} x_i = n \text{ (sample size)} \]

The prior survey, which is coarse, is only related to the first-stage model. Let \( x_i^* \) denote the number of observations in the prior survey that belong to \( C_i \) \((i = 1, \ldots, I)\). The sample size of the prior survey is denoted by \( n^* \). The notations of \( X_{ij}, X_i, X_i^* \) are similarly defined as the corresponding random variables.

With two sets of samples \( x_i, i = 1, \ldots, I \) and \( x_i^*, i = 1, \ldots, I \), there are three possible MLEs for the first-stage parameter \( m = (m_1, \ldots, m_I) \). (We use capital letters for the corresponding random variables.)

1. MLE from the present sample:

\[ \hat{m}_f \triangleq (\hat{m}_1, \ldots, \hat{m}_I), \quad \hat{m}_i = X_i / n, i = 1, \ldots, I \]
MLE from the prior sample:
\[ \hat{m}_i^* \triangleq (\hat{m}_1^*, ..., \hat{m}_t^*), \quad \hat{m}_i^* = X_i^* / n^*, i = 1, ..., I \]

(3) MLE from the pooled sample:
\[ \hat{m}_i^p \triangleq (\hat{m}_1^p, ..., \hat{m}_t^p), \quad \hat{m}_i^p = X_i + X_i^* / (n + n^*) \]

MLE for the \( i \)-th second-stage model (\( i = 1, ..., I \)) based on the present sample is given by
\[ \hat{p}_{ij} \triangleq X_{ij}/X_i, \quad j = 1, ..., J_i \]

If we combine this estimator with each estimator for the first-stage model, we have three estimators for \( m_{ij} (i = 1, ..., I, j = 1, ..., J_i) \):
\[ \hat{m}_{ij} \triangleq \hat{m}_{i} \hat{p}_{ij}, \quad \hat{m} (\hat{m}_{ij})_{1 \leq i \leq I, 1 \leq j \leq J} \]
\[ \hat{m}_{ij}^* \triangleq \hat{m}_{i}^* \hat{p}_{ij}, \quad \hat{m}^* (\hat{m}_{ij}^*)_{1 \leq i \leq I, 1 \leq j \leq J} \]
\[ \hat{m}_{ij}^p \triangleq \hat{m}_{i}^p \hat{p}_{ij}, \quad \hat{m}^p (\hat{m}_{ij}^p)_{1 \leq i \leq I, 1 \leq j \leq J} \]

This study aims to compare the risks of these three estimators. Asymptotic expansion of the risk for the three estimators is given in Section 2.1. Comparison of the first and second-order terms give us some insight on the asymptotic behavior of the estimators. In Section 2.2, we observe the performance of the estimators through three examples. The findings from our simulation studies supplement the theoretical results.

2. Main results

2.1. Theoretical result

In this section, we derive an asymptotic expansion of the risk of three estimators \( \hat{m}, \hat{m}^*, \) and \( \hat{m}^p \). We use the following decomposition of Kullback–Leibler divergence \( D^1[\hat{m} : m] \). Let

\[ m \triangleq (m_{ij})_{1 \leq i \leq I, 1 \leq j \leq J}, \quad \hat{m} \triangleq (\hat{m}_{ij})_{1 \leq i \leq I, 1 \leq j \leq J} \]

then from (7), we have
\[
\begin{align*}
\bar{D}^{-1} \left[ \hat{m} : m \right] &= \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{m}_{ij} \log \left( \hat{m}_{ij} / m_{ij} \right) \\
&= \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{m}_{i} \hat{p}_{ij} \log \left( \hat{m}_{i} \hat{p}_{ij} / m_{i} p_{ij} \right) \\
&= \sum_{i=1}^{I} \hat{m}_{i} \log \left( \hat{m}_{i} / m_{i} \right) + \sum_{i=1}^{I} \hat{m}_{i} \sum_{j=1}^{J} \hat{p}_{ij} \log \left( \hat{p}_{ij} / p_{ij} \right) \\
&= \bar{D}^{-1} \left[ \hat{m}_f : m_f \right] + \sum_{i=1}^{I} \hat{m}_{i} \bar{D}^{-1} \left[ \hat{p}_i : p_i \right],
\end{align*}
\]

where
The decomposition is called the “chain rule” and is one of the appealing characteristics of the Kullback–Leibler divergence. It should be noted that it is possible that 

\[ X_i = 0 \]

Hence we can encounter the situation we are unable to estimate \( p_{ij}, j = 1, \ldots, \), the parameter of the \( i \)-th second-stage model because of no sample. As a remedy for this problem, we make it a rule to discard such samples with no estimation. In the following theorems, we adopt this rule, hence expectations taken are all conditional on the state 

\[ X_i \neq 0, 1 \]

However, as Lemma 1 in Appendix shows, the conditional expectations of Kullback–Leibler divergence and 

\[ (X_i - nm_{ik})^k (k = 1, 2, \ldots) \]

differ from those unconditional by \( o(n^{-s}) \) for any \( s > 0 \). Therefore, we use the same notation as that for the unconditional distribution.

The risk for each estimator is given in the following three theorems. We use the notation \( A(i), i = 1, \ldots, I \) for the coefficient of \( 1/(24n^2) \) in the asymptotic expansion of the risk of MLE for the \( i \)-th second-stage model when the sample size is given as \( n \). Because the dimension of the \( i \)-th model equals \( s_i \), the asymptotic expansion of the risk of MLE with respect to \( n \) is given by

\[
\frac{1}{ED} [\hat{p}_i : p_i] = \frac{1}{2n} + \frac{1}{24n^2} A(i) + o(n^{-2})
\]

(See (25) of Sheena (2018).)

**Theorem 1.** If we use the estimator (8), then the risk is given by

\[
\frac{1}{ED} [\hat{m} : m] = \frac{p'}{2n} + \frac{1}{24n^2} \left\{ \sum_{i=1}^{I} m_i^{-1} \left( A(i) + 2 \right) - 2 \right\} + o(n^{-2})
\]

\[
= \frac{1}{2n} \left( I - 1 + \sum_{i=1}^{I} s_i \right) + \frac{1}{12n^2} (M_f - 1) + \frac{1}{24n^2} \left( \sum_{i=1}^{I} m_i^{-1} A(i) \right) + o(n^{-2}),
\]

(13)

where

\[ p' \triangleq I - 1 + \sum_{i=1}^{I} s_i, \quad M_f \triangleq \sum_{i=1}^{I} m_i^{-1} \]

In the special case, when the second-stage models are all full models, the risk is given by

\[
\frac{1}{ED} [\hat{m} : m] = \frac{p}{2n} + \frac{1}{12n^2} \left\{ \sum_{i=1}^{I} \sum_{j=1}^{J_i} m_{ij}^{-1} - 1 \right\} + o(n^{-2})
\]

(14)

where

\[ p \triangleq \sum_{i=1}^{I} J_i - 1 \]
Proof. From (11), we have

\[
\frac{-1}{ED [\hat{m} : m]}
\]

\[
\Delta E \left[ \frac{-1}{D [\hat{m} : m]} \right]
\]

\[
= E_{X_1, \ldots, X_l} \left[ E \left[ \frac{-1}{D [\hat{m} : m]} | X_1, \ldots, X_l \right] \right]
\]

\[
= E_{X_1, \ldots, X_l} \left[ E \left[ \frac{-1}{D [\hat{m}_f : m_f]} + \sum_{i=1}^{l} (X_i / n) \frac{-1}{D [\hat{p}_i : p_i] | X_1, \ldots, X_l] \right] \right]
\]

\[
= \frac{-1}{ED_f} + \sum_{i=1}^{l} E_{X_1, \ldots, X_l} \left[ (X_i / n) \frac{-1}{ED_{(i)}(X_i)]} \right]
\]

where

\[
\frac{-1}{ED_f} \Delta E \left[ \frac{-1}{D [\hat{m}_f : m_f]} \right]
\]

\[
\frac{-1}{ED_{(i)}(X_i)]} \Delta E \left[ \frac{-1}{D [\hat{p}_i : p_i] \right] |_{n=X_i}
\]

Note that in general, \( E[D [\hat{p}_i : p_i]] \) is the function of the size \( n \) of the sample on which \( \hat{p}_i \) are based. Here, in the two-stage model, size \( n \) equals to \( X_i \).

