Title

Middle censoring in the multinomial distribution with applications

Authors

Jammalamadaka, S Rao

Bapat, Sudeep R

Publication Date

2020-12-01

DOI

10.1016/j.spl.2020.108916

Peer reviewed

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Statistics and Probability Letters 167 (2020) 108916

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Middle censoring in the multinomial distribution with applications

S. Rao Jammalamadaka [a], Sudeep R. Bapat [b,*]

[a] Department of Statistics and Applied Probability, University of California, Santa Barbara, USA
[b] Department of Operations Management & Quantitative Techniques, Indian Institute of Management, Indore, INDIA

ARTICLE INFO

Article history:
Received 13 May 2019
Received in revised form 27 April 2020
Accepted 16 August 2020
Available online 27 August 2020

Keywords:
Middle-censoring
Multinomial
Dirichlet
Bayes Estimation

ABSTRACT

In a multinomial set-up with k possible outcomes, we develop estimation under a "middle censoring" paradigm, which is as defined in Jammalamadaka and Mangalam (2003). This problem has many special features because of the inter-dependent probabilities, which we explore here.

1. Introduction

In this paper we discuss a "middle-censoring" scheme when the data comes from a multinomial experiment. Middle censoring occurs if the actual value of a data point is not observed but is known to fall inside a specific interval. In particular for our multinomial setup, some individuals choose exactly one of the k possible categories whereas some others, choose intervals covering several categories. Well known censoring schemes such as right- and left-censoring can be seen as special cases of such a middle censoring by picking suitable censoring intervals.

Considerable ground has been covered with regard to middle censoring problems over the last decade and a half. One may refer to Jammalamadaka and Mangalam (2003) where the authors develop self-consistent and non-parametric maximum likelihood estimators (MLEs) for the unknown Cumulative Distribution Function (CDF) for such middle censored data. Jammalamadaka and Iyer (2004) establish approximate self consistency for middle censored data. Iyer et al. (2008) considered a parametric middle censoring scheme using exponential lifetime data. Davarzani and Parsian (2011) discussed middle censoring in a discrete setup by taking observations from a geometric distribution. More recent references include Jammalamadaka and Leong (2015) where the authors discuss a middle censoring scheme for geometric random variables in the presence of covariates, and Ahmadi et al. (2017) who consider middle censoring in the context of competing risks.

An outline of the paper is as follows. In Section 2 we develop the likelihood function for middle censored data from a multinomial model, in the most general setup. However, because of the complicated dependencies between the multinomial probabilities as well as the observed frequencies, explicit expressions for the MLEs for individual probabilities and their large-sample variances in such a general setup are not easy to get, and may have to be obtained numerically. To illustrate these ideas, we consider three different scenarios covering the middle censoring scheme—one where there is just one interval allowed, a second one where there are 2 non-overlapping intervals, and the third case that allows

* Corresponding author.
E-mail address: sudeepb@iimidr.ac.in (S.R. Bapat).

https://doi.org/10.1016/j.spl.2020.108916
0167-7152/

2 intervals that overlap. Section 3 develops a Bayesian framework for estimating the required probabilities. The final Section 4 contains bootstrap estimates and variances of the unknown probability vector. This section also provides a simulation analysis comparing the Bayes estimates, and the estimates one gets from the different methods proposed here. We also present an example using real data, in the form of ratings given by a group of students for their experience in using a particular software for remote lectures.