Because

\[
\frac{-1}{ED_{(i)}(X_i)} = \frac{s_i}{2X_i} + \frac{1}{24X_i^2} \frac{-1}{A(i)} + o(n^{-2}) |_{n=X_i}
\]

we have

\[
E_{X_1, \ldots, X_l} \left[ (X_i / n) \frac{-1}{ED_{(i)}(X_i)]} \right]
\]

\[
= \frac{s_i}{2n} + \frac{1}{24n} \frac{-1}{A(i)} E[X_i^{-1}] + E[o_p(n^{-2})]
\]

For the derivation of the relation

\[
o(n^{-2}) |_{n=X_i} = o_p(n^{-2})
\]

we used the Lemma 2 in Appendix and the fact \( X_i / n \) converges to \( m_i \) in probability.

Taylor expansion of \( 1/X_i \) around \( 1/(nm_i) \) is given by

\[
\frac{1}{X_i} = \frac{1}{nm_i} - \frac{1}{(nm_i)^2} (X_i - nm_i) + O_p(n^{-2})
\]

Because \( E[X_i - nm_i] = 0 \), we have

\[
E[X_i^{-1}] = \frac{1}{nm_i} + O(n^{-2})
\]

and

\[
E_{X_1, \ldots, X_l} \left[ (X_i / n) \frac{1}{ED_{(i)}(X_i)} \right] = \frac{s_i}{2n} + \frac{1}{24n^2} m_i^{-1} A(i) + o(n^{-2})
\]

(We omit the proof that \( E[O_p(n^{-2})] = O(n^{-2}) \) and \( E[o_p(n^{-2})] = o(n^{-2}) \) hold for the present case.)
Consequently,

$$\bar{ED} \left[ \hat{m} : m \right] = \bar{ED}_I + \sum_{i=1}^{l} \left( \frac{s_i}{2n} + \frac{1}{24n^2} m_i^{-1} \bar{A}_{(i)} \right) + o(n^{-2}) \quad (17)$$

If we apply (3) to $\bar{ED}_I$, we obtain

$$\bar{ED}_I = \frac{I - 1}{2n} + \frac{1}{12n^2} (M_f - 1) + o(n^{-2})$$

If we insert this result into (17), we have (13).

When the second-stage models are all full models, $s_i = J_i - 1, i = 1, ..., I$. From (3), it turns out that

$$\sum_{i=1}^{l} m_i^{-1} \bar{A}_{(i)} = \sum_{i=1}^{l} m_i^{-1} \left( \sum_{j=1}^{l} 2p_{ij}^{-1} - 2 \right) = 2 \sum_{i=1}^{l} \sum_{j=1}^{l} m_i^{-1} - 2M_f$$

If we insert these results into (13), we have (14). (Because the total model is a full model in this case, we notice that (14) could be derived from (3) straightforwardly.)

**Theorem 2.** If we use the estimator (9), then the risk is given by

$$\bar{ED} \left[ \hat{m}^* : m \right] = \frac{1}{2n^*} (I - 1) + \frac{1}{2n} s_i + \frac{1}{12n^{*2}} (M_f - 1)$$

$$+ \frac{1}{24n^2} \sum_{i=1}^{l} m_i^{-1} \left( \bar{A}_{(i)} + 12(1 - m_i) s_i \right) + o(n^{-2}) + o(n^{*-2}) \quad (18)$$

In the special case, when the second-stage models are all full models, the risk is given by

$$\bar{ED} \left[ \hat{m}^* : m \right] = \frac{1}{2n^*} (I - 1) + \frac{1}{2n} (p + 1 - I)$$

$$+ \frac{1}{12n^{*2}} (M_f - 1) + \frac{1}{12n^2} \left( M + \sum_{i=1}^{l} m_i^{-1} (6J_i - 7) - 6(p + 1 - I) \right)$$

$$+ o(n^{-2}) + o(n^{*-2}) \quad (19)$$

**Proof.** From (11), we have

$$\bar{ED} \left[ \hat{m}^* : m \right]$$

$$\Delta \bar{D}[\hat{m}^* : m]$$

$$= E_{X_1, ..., X_l} \left[ E_{X_1, ..., X_l} \left[ \bar{D}[\hat{m}^* : m] | X_1, ..., X_l \right] | X_1^*, ..., X_l^* \right]$$

$$= E_{X_1, ..., X_l} \left[ E_{X_1, ..., X_l} \left[ \bar{D}[\hat{m}^* : m_f] + \sum_{i=1}^{l} \left( X_i^* / n^* \right) \bar{D}[\hat{p}_i : p_i] | X_1, ..., X_l | X_1^*, ..., X_l^* \right] \right]$$

$$= \bar{ED}_f + \sum_{i=1}^{l} E_{X_1, ..., X_l} \left[ X_i^* \left( X_i^* / n^* \right) \bar{ED}_i(X_i) | X_1^*, ..., X_l^* \right]$$
where
\[
\begin{align*}
\bar{ED}^{1}_{j} & \triangleq E \left[ D \left[ \hat{m}^{*}_{j} : m_{j} \right] \right] \\
\bar{ED}^{1}_{(i)}(X_{i}) & \triangleq E \left[ D \left[ \hat{p}_{i} : p_{i} \right] \right] \big|_{n=X_{i}}
\end{align*}
\]
Because
\[
\bar{ED}^{1}_{(i)}(X_{i}) = \frac{s_{i}}{2X_{i}} + \frac{1}{24X_{i}^{2}} A^{1}_{(i)} + o(n^{-2}) \big|_{n=X_{i}}
\]
we have
\[
\begin{align*}
E_{X_{1},\ldots,X_{i}} \left[ E_{X_{1},\ldots,X_{i}} \left[ (X_{i}^{*}/n^{*}) \bar{ED}^{1}_{(i)}(X_{i}) | X_{1}^{*},\ldots,X_{i}^{*} \right] \right] \\
& = \frac{s_{i}}{2n^{*}} E[X_{i}^{*}] E[X_{i}^{-1}] + \frac{1}{24n^{*}} E[X_{i}^{*}] A^{1}_{(i)} E[X_{i}^{-2}] + E[o_{p}(n^{-2})] \\
& = \frac{s_{i}m_{i}}{2} E[X_{i}^{-1}] + \frac{m_{i}}{24} A^{1}_{(i)} E[X_{i}^{-2}] + E[o_{p}(n^{-2})]
\end{align*}
\]
Taylor expansions of $1/X_{i}$ and $1/X_{i}^{2}$ are given by
\[
1/X_{i}^{2} = \frac{1}{nm_{i}} - \frac{1}{(nm_{i})^{2}} (X_{i} - nm_{i}) + \frac{1}{(nm_{i})^{3}} (X_{i} - nm_{i})^{2} + O_{p}(n^{-5/2})
\]
Using $E[X_{i} - nm_{i}] = 0$ and $E[(X_{i} - nm_{i})^{2}] = nm_{i}(1 - m_{i})$, we have
\[
\begin{align*}
E[X_{i}^{-1}] &= \frac{1}{nm_{i}} + \frac{1}{(nm_{i})^{2}} + O(n^{-5/2}) \\
E[X_{i}^{-2}] &= \frac{1}{(nm_{i})^{2}} + O(n^{-5/2})
\end{align*} \tag{20}
\]
(We omit the proof that $E[o_{p}(n^{-5/2})] = O(n^{-5/2})$ holds for the present case.) From (20) and (21), we have
\[
\begin{align*}
E_{X_{1},\ldots,X_{i}} \left[ E_{X_{1},\ldots,X_{i}} \left[ (X_{i}^{*}/n^{*}) \bar{ED}^{1}_{(i)}(X_{i}) | X_{1}^{*},\ldots,X_{i}^{*} \right] \right] \\
& = \frac{s_{i}}{2n} + \frac{1}{24n^{2}} m_{i}^{-1}(12s_{i}(1 - m_{i}) + A^{1}_{(i)}) + o(n^{-2})
\end{align*}
\]
(We omit the proof that $E[o_{p}(n^{-2})] = o(n^{-2})$ holds for the present case.) Consequently,
\[
\bar{ED} \left[ \hat{m}^{*} : m \right] = \bar{ED}^{1}_{j} + \sum_{i=1}^{l} \left( \frac{s_{i}}{2n} + \frac{1}{24n^{2}} m_{i}^{-1}(12s_{i}(1 - m_{i}) + A^{1}_{(i)}) \right) + o(n^{-2}) \tag{22}
\]
If we apply (3) to $ED_j$, we obtain

$$ED_j^{-1} = \frac{I-1}{2n^*} + \frac{1}{12(n^*)^2}(M_f - 1) + o((n^*)^{-2})$$

If we insert this result in the right-hand side of (22), we obtain (18).

By inserting $s_i = I_i - 1$ and $A(i) = 2m_i \sum_{j=1}^{h} m_{ij}^{-1} - 2$ into (18), we obtain (19).

We compare the risks of the two estimators $\hat{m}$ and $\hat{m}^*$. From (13) and (18), we notice that

$$ED[\hat{m} : m] - ED[\hat{m}^* : m] = \frac{I-1}{2} \left( \frac{1}{n} - \frac{1}{n^*} \right) + \frac{M_f-1}{12} \left( \frac{1}{n^*} - \frac{1}{(n^*)^2} \right) - \frac{1}{2n^2} \left( \sum_{i=1}^{I} s_i m_i^{-1} (1 - m_i) \right)$$

$$+ o(n^{-2}) + o(n^{-2})$$

$$= \frac{I-1}{2n} \frac{\beta - 1}{\beta} + \frac{1}{12n^2} \left( \frac{\beta^2 - 1}{\beta^2} (M_f - 1) - 6 \sum_{i=1}^{I} s_i m_i^{-1} (1 - m_i) \right)$$

$$+ o(n^{-2})$$

where $\beta \triangleq n^*/n$.

This equation gives us the following suggestions on whether we should use the prior information or not:

- The difference in the first-order term, $(I - 1)(n^{-1} - n^{*-1})/2$, leads us to the simple conclusion that if $n^* > n$ and both are large enough, it is better to use $\hat{m}^*$.
- To see the effect of the difference in the second-order term, we let $n^* = n$, then the difference equals

$$- \frac{1}{2n^2} \left( \sum_{i=1}^{I} s_i (m_i^{-1} - 1) \right) < 0$$

This suggests that if the prior survey is done in a similar scale $(n^* = n)$, then it is better to use $\hat{m}$. The "unified" estimation using the same sample for both stages has an advantage over the "non-unified" estimation that uses the different sample sets for the two stages. (We confirm this observation in the next subsection by simulation.) Notice that in (11), each $D[p_i : p_i]$ is multiplied by $\hat{m}_i$. If $x_i$, the sample size used for the $i$-th second-stage model estimation, is small, then $\hat{m}_i = x_i/n$ also gets small. Hence, the loss $D[p_i : p_i]$ is devalued in the total loss evaluation $D[\hat{m} : \hat{m}]$ in the unified approach (8). This is natural because when the sample size is small, we should not rely much on the result. This mechanism does not work if we use independent sets of samples for the first and second stages as in the non-unified approach. From (24), we notice that this
negative effect of non-unified approach is strong when the first-stage model contains a
category of small probability (i.e., \( m_i^{-1} \) is large) and especially if the corresponding
second-stage model has many parameters to be estimated (i.e., \( s_i \) is large).

The risk for the estimation using \( \hat{m}^p \) is given by the following theorem.

**Theorem 3.** If we use the estimator (10), then the risk is given by

\[
\text{ED}[\hat{m}^p : m] = \frac{1}{2(n + n^*)}(I - 1) + \frac{1}{2n} \sum_{i=1}^{I} s_i + \frac{1}{12(n + n^*)^2}(M_f - 1) \\
+ \frac{1}{24n^2} \sum_{i=1}^{I} m_i^{-1} \left( A_{(i)} - 1 + 12(1 - m_i) s_in^*/(n + n^*) \right) \\
+ o(n^{-2}) + o((n + n^*)^{-2})
\]

(25)

In the special case, when the second-stage models are all full models, the total risk is given
by

\[
\text{ED}[\hat{m}^p : m] = \frac{1}{2(n + n^*)}(I - 1) + \frac{1}{2n} (p - I + 1) + \frac{1}{12(n + n^*)^2}(M_f - 1) \\
+ \frac{1}{12n^2} \left( M - M_f + \frac{6n^*}{n + n^*} \left( \sum_{i=1}^{I} m_i^{-1}I_i - M_f - (p - I + 1) \right) \right) \\
+ o(n^{-2}) + o((n + n^*)^{-2})
\]

(26)

**Proof.** As in (15), we have

\[
\begin{align*}
\text{ED}[\hat{m}^p : m] & \triangleq E[\text{ED}[\hat{m}^p : m]] \\
& = E[X_1, \ldots, X_I] [E[\text{ED}[\hat{m}^p : m] | X_1, \ldots, X_I] | X_1^*, \ldots, X_I^*] \\
& = E[X_1, \ldots, X_I] [E[\text{ED}[\hat{m}^p : m_f]] \\
& + \sum_{i=1}^{I} (X_i + X_i^*) / (n + n^*) \text{ED}_{(i)}(X_i | X_1^*, \ldots, X_I^*)] \\
& = \text{ED}_f + \sum_{i=1}^{I} E[X_1, \ldots, X_I] [(X_i + X_i^*) / (n + n^*) \text{ED}_{(i)}X_i] \\
& = \text{ED}_{(i)}(X_i) = \frac{s_i}{2X_i} + \frac{1}{24X_i^2} A_{(i)} + o(n^{-2}) |_{n=X_i}
\end{align*}
\]
we have
\[
EX_i \ldots X_j \left[ EX_i \ldots X_j \left( \frac{X_i + X_j}{n + n^*} ED_{(j)}(X_i) | X_j^*, \ldots, X_i^* \right) \right] = \frac{s_i}{2(n + n^*)} E[X_j^*] E[X_i^{-1}] + \frac{s_i}{2(n + n^*)} + \frac{1}{24(n + n^*)} A_{(i)}^{-1} E[X_j^*] E[X_i^{-2}] + \frac{1}{24(n + n^*)} A_{(i)}^{-1} E[X_i^{-1}] + o\left(n^{-2}\right)
\]