1.1. The problem

Consumers are constantly asked to rate products that they buy on a website like Amazon. Or in market research, a company which plans to launch a new product, wants to gauge the user response in terms of the preference-ratings or the "star-ratings" the product gets, as part of a pilot study. Assume that the company contacts n individuals, each of them being asked to rate the product in terms of $\{1, 2, \ldots, k\}$ stars, according to his/her liking for the product. Let f_j stand for the number/frequency of people giving j stars. If we denote the true probabilities of giving $1, 2, \ldots, k$ stars by $p_1, p_2, \ldots, p_k$ respectively, we have the standard multinomial scheme with $\sum_{j=1}^{k} f_j = n$ and $\sum_{j=1}^{k} p_j = 1$, which is a classical and well-studied problem. Alternatively, assume that out of these n individuals some of them hedge their bets, and assign an "interval rating" for the item. To get started and to illustrate things, let us say e.g. a given number f_{12} of people are undecided between the ratings 1 and 2, and say their rating falls in the interval $[1, 2]$ comprising both the ratings between 1 and 2. This refers to either 1 or 2 stars but s/he is not convinced over one particular rating between these two. This is what we shall refer to as an "interval rating" from now on. Given this new additional category, say with probability p_{12}, we now have $p_{12} + \sum_{j=1}^{k} p_j = 1$ and the total frequency $f_{12} + \sum_{j=1}^{k} f_j = n$. We are interested in determining how the estimated probabilities for each individual category would change if the scheme also allows such interval ratings. In other words, we wish to figure out the estimated probabilities, $\hat{p_1}, \hat{p_2}, \ldots, \hat{p_k}$ under this new scheme.

1.2. The likelihood under a general scheme

Developing the maximum likelihood estimates along with their properties such as asymptotic variances under the standard multinomial setup, has been considered extensively in the literature. One may, for instance, refer to Alam (1979) or Kunte and Upadhya (1996) where the authors have discussed both the MLEs and UMVUEs under the classical multinomial setup.

First we consider the likelihood function under our general multinomial scheme which allows interval ratings, for which we introduce some notations. Let I represent an interval (say e.g. j_1 to j_2) of categories/scores with corresponding probability $P_I = \sum_{j \in I} p_j$ for this interval. When such interval scores are allowed, out of the n individuals, let us say $m\ (\leq n)$ of them provide interval-ratings that belong to the intervals $\{I_j; j = 1, 2, \ldots, m\}$ with r of these intervals being distinct. The remaining $(n - m)$ individuals provide specific single ratings, of which let us say there are k. Then the probabilities satisfy

$$\sum_{j=1}^{r} P_{I_j} + \sum_{i=1}^{k} p_i = 1. \tag{1.1}$$

Further assume that the frequency in the interval I_j is F_j and the frequency in the k individual categories is f_i. Then $\sum_{j=1}^{r} F_j = m$ and $\sum_{i=1}^{k} f_i = n - m$, so that

$$\sum_{j=1}^{r} F_j + \sum_{i=1}^{k} f_i = n. \tag{1.2}$$

Then the likelihood for the vector $\boldsymbol{p}$ given $m, n - m, \{f_i\}$, and $\{F_j\}$ is given by:

$$L \propto \prod_{j=1}^{r} P_{I_j}^{F_j} \times \prod_{i=1}^{k} p_i^{f_i} \tag{1.3}$$

subject to the conditions (1.1) and (1.2) with the corresponding Log-likelihood

$$logL = constt. + \sum_{j=1}^{r} F_j.logP_{I_j} + \sum_{i=1}^{k} f_i.logp_i. \tag{1.4}$$

This likelihood in Eq. (1.3) is comparable to Eq. (4) in Iyer et al. (2008) or Eqn. (1) in Jammalamadaka and Leong (2015), except for the additional restrictions imposed by the conditions (1.1) and (1.2) due to the dependence among the categories, and their frequencies. Estimation for individual $p_i's$ which is our main goal, becomes even more cumbersome when some of the intervals overlap. In such cases, analytical solutions may not be possible, but one can obtain estimates through numerical methods.

To illustrate these ideas, we develop three successively more complex scenarios—labeled Cases 1, 2, and 3, and show how they can be handled. The following sections introduce corresponding likelihood functions for these three cases, provide estimators for $\boldsymbol{p} \equiv (p_1, p_2, \ldots, p_k)$ and discuss their asymptotic variances, in each of these cases.