Because of (20) and (21), we have
\[
EX_i \ldots X_j \left[ EX_i \ldots X_j \left( \frac{X_i + X_j}{n + n^*} ED_{(j)}(X_i) | X_j^*, \ldots, X_i^* \right) \right] = \frac{s_i m_i}{2(n + n^*)} E[X_i^{-1}] + \frac{s_i}{2(n + n^*)} + \frac{m_i n^*}{24(n + n^*)} A_{(i)}^{-1} E[X_i^{-2}] + \frac{1}{24(n + n^*)} A_{(i)}^{-1} E[X_i^{-2}] + o\left(n^{-2}\right)
\]

(We omit the proof that \(E[O_p(n^{-2})] = o(n^{-2})\) holds for the present case.) Consequently,
\[
ED_{\hat{m}^p \mid m} = ED_{\hat{m}} + \sum_{i=1}^{I} \left( \frac{s_i}{2n} + \frac{1}{24n^2} m_i^{-1} \left( 12s_i(1 - m_i)n^*/(n + n^*) + A_{(i)}^{-1} \right) + o\left(n^{-2}\right) \right) \tag{27}
\]

If we apply (3) to \(ED_{\hat{m}}^{-1}\), we obtain
\[
ED_{\hat{m}^p}^{-1} = \frac{I - 1}{2(n + n^*)} + \frac{1}{12(n + n^*)^2} \left( M_f - 1 \right) + o\left((n + n^*)^{-2}\right)
\]

If we insert this result into (27), we obtain (25).

By inserting \(s_i = I_i - 1\) and
\[
A_{(i)}^{-1} = 2m_i \sum_{j=1}^{I_i} m_j^{-1} - 2
\]
into (25), we obtain (26). \(\square\)

Let us compare the risk of \(\hat{m}^p\) with that of \(\hat{m}\). From (13) and (25), we notice that
\[
ED_{\hat{m} \mid m}^{-1} - ED_{\hat{m}^p \mid m}^{-1} = \frac{I - 1}{2} \left( n^{-1} - (n + n^*)^{-1} \right) + \frac{M_f - 1}{12} \left( n^2 - (n + n^*)^{-2} \right) \frac{n^*}{2n^2(n + n^*)} \sum_{i=1}^{I} m_i^{-1} \left( 1 - m_i \right) s_i + o\left(n^{-2}\right) + o\left((n + n^*)^{-2}\right)
\]
\[
= \frac{I - 1}{2n} \frac{1}{1 + \beta} + \frac{1}{n^2} \frac{1}{1 + \beta} \left( M_f - 1 \right) \frac{1}{12} \beta \sum_{i=1}^{I} m_i^{-1} \left( 1 - m_i \right) s_i + o\left(n^{-2}\right) + o\left((n + n^*)^{-2}\right)
\]
From the first-order term of (28), it is easily confirmed that \( \hat{m}^p \) is better than \( \hat{m} \) with large enough size of \( n \) and \( n^* \).

Taking the second-order terms into account, (28) could be negative. This could happen when large \( s_i \) exists, that is, some of the second-stage models contain a large number of parameters. For example, let \( m_1 = \cdots = m_I = 1/I \) and suppose that \( n^* \) is relatively so large to \( n \) that \( \beta/(1 + \beta) \div 1 \) and \( (\beta^2 + \beta)/(1 + \beta)^2 \div 1 \), then (28) approximately equals

\[
(I - 1) \left( \frac{1}{2n} + \frac{I + 1}{12n^2} - \frac{1}{2n^2} \sum_{i=1}^{I} s_i \right)
\]

up to the second-order terms. This is negative when \( n < \sum_{i=1}^{I} s_i - (I + 1)/6 \).

Because we are neglecting the higher-order terms, we are unsure if the situation \( ED^{\hat{m}}[\hat{m} : m] < ED^{\hat{m}^p}[\hat{m}^p : m] \) could happen with some \( n \). (We confirm this later by simulation in Section 2.2.) Anyway, we notice that superiority of the pooled estimator fades when the second-stage models are highly complicated.

As in the estimator \( \hat{m}^* \), the use of another set of sample for the first-stage model estimation has an adverse effect on the risk. To see this phenomenon, consider the case that the present sample size equals to \( n + n^*(\hat{A} \hat{n}) \). In this case, (13) with \( n \) substituted with \( \tilde{n} \) turns out

\[
\frac{1}{2\tilde{n}} \left( I - 1 + \sum_{i=1}^{I} s_i \right) + \frac{1}{12\tilde{n}^2} (M_f - 1) + \frac{1}{24\tilde{n}^2} \left( \sum_{i=1}^{I} m_i^{-1} A_{(i)} \right) + o(\tilde{n}^{-2})
\]

(29)

and if \( \hat{A}_{(i)} > 0, i = 1, \ldots, I \), then neglecting the higher-order terms, we have

\[
ED^{\hat{m}}[\hat{m} : m](\tilde{n}) - ED^{\hat{m}^p}[\hat{m}^p : m](n, n^*)
\]

\[
= - \frac{1}{2n} \frac{\beta}{1 + \beta} \sum_{i=1}^{I} s_i - \frac{1}{24n^2} \frac{\beta^2 + 2\beta}{(1 + \beta)^2} \sum_{i=1}^{I} m_i^{-1} \hat{A}_{(i)}
\]

(30)

\[
- \frac{1}{2n^2} \frac{\beta}{1 + \beta} \sum_{i=1}^{I} m_i^{-1}(1 - m_i)s_i < 0
\]

Under the condition of the same total sample size \( \tilde{n} \), unified approach \( \hat{m} \) is superior to non-unified approach \( \hat{m}^p \).

As for the comparison between \( \hat{m}^* \) and \( \hat{m}^p \), it is obvious from (25) and (18) that the latter is always asymptotically (with respect to both of the first and the second-order terms) better than the former.

2.2. Examples

We consider three examples to check the theoretical results in 2.1 for the comparison of three estimators \( \hat{m}, \hat{m}^*, \) and \( \hat{m}^p \). Each of three examples is a multinomial distribution.
over a contingency table, where there exists a prior survey on the marginal distribution over the column categories.

First, we use an artificial example to illustrate some points mentioned in the previous subsection. We focus on the situation when the sample size $n$ is relatively small ($n < 10p$). It is rather impractical setting, but wherein the second-order term works in a non-negligible way and interesting (rather pathological) behavior of the three estimators can be watched.

The next two examples are the distributions derived from real datasets. We will observe the behavior of the three estimators under practical conditions, where $n$ and $n^*$ are of relatively large size ($n > 10p$). We can confirm the relative performance of the three estimators that is expected from the theoretical result on their asymptotic risks.

We introduce the common notations among the examples. The abbreviations “pre.*”, “pri.*,” “pool.*” “correspond to $\hat{m}, \hat{m}^*, \hat{m}^p$, respectively. The notation”* sim” shows the risk (or the “required sample size” defined later) as obtained by simulation. We generated samples for the present and prior surveys with the sizes of given $n$ and $n^*$, respectively, and calculated the loss of the estimators. Repeating this $10^4$ (or $10^6$) times, we obtained the simulated risk by averaging the losses. We also calculated the approximated risk from (13, 18), and (25) by neglecting the terms $o(n^{-2})$ and $o((n^*)^{-2})$, which are denoted by “* app”.

The “required sample size (r.s.s.)” is defined as follows, according to estimators to be compared. Let the risks of the three estimators as the functions of $n$ and $n^*$ be denoted by $ED[\hat{m} : m](n), ED[\hat{m}^* : m](n, n^*), ED[\hat{m}^p : m](n, n^*)$.

1. Comparison between $\hat{m}$ and $\hat{m}^*$ through r.s.s. The “r.s.s for $\hat{m}^*$ to $\hat{m}$ under the condition $n = n_0$” is the solution of $n^*$ for the equation

$$ED[\hat{m}^* : m](n_0, n^*) = ED[\hat{m} : m](n_0)$$

(31)

As we mentioned in the previous subsection, if $n^* = n_0$, then (at least asymptotically) the left side is larger than the right side in (31). Therefore, $n^*$ must be larger than $n_0$ so that the equation holds. The r.s.s. shows the degree of inefficiency of $\hat{m}^*$ (non-unified approach) to $\hat{m}$ (unified approach).

2. Comparison between $\hat{m}$ and $\hat{m}^p$ through r.s.s. The “r.s.s for $\hat{m}$ to $\hat{m}^p$ under the condition $(n, n^*) = (n_0, n_0^*)$” is the solution of $n$ for the equation

$$ED[\hat{m} : m](n) = ED[\hat{m}^p : m](n_0, n_0^*)$$

(32)

In the previous subsection, we noticed that (at least asymptotically) if $n = n_0 + n_0^*$, then the left-hand side is smaller than the right-hand side. Therefore, the solution $n$ for the equation is smaller than $n_0 + n_0^*$. The r.s.s. indicates the degree of efficiency of $\hat{m}$ to $\hat{m}^p$ in view of the total sample size.

Note that the risks that appear in (31) and (32) are calculated from either simulation or approximation.

**Example 1. 100 by 2 Contingency Table**
Consider a 100 by 2 contingency table and suppose that the distribution over the table is uniform, that is, each category (cell) has the probability 1/200. Suppose that we have a prior survey on the two column categories, hence \( I = 2 \) and \( J_1 = J_2 = 100 \). The second-stage model is a full model over the row categories with a column fixed.

Table 1 shows the risks of the three estimators calculated from both simulation and approximation, where * . sim is the average of \( 10^6 \) simulated losses (its standard deviation is at most 6.13E - 6).

The r.s.s. of \( \hat{m}^* \) to \( \hat{m} \) under the condition \( n = n_0 \) is given in Table 2 with several values of \( n_0 \). Table 3 shows the r.s.s. of \( \hat{m} \) to \( \hat{m}^p \) under the condition \( (n, n^*) = (n_0, n_0^*) \) with several values \( (n_0, n_0^*) \) satisfying \( n_0 = n_0^* \).

We can observe the following points from this result.

- We confirm that \( \hat{m} \) outperforms \( \hat{m}^* \) under same sample size \( n = n^* \). (See the rows for \( (n, n^*) = (200, 200), \ldots, (1000, 1000) \) of Table 1.) When \( n \) is small as \( n = 100, 150, 200 \), even \( n^* \) as large as 100 is not enough to compensate the inferiority of non-unified approach. (See the first three rows of Table 1.) The r.s.s. of Table 2 shows more clearly the inferiority of \( \hat{m}^* \) to \( \hat{m} \). Here, \( \hat{m}^* \) needs approximately 110% – 250% larger sample size compared with \( \hat{m} \) to attain the same risk.

- It is confirmed that \( \hat{m}^p \) is outperformed by \( \hat{m} \) under some situations like \( (n, n^*) = (100, 105^2), (150, 10^5), (200, 10^5), (200, 200) \) in Table 1. The relation \( \frac{\hat{m}(n + n^*)}{n} < \frac{\hat{m}^p}{n^*} \) is also confirmed from the fact \( \frac{1}{\hat{m}} \hat{m}(m : m)(n + n^*) < \frac{1}{\hat{m}^p} \hat{m}(m : m)(200, 200) \) and \( \frac{1}{\hat{m}} \hat{m}(m : m)(800) = 0.132419 < 0.283377 = \frac{1}{\hat{m}^p} \hat{m}(m : m)(400, 400) \). The r.s.s. in Table 3 shows that \( \hat{m} \) needs only 1 to 5 more observations than \( n_0^* \) to be par with \( \hat{m}^p \) with the sample size of \( (n_0^*, n_0^*) \).

- The last 10 rows of Table 1 shows an interesting aspect of the risks of \( \hat{m}^* \) and \( \hat{m}^p \). As \( n \) is fixed to be 90 and \( n^* \) increases from 100 to 1000, the risk of \( \hat{m}^* \) increases from 0.108104 to 1.517068.
decreases, while the risk of $\hat{m}^p$ increases. Increasing the sample size $n^*$ naturally reduces the risk for $\hat{m}^*$. On the contrary, in the case of $\hat{m}^p$, the adverse effect of blending two sets of samples outweigh the positive effect of increasing the sample size.

- The superiority of $\hat{m}^p$ to $\hat{m}^*$ is confirmed from the simulated risk.
- The discrepancy between the simulated risk ($\hat{m}^*$) and the approximated risk ($\hat{m}^p$) could be large when $n$ is small. However, as for the order of the three estimators' risks, the orders in the simulation and approximation coincide in most cases in Table 1.

**Example 2. Breast cancer classification**

The second example is the breast cancer data from Zwitter and Soklic (1988). We made a cross-tabulation w.r.t. the variables "age group" (5 groups: 30–39, …, 70–79) and "the degree of malignancy" (3 levels: 1, 2, 3) by excluding the only person that is in his/her twenties from the original dataset. By dividing each cell by 285 (the total number of observations), we obtained the relative frequency (see Table 4). We suppose that they are the true probability $m_{ij}$ for the categories $C_{ij}$ for $i = 1, ..., 5, j = 1, ..., 3$ (the first and second index represent the age group and the malignancy, respectively).

We consider a two-stage model, where the first-stage model is the full model over the age groups, and each of the second-stage models is a full model over malignancy levels within an age group. Hence, $I = 5$ and $J_i = 3, i = 1, ..., 5$. The marginal probability (the column sum in the table) of each age group corresponds to $m_i, i = 1, ..., 5$, the first-stage parameter. Suppose that we have a prior survey on the ratio of age groups of the sample size $n^*$. 