2. Maximum likelihood estimators in some special cases

We now propose three interesting scenarios with increasing levels of complexity and provide appropriate MLEs for the probability vector $\boldsymbol{p}$. First, in "Case 1", we start by assuming that the individuals are allowed just one pre-specified interval rating besides the singleton ratings. Similarly "Case2" assumes that two such "non-overlapping" interval ratings are allowed besides the singleton ratings, whereas "Case 3" assumes that two such "overlapping" interval ratings are possible. More general scenarios are possible, and follow similar ideas.

2.1. Case 1

Assume that we only have a single "interval rating" namely $[i, j]$, with f_{ij} number of individuals opting for that. Clearly the probability of any individual giving that rating is $p_{ij} = p_i + p_{i+1} + \cdots + p_j$. Given that this is an additional category that is being allowed in the multinomial scheme, we further have

$$p_{ij} + \sum_{i=1}^{k} p_i = p_1 + \cdots p_{i-1} + 2(p_i + p_{i+1} + \cdots + p_j) + \cdots + p_{k-1} + p_k = 1. \tag{2.1}$$

The likelihood function is then proportional to

$$p_1^{f_1} \ldots p_i^{f_i} \ldots p_j^{f_j} \ldots p_k^{f_k} p_{ij}^{f_{ij}}, \tag{2.2}$$

where $1 \le i < j \le k$, with the log-likelihood

$$\log L = f_1 \log p_1 + \cdots + f_k \log p_k + f_{ij} \log p_{ij}, \tag{2.3}$$

where $p_k = (1 - p_1 - \cdots - 2p_i - 2p_{i+1} - \cdots - 2p_j - \cdots - p_{k-1})$. To obtain the MLEs, one needs to solve the following simultaneous equations,

$$\frac{f_1}{p_1} + \frac{-f_k}{p_k} = 0, \ldots, \frac{f_{i-1}}{p_{i-1}} + \frac{-f_k}{p_k} = 0, \quad \frac{f_{j+1}}{p_{j+1}} + \frac{-f_k}{p_k} = 0, \ldots, \frac{f_{k-1}}{p_{k-1}} + \frac{-f_k}{p_k} = 0$$

and,

$$\frac{f_i}{p_i} + \frac{-2f_k}{p_k} + \frac{f_{ij}}{p_{ij}} = 0, \ldots, \frac{f_j}{p_j} + \frac{-2f_k}{p_k} + \frac{f_{ij}}{p_{ij}} = 0$$

which lead to the following MLEs,

$$\hat{p_1} = \frac{f_1}{n}, \ldots, \hat{p}_{i-1} = \frac{f_{i-1}}{n}, \hat{p}_{j+1} = \frac{f_{j+1}}{n}, \ldots, \hat{p}_k = \frac{f_k}{n} \tag{2.4}$$

and,

$$\hat{p}_l = \frac{f_l}{2n}\left(\frac{f_i + f_{i+1} + \cdots + f_j + f_{ij}}{f_i + f_{i+1} + \cdots + f_j}\right) = \frac{f_l}{2n}\left(1 + \frac{f_{ij}}{f_i + f_{i+1} + \cdots + f_j}\right), \tag{2.5}$$

where $l = i, (i+1), \ldots, j$.

Remark 1. Now if $f_{ij} = 0$, i.e. no one opts for the interval rating even after being given that choice, the MLEs for the p_l in this interval will suffer because of that and reduce to become $\hat{p}_l = f_l/2n$. This is justified in view of Eq. (2.5).

Remark 2. If all the individual frequencies in the interval $[i, j]$ are zero except for one category, say just the $f_i \neq 0$, then

$$\hat{p}_i = \frac{f_i}{2n}\left(\frac{f_i + f_{ij}}{f_i}\right) = \frac{f_i + f_{ij}}{2n},$$

i.e. the ith category gets all the added benefit of this interval frequency f_{ij}. This is in agreement with the Proposition 1 of Jammalamadaka and Mangalam (2003).