For several values of $(n, n^*)$, we calculated the approximated risks and the simulated risks. The simulated risk is the average of $10^4$ losses calculated from each generated sample. The result is given in Table 5, where the simulated risk is in the parenthesis. Figure 1 shows how the simulated risks change as $n$ varies. We calculated two kinds of r.s.s.'s. One is the r.s.s. of $\hat{m}^*$ to $\hat{m}$ under the condition $n = n_0$. Table 6 shows the r.s.s. for several values of $n_0$. The other r.s.s. is that of $\hat{m}$ to $\hat{m}^p$ under the condition $(n, n^*) = (n_0, n_0^*)$. In Table 7, we show the r.s.s. for several cases where $n_0 = n_0^*$.

We make the following observations:

- In the lower left part (including the diagonal part) of Table 5, $\hat{m}$ is superior to $\hat{m}^*$, as is indicated in Section 2.1. When $(n, n^*) = (200, 600), (200, 1000), (600, 1000)$, the result is opposite. From Table 6, we notice that when $n = 200$, the prior survey needs 28% more samples ($n^* = 255$) than the present survey, while just 3% more samples ($n^* = 1029$) suffice when $n = 1000$.
- In all cases of Table 5, $\hat{m}^p$ is superior to $\hat{m}$. It seems that each $(n, n^*)$ in Table 5 is large enough to guarantee the superiority of $\hat{m}^p$ to $\hat{m}$. In view of the total sample size, the performance of $\hat{m}^p$ is inferior to $\hat{m}$ as is shown in Section 2.1. Table 7 shows that $\hat{m}$ with the sample size of $1.12n_0 - 1.16n_0$ attain the same risk as $\hat{m}^p$ with the total sample size $2n_0$.

### Table 2. Example 1. R.s.s. of $\hat{m}^*$ to $\hat{m}$.

| $n_0$ | 400  | 600  | 800  | 1000 | 1200 | 1400 | 1600 | 1800 | 2000 |
|-------|------|------|------|------|------|------|------|------|------|
| r.s.s.sim | 995  | 1014 | 1142 | 1296 | 1467 | 1658 | 1854 | 2034 | 2229 |
| r.s.s.app | 791  | 895  | 1063 | 1247 | 1437 | 1631 | 1826 | 2023 | 2220 |
As the theoretical result indicates, $b^p$ is always superior to $b^m$:

There is no large discrepancy between the simulation and the approximation for the risk and the r.s.s. under the given sample sizes.

Example 3. Household classification

As the third example, we use data from “2014 national survey of family income and expenditure” by Statistics Bureau of Japan (Statistics Bureau of Japan [2016]). Table 8 is the relative frequency obtained from the cross tabulation of 100006 households by “Yearly income group $(Y_1, \ldots, Y_{10})$” and “Household age group $(H_1, \ldots, H_6)$“. We use this relative frequency as the population parameter $m_{ij}$, $1 \leq i \leq 10, 1 \leq j \leq 6$.

The two-staged model is considered where the first-stage model is the full model over age groups $H_1, \ldots, H_6$, and each second-stage model is a full model over the income groups $Y_1, \ldots, Y_{10}$ within an age group. We suppose that there is a prior survey on the first-stage model. For the risk of the three estimators under several values of $(n, n^*)$, see Table 9. Figure 2 shows how the simulated risks change as $n$ varies. The r.s.s. of $\hat{m}^*$ to $\hat{m}$ under the condition $n = n_0$ is shown in Table 10. The r.s.s. of $\hat{m}$ to $\hat{m}^p$ under the condition $(n, n^*) = (n_0, n_0^*)$ is presented in Table 11. Because the results are similar to Example 2, we omit the comments.

3. Conclusion

The golden rule “the larger sample, the less risk” holds true under sufficient sample sizes. Actually, we observed the following points theoretically, and numerically for three examples:

- As the theoretical result indicates, $\hat{m}^p$ is always superior to $\hat{m}^*$.  

- There is no large discrepancy between the simulation and the approximation for the risk and the r.s.s. under the given sample sizes.

Example 3. Household classification

As the third example, we use data from “2014 national survey of family income and expenditure” by Statistics Bureau of Japan (Statistics Bureau of Japan [2016]). Table 8 is the relative frequency obtained from the cross tabulation of 100006 households by “Yearly income group $(Y_1, \ldots, Y_{10})$” and “Household age group $(H_1, \ldots, H_6)$“. We use this relative frequency as the population parameter $m_{ij}$, $1 \leq i \leq 10, 1 \leq j \leq 6$.

The two-staged model is considered where the first-stage model is the full model over age groups $H_1, \ldots, H_6$, and each second-stage model is a full model over the income groups $Y_1, \ldots, Y_{10}$ within an age group. We suppose that there is a prior survey on the first-stage model. For the risk of the three estimators under several values of $(n, n^*)$, see Table 9. Figure 2 shows how the simulated risks change as $n$ varies. The r.s.s. of $\hat{m}^*$ to $\hat{m}$ under the condition $n = n_0$ is shown in Table 10. The r.s.s. of $\hat{m}$ to $\hat{m}^p$ under the condition $(n, n^*) = (n_0, n_0^*)$ is presented in Table 11. Because the results are similar to Example 2, we omit the comments.

3. Conclusion

The golden rule “the larger sample, the less risk” holds true under sufficient sample sizes. Actually, we observed the following points theoretically, and numerically for three examples:

- As the theoretical result indicates, $\hat{m}^p$ is always superior to $\hat{m}^*$.  

- There is no large discrepancy between the simulation and the approximation for the risk and the r.s.s. under the given sample sizes.

Example 3. Household classification

As the third example, we use data from “2014 national survey of family income and expenditure” by Statistics Bureau of Japan (Statistics Bureau of Japan [2016]). Table 8 is the relative frequency obtained from the cross tabulation of 100006 households by “Yearly income group $(Y_1, \ldots, Y_{10})$” and “Household age group $(H_1, \ldots, H_6)$“. We use this relative frequency as the population parameter $m_{ij}$, $1 \leq i \leq 10, 1 \leq j \leq 6$.

The two-staged model is considered where the first-stage model is the full model over age groups $H_1, \ldots, H_6$, and each second-stage model is a full model over the income groups $Y_1, \ldots, Y_{10}$ within an age group. We suppose that there is a prior survey on the first-stage model. For the risk of the three estimators under several values of $(n, n^*)$, see Table 9. Figure 2 shows how the simulated risks change as $n$ varies. The r.s.s. of $\hat{m}^*$ to $\hat{m}$ under the condition $n = n_0$ is shown in Table 10. The r.s.s. of $\hat{m}$ to $\hat{m}^p$ under the condition $(n, n^*) = (n_0, n_0^*)$ is presented in Table 11. Because the results are similar to Example 2, we omit the comments.

3. Conclusion

The golden rule “the larger sample, the less risk” holds true under sufficient sample sizes. Actually, we observed the following points theoretically, and numerically for three examples:
When $n$ and $n^*$ are large enough, the risk of $\hat{m}_p$ (which uses the sample of size $n + n^*$ for the estimation of the first-stage model) is lower than that of $\hat{m}$ (which uses the sample of size $n$ for the estimation of the first-stage model) or $\hat{m}^*$ (which uses the sample of size $n^*$ for the estimation of the first-stage model). When $n$ and $n^*$ are large enough and additionally $n^* > n$, then the risk of $\hat{m}^*$ is lower than that of $\hat{m}$.

**Figure 1.** Example 2. Simulated risk.

**Table 6.** Example 2. R.s.s. of $\hat{m}^*$ to $\hat{m}$.

| $n_0$ | 200   | 400   | 600   | 800   | 1000  |
|-------|-------|-------|-------|-------|-------|
| r.s.s.app(r.s.s.sim) | 236(255) | 433(445) | 633(652) | 832(850) | 1032(1029) |
However, blending samples from two surveys has a negative effect in risk reduction. We notice this from the following theoretical or numerical observations:

- When \( n = n^* \), the risk of \( \hat{m} \) is lower than that of \( \hat{m}^* \). The latter is using two independent sample sets separately for the first and second-stage models, while the former uses the common sample for the two stages. Unified approach \( \hat{m} \) is preferable, because the estimation at the first stage gives an appropriate weight to each result of the second-stage estimations.

- If you compare \( \hat{m} \) using the present sample of size \( n + n^* \) to \( \hat{m}^P \) (which uses the sample of a total size \( n + n^* \)), the former performs better with respect to the risk. In view of the r.s.s. of \( \hat{m} \) to \( \hat{m}^P \) under the condition \( (n, n^*) = (n_0, n_0) \), it is far less than \( 2n_0 \).

- If \( n \) is small, we observe a rather pathological phenomenon that increasing \( n^* \) causes the higher risk of \( \hat{m}^P \).

The boundary between the area of \( (n, n^*) \) where the golden rule holds, and the pathological area, can be estimated using the approximated risks. If the approximated risk of \( \hat{m} \) is smaller than that of \( \hat{m}^P \), that is, (28) is minus, then it indicates \( n \) or \( n^* \) is too small and we might be in the pathological area. Consequently, we propose the following procedure.

### Table 7. Example 2. R.s.s. of \( \hat{m} \) to \( \hat{m}^P \).

| \( n_0 = n_0^* \) | 200 | 400 | 600 | 800 | 1000 |
|-----------------|-----|-----|-----|-----|------|
| r.s.s.app(r.s.s.sim) | 226(223) | 459(456) | 692(686) | 925(921) | 1158(1159) |

### Table 8. Classification of households by income and age.

|   | H1  | H2  | H3  | H4  | H5  | H6  |
|---|-----|-----|-----|-----|-----|-----|
| Y1 | 0.00161 | 0.00232 | 0.00512 | 0.00395 | 0.00468 | 0.00066 |
| Y2 | 0.00331 | 0.0081 | 0.00953 | 0.00783 | 0.01145 | 0.00278 |
| Y3 | 0.00974 | 0.02109 | 0.02046 | 0.01499 | 0.02536 | 0.00494 |
| Y4 | 0.00799 | 0.03519 | 0.03229 | 0.02017 | 0.0338 | 0.00708 |
| Y5 | 0.00547 | 0.0376 | 0.04362 | 0.02442 | 0.03723 | 0.00352 |
| Y6 | 0.00494 | 0.05082 | 0.09003 | 0.05772 | 0.03732 | 0.00452 |
| Y7 | 0.00126 | 0.02106 | 0.0543 | 0.05531 | 0.01999 | 0.00234 |
| Y8 | 0.00071 | 0.00961 | 0.0323 | 0.04043 | 0.0108 | 0.00122 |
| Y9 | 0.00011 | 0.00201 | 0.01204 | 0.02184 | 0.00466 | 0.00052 |
| Y10 | 0.00006 | 0.00139 | 0.00697 | 0.01582 | 0.00344 | 0.00022 |
| Col. Sum | 0.03520 | 0.18919 | 0.30666 | 0.26248 | 0.17825 | 0.02826 |

### Table 9. Example 3. Risk of estimators.

|   | pre.app | pre.sim | pri.app | pri.sim | pool.app | pool.sim | n = 1000 | n = 2000 | n = 3000 |
|---|---------|---------|---------|---------|----------|----------|----------|----------|----------|
| \( n^* = 1000 \) | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.033 | 0.033 | 0.033 |
| \( n = 1000 \) | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.033 | 0.033 | 0.033 |
| \( n = 2000 \) | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.033 | 0.033 | 0.033 |
| \( n = 3000 \) | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.0332 | 0.033 | 0.033 | 0.033 |

However, blending samples from two surveys has a negative effect in risk reduction. We notice this from the following theoretical or numerical observations:

- When \( n = n^* \), the risk of \( \hat{m} \) is lower than that of \( \hat{m}^* \). The latter is using two independent sample sets separately for the first and second-stage models, while the former uses the common sample for the two stages. Unified approach \( \hat{m} \) is preferable, because the estimation at the first stage gives an appropriate weight to each result of the second-stage estimations.

- If you compare \( \hat{m} \) using the present sample of size \( n + n^* \) to \( \hat{m}^P \) (which uses the sample of a total size \( n + n^* \)), the former performs better with respect to the risk. In view of the r.s.s. of \( \hat{m} \) to \( \hat{m}^P \) under the condition \( (n, n^*) = (n_0, n_0) \), it is far less than \( 2n_0 \).

- If \( n \) is small, we observe a rather pathological phenomenon that increasing \( n^* \) causes the higher risk of \( \hat{m}^P \).
1. If the present survey is already finished, substitute $m_i$ in (28) with the estimate from $\hat{m}^p$ and calculate the value. If the estimated value is plus (minus), use the estimator $\hat{m}^p$ ($\hat{m}$).

Table 10. Example 3. R.s.s. of $\hat{m}^*$ to $\hat{m}$.

| $n_0$ | 1000 | 1500 | 2000 | 2500 | 3000 |
|-------|------|------|------|------|------|
| r.s.s.app(r.s.s.sim) | 1157(1183) | 1650(1668) | 2146(2145) | 2644(2696) | 3143(3155) |

Table 11. Example 3. R.s.s. of $\hat{m}$ to $\hat{m}^p$.

| $n_0 = n_0^*$ | 1000 | 1500 | 2000 | 2500 | 3000 |
|---------------|------|------|------|------|------|
| r.s.s.app(r.s.s.sim) | 1031(1035) | 1552(1562) | 2074(2079) | 2595(2604) | 3117(3129) |
2. If the present survey is still in the planning stage, then substitute \( \hat{m}_i \) in (28) with the estimate from \( \hat{m}_i \). With \( n \) given, if the estimated value of (28) is minus, it indicates that the \( n \) is not large enough for reliable estimation. We should try to increase \( n \) (or alternatively decrease the number of the categories within each second-stage model) until the value gets positive.

**Acknowledgment**

We appreciate valuable comments of the referees, which enhanced the quality and readability of the paper.

**References**

Amari, S. 2016. *Information geometry and its applications*. Japan: Springer.

Cochran, W. G. 1954. The combination of estimates from different experiments. *Biometrics* 10 (1):101–29. doi:10.2307/3001666.

Cochran, W. G., and G. M. Cox. 1957. *Experimental design*. 2nd ed. New York: Wiley.

Good, I. J. 1950. *Probability and the weighting of evidence*. Scotland: Griffin.

Ridder, G., and R. Moffitt. 2007. The econometrics of data combination. In *Handbook of Econometrics*, ed. J. Heckman and E. Leamer, vol. 6B, 5469–547. Holland: Elsevier.