2.2. Case 2

Now assume that we allow for two disjoint "interval ratings" namely, $[i_1, j_1]$ and $[i_2, j_2]$ with corresponding observed frequencies $f_{i_1 j_1}$ and $f_{i_2 j_2}$ and respective probabilities $p_{i_1 j_1}, p_{i_2 j_2}$. The forms of $p_{i_1 j_1}$ and $p_{i_2 j_2}$ are similar to those given in Section 2.1. We further have, $p_{i_1 j_1} + p_{i_2 j_2} + \sum_{i=1}^{k} p_i = 1$. The likelihood function will then be,

$$p_1^{f_1} \ldots p_{i_1}^{f_{i_1}} \ldots p_{j_1}^{f_{j_1}} \ldots p_{i_2}^{f_{i_2}} \ldots p_{j_2}^{f_{j_2}} \ldots p_k^{f_k} p_{i_1 j_1}^{f_{i_1 j_1}} p_{i_2 j_2}^{f_{i_2 j_2}}, \tag{2.6}$$

where $1 \le i_1 < j_1 < i_2 < j_2 \le k$. The log-likelihood function is clearly,

$$\log L = f_1 \log p_1 + \cdots + f_k \log p_k + f_{i_1 j_1} \log p_{i_1 j_1} + f_{i_2 j_2} \log p_{i_2 j_2} \tag{2.7}$$

where $p_k = (1 - p_1 - \cdots - 2p_{i_1} - \cdots - 2p_{j_1} - p_{j_1+1} - \cdots - 2p_{i_2} - \cdots - 2p_{j_2} - \cdots - p_{k-1})$. To obtain the MLEs, one needs to solve the following simultaneous equations,

$$\frac{f_1}{p_1} + \frac{-f_k}{p_k} = 0, \ldots, \frac{f_{i_1-1}}{p_{i_1-1}} + \frac{-f_k}{p_k} = 0, \quad \frac{f_{j_1+1}}{p_{j_1+1}} + \frac{-f_k}{p_k} = 0, \ldots, \frac{f_{i_2-1}}{p_{i_2-1}} + \frac{-f_k}{p_k} = 0, \quad \frac{f_{j_2+1}}{p_{j_2+1}} + \frac{-f_k}{p_k} = 0, \ldots,$$
$$\frac{f_{k-1}}{p_{k-1}} + \frac{-f_k}{p_k} = 0$$

and,

$$\frac{f_{i_1}}{p_{i_1}} + \frac{-2f_k}{p_k} + \frac{f_{i_1 j_1}}{p_{i_1 j_1}} = 0, \ldots, \frac{f_{j_1}}{p_{j_1}} + \frac{-2f_k}{p_k} + \frac{f_{i_1 j_1}}{p_{i_1 j_1}} = 0, \quad \frac{f_{i_2}}{p_{i_2}} + \frac{-2f_k}{p_k} + \frac{f_{i_2 j_2}}{p_{i_2 j_2}} = 0, \ldots, \frac{f_{j_2}}{p_{j_2}} + \frac{-2f_k}{p_k} + \frac{f_{i_2 j_2}}{p_{i_2 j_2}} = 0$$

which lead to the following MLEs,

$$\hat{p_1} = \frac{f_1}{n}, \ldots, \hat{p}_{i_1-1} = \frac{f_{i_1-1}}{n}, \hat{p}_{j_1+1} = \frac{f_{j_1+1}}{n}, \ldots, \hat{p}_{i_2-1} = \frac{f_{i_2-1}}{n}, \hat{p}_{j_2+1} = \frac{f_{j_2+1}}{n}, \ldots, \hat{p}_k = \frac{f_k}{n} \tag{2.8}$$

and,

$$\hat{p}_{l_1} = \frac{f_{l_1}}{2n}\left(1 + \frac{f_{i_1 j_1}}{f_{i_1} + \cdots + f_{j_1}}\right), \qquad \hat{p}_{l_2} = \frac{f_{l_2}}{2n}\left(1 + \frac{f_{i_2 j_2}}{f_{i_2} + \cdots + f_{j_2}}\right) \tag{2.9}$$

where $l_1 = i_1, (i_1 + 1), \ldots, j_1$ and $l_2 = i_2, (i_2 + 1), \ldots, j_2$.

Again, if $f_{i_1 j_1} = f_{i_2 j_2} = 0$ i.e. no individual opts for either of these interval ratings even after being given the option, then the MLEs will become $\hat{p}_{l_1} = f_{l_1}/2n$ and $\hat{p}_{l_2} = f_{l_2}/2n$, where l_1, l_2 belong to intervals given above.

Asymptotic Variances of the Estimates

Next we consider the large-sample variances of the estimates given in Eqs. (2.4), (2.5) for Case 1, and Eqs. (2.8) and (2.9) for Case 2. Since these estimates are all MLEs, the asymptotic standard errors of $\hat{p}_i$ can be computed using the corresponding information matrix. For a vector of N parameters, say $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_N]$ in the model, a typical ij^{th} element in the Fisher information matrix is given by,

$$[I(\theta)]_{ij} = E_\theta^X \left[\left(\frac{\partial}{\partial \theta_i} \log L(X; \theta) \right) \left(\frac{\partial}{\partial \theta_j} \log L(X; \theta) \right) \right].$$

Then one can obtain the large-sample variances as $V(\hat{\theta}_i) = [I^{-1}(\theta)]_{ii}, i = 1, \ldots, N$. Deriving these asymptotic variances for a general scheme is not straightforward and we provide derivations for some special cases in Appendices A.1 and A.2 (found in the supplement) corresponding to Cases 1 and 2 respectively.

2.3. Case 3

For some $i_1 < i_2 < j_1 < j_2$, if we now allow for two overlapping "interval ratings" say, $[i_1, j_1]$ and $[i_2, j_2]$, with an overlap of $[i_2, j_1]$, the likelihood function can be written similar to Eq. (2.9) and is given by:

$$p_1^{f_1} \ldots p_{i_1}^{f_{i_1}} \ldots p_{j_1}^{f_{j_1}} \ldots p_{i_2}^{f_{i_2}} \ldots p_{j_2}^{f_{j_2}} \ldots p_k^{f_k} p_{i_1 j_1}^{f_{i_1 j_1}} p_{i_2 j_2}^{f_{i_2 j_2}}, \tag{2.10}$$

where $1 \le i_1 < i_2 \le j_1 < j_2 \le k$. The log-likelihood function is clearly,

$$f_1 \log p_1 + \cdots + f_k \log p_k + f_{i_1 j_1} \log p_{i_1 j_1} + f_{i_2 j_2} \log p_{i_2 j_2} \tag{2.11}$$

where $\sum_{i=1}^{k} p_i + p_{i_1 j_1} + p_{i_2 j_2} = 1$. However because of this overlap, finding even the MLEs, leave alone their asymptotic variances, becomes very cumbersome and easy analytical solutions do not exist. However they can be obtained numerically, as we demonstrate in Section 4.

3. Bayes estimation under Dirichlet priors

Bayes estimation in a multinomial setup has been discussed by several authors—see e.g. Lehmann and Casella (1998) or Ferrie and Blume-Kohout (2016).