Sheena, Y. 2018. Asymptotic expansion of the risk of maximum likelihood estimator with respect to \( \alpha \)-divergence. *Communications in Statistics—Theory and Methods* 47 (16):4059–87. doi:10.1080/03610926.2017.1380828.

Valliant, R., J. A. Dever, and F. Kreuter. 2013. *Practical tools for designing and weighting survey samples*. New York: Springer.

Zwitter, M., and M. Soklic. 1988. Breast cancer data set. UCI machine learning repository. Accessed September 9, 2018. [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer).

Statistics Bureau of Japan. 2016. 2014 national survey of family income and expenditure. Accessed September 11, 2018. [https://www.stat.go.jp/english/data/zensho/index.html](https://www.stat.go.jp/english/data/zensho/index.html).

**Appendix**

**Lemma 1.** Let \( X = (X_1, ..., X_{p+1}) \) be the random vector whose distribution is defined as the multinomial distribution with (1) and the sample size \( n \). The distribution under the condition \( X_i \neq 0, 1 \leq i \leq p+1 \) is considered. Let the unconditional and conditional expectations of a random variable \( Y(X) \) be denoted by \( E[Y(X)] \) and \( E^c[Y(X)] \), respectively. If

\[
|Y(X)| \leq a + bn^c
\]

holds with some nonnegative numbers \( a, b, c \), the difference between the two expectations decreases to zero with exponential speed as \( n \) goes to infinity, namely for any \( s > 0 \),

\[
E^c[Y(X)] = E[Y(X)] + o(n^{-s})
\]

(33)

In the special case, MLE of \( m \), \( \hat{m}(X) \), and the moments of \( X_i, i = 1, ..., p+1 \), the following equations hold for any \( s > 0 \).

\[
E^c\left[\hat{D} \hat{m}(X) : m\right] = E\left[\hat{D} \hat{m}(X) : m\right] + o(n^{-s})
\]

\[
E^c\left[(X_i - n\hat{m}_i)^k\right] = E\left[(X_i - n\hat{m}_i)^k\right] + o(n^{-s}), \quad k = 1, 2, ....
\]
Proof. For \( i(= 1, ..., p + 1) \) and \( s(>0) \), because the following equivalence relation holds

\[ n!(1 - m_i)^n \to 0 \]

\[ \iff s \log n + n \log (1 - m_i) \to -\infty \iff n \left( \frac{s \log n}{n} + \log (1 - m_i) \right) \to -\infty \]

we have

\[ P(X_i = 0) = (1 - m_i)^n = o(n^{-s}) \tag{34} \]

Let \( \mathbb{Z}^{p+1}_+ \) be the set of all \( p + 1 \)-dimensional vectors whose elements are nonnegative integers.

\[ \mathcal{X} \triangleq \left\{ x = (x_1, ..., x_{p+1}) \in \mathbb{Z}^{p+1}_+ \mid \sum_{j=1}^{p+1} x_j = n, x_j > 0, 1 \leq i \leq p + 1 \right\} \]

\[ \mathcal{X}^* \triangleq \left\{ x = (x_1, ..., x_{p+1}) \in \mathbb{Z}^{p+1}_+ \mid \sum_{j=1}^{p+1} x_j = n, x_i = 0, 1 \leq \exists i \leq p + 1 \right\} \]

Notice that

\[ m^* \triangleq P(X_i = 0, 1 \leq i \leq p + 1) \leq \sum_{i=1}^{p+1} P(X_i = 0) \]

which means \( m^* = o(n^{-s}) \). Because

\[ E^*[Y(X)] = \sum_{x \in \mathcal{X}} Y(x)P(X = x)/(1 - m^*) \]

\[ E[Y(X)] = \sum_{x \in \mathcal{X}} Y(x)P(X = x) + \sum_{x \in \mathcal{X}^*} Y(x)P(X = x) \]

we have

\[ E^*[Y(X)] - E[Y(X)] = \sum_{x \in \mathcal{X}} Y(x)P(X = x) \frac{m^*}{1 - m^*} - \sum_{x \in \mathcal{X}^*} Y(x)P(X = x) \]

Choose arbitrary \( s(>0) \). Because \( \mathcal{X} \) and \( \mathcal{X}^* \) are both finite sets, for some nonnegative constants \( a', a'', b', b'' \),

\[ n^s \left| \sum_{x \in \mathcal{X}^*} Y(x)P(X = x) \frac{m^*}{1 - m^*} \right| \leq n^s \sum_{x \in \mathcal{X}^*} (a + bn^2) \frac{m^*}{1 - m^*} \]

\[ \leq a' n^s m^* \frac{1}{1 - m^*} + b' n^s m^* \frac{m^{s+c^*}m^*}{1 - m^*} \]

\[ n^s \left| \sum_{x \in \mathcal{X}^*} Y(x)P(X = x) \right| \leq n^s \sum_{x \in \mathcal{X}^*} (a + bn^2) m^* \leq a'' n^s m^* + b'' (n^{s+c^*}m^*). \]

Because of \( m^* = o(n^{-s}) \) for any \( t(>0) \),

\[ \frac{n^s m^*}{1 - m^*}, \frac{n^s m^{s+c^*}m^*}{1 - m^*}, n^s m^*, n^s m^{s+c^*}m^* \to 0 \]

as \( n \to \infty \), which shows (33). The rest is obvious from the fact \( \bar{D}[\hat{m}(X) : m] \) is bounded (notice that \( x \log x \to 0 \) as \( x \to 0 \)) and

\[ |(X_i - nm_i)|^k \leq \max(m_i^k, (1 - m_i)^k)^n, \quad k = 1, 2, \ldots \]
Lemma 2. Suppose that $X_n, n = 1, 2, \ldots$ converges in probability to some positive constant as $n \to \infty$. If $F(x) = O(1)$ as $x \to \infty$, then $F(nX_n) = O_p(1)$. If $F(x) = o(1)$ as $x \to \infty$, then $F(nX_n) = o_p(1)$.

Proof. The first statement is proved as follows. If $F(x) = O(1)$, there exists some $M > 0$ and $x_0$ such that
\[ |F(x)| < M, \quad \forall x > x_0 \]  
(35)

Let $b > 0$ denote the point to which $X_n$ converges in probability. Choose an arbitrary $\epsilon(> 0)$ and $\tau(b > \tau > 0)$, then because of the convergence of $X_n$, we have $n_0$ such that
\[ P(b - \tau < X_n < b + \tau) > 1 - \epsilon, \quad \forall n > n_0 \]
which means
\[ P(n(b - \tau) < nX_n) > 1 - \epsilon, \quad \forall n > n_0 \]
Let $N = \max(x_0/(b - \tau), n_0)$. Note that if $n > N$, then $n(b - \tau) > x_0$ and $n > n_0$. Therefore,
\[ P(x_0 < nX_n) > 1 - \epsilon, \quad \forall n > N \]  
(36)

Combining (35) and (36), we have
\[ P(|F(nX_n)| < M) > 1 - \epsilon, \quad \forall n > N \]
Now we prove the second statement. Choose an arbitrary $\epsilon(> 0)$. If $F(x) = o(1)$, there exists $x_0$ such that
\[ |F(x)| < \epsilon, \quad \forall x > x_0 \]  
(37)
Combining (36) and (37), we have
\[ P(|F(nX_n)| < \epsilon) > 1 - \epsilon, \quad \forall n > N \]