In this section we will adopt a Bayesian framework to estimate the unknown probability vector. Now using notations from Section 1.2, the unknown probability vector is $\boldsymbol{p} \equiv (P_{l_1}, P_{l_2}, \ldots, P_{l_r}, p_1, p_2, \ldots, p_k)$ We will now assume a prior distribution for $\boldsymbol{p}$. A natural choice would be the conjugate prior, namely the Dirichlet distribution. The setup is as follows: Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ denote the choices of the n individuals with each X_i taking either an interval or a specific score. Hence we can assume,

$$X|\boldsymbol{p} \sim \textit{Multinomial}(\boldsymbol{p})$$
$$\boldsymbol{p}|\boldsymbol{\alpha} \sim \pi(\boldsymbol{p}) \equiv \textit{Dir}(\boldsymbol{\alpha}),$$

where $Dir(\boldsymbol{\alpha})$ stands for a Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha_1^*, \ldots, \alpha_r^*, \alpha_1, \alpha_2, \ldots, \alpha_k)$, where $\alpha_i^*(> 0)$ corresponds to the respective prior parameter on each $P_{l_j}, j = 1, 2, \ldots, r$. The Dirichlet density is then given as:

$$Dir(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{r} \alpha_j^* + \sum_{i=1}^{k} \alpha_i)}{\prod_{j=1}^{r} \Gamma(\alpha_j^*) \prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{j=1}^{r} P_{l_j}^{\alpha_j^* - 1} \times \prod_{i=1}^{k} p_i^{\alpha_i - 1}. \tag{3.1}$$

Now the likelihood function ($L(data|\boldsymbol{p})$) is exactly similar to Eq. (1.1). Hence the posterior density is given by:

$$\pi(\boldsymbol{p}|data) = \frac{L(data|\boldsymbol{p})\pi(\boldsymbol{p})}{\int_{\boldsymbol{p}} L(data|\boldsymbol{p})\pi(\boldsymbol{p})d\boldsymbol{p}} \tag{3.2}$$

The numerator of (3.2) can be written as,

$$L(data|\boldsymbol{p})\pi(\boldsymbol{p}) \propto \prod_{j=1}^{r} P_{l_j}^{F_j + \alpha_j^* - 1} \times \prod_{i=1}^{k} p_i^{f_i + \alpha_i - 1}, \tag{3.3}$$

from which the posterior density can be easily written down.

We now illustrate the ideas in a simple special case namely when $k = 5$ and $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5, p_{12})$, where there exists a single "interval rating" viz. [1, 2] with a frequency of $f_{12}(> 0)$ and having a probability of $p_{12} = p_1 + p_2$. Also hence, $p_{12} + \sum_{i=1}^{5} p_i = 1$. We will now build upon the likelihood function along with the appropriate posterior distribution for $\boldsymbol{p}$. Now as in previous sections, the likelihood function can be written as,

$$L(data|\boldsymbol{p}) = \prod_{i=1}^{5} p_i^{f_i} (p_1 + p_2)^{f_{12}}. \tag{3.4}$$

Further, similar to (3.3),

$$L(data|\boldsymbol{p})\pi(\boldsymbol{p}) = (p_1 + p_2)^{f^*} \prod_{i=1}^{5} p_i^{f_i + \alpha_i - 1}, \tag{3.5}$$

where $f^* = f_{12} + \alpha_1^* - 1$, which is > 0 since we assume $f_{12} \geq 1$ and $\alpha_1^* > 0$. The expression in (3.5) can be thought of as a Dirichlet distribution with a different set of parameters. Further,

$$\int_{\boldsymbol{p}} L(data|\boldsymbol{p})\pi(\boldsymbol{p})d\boldsymbol{p} = \frac{\Gamma(f^* + 1) \prod_{i=1}^{5} \Gamma(f_i + \alpha_i)}{\Gamma(n + \sum_{i=1}^{5} \alpha_i + \alpha_1^*)} = c^*(say). \tag{3.6}$$

Combining (3.5) and (3.6) we have the posterior density, from which one can obtain the Bayes estimate of $\boldsymbol{p}$. Under Squared Error Loss, it is given by the mean of this posterior. In particular, the Bayes estimator of p_i is given by

$$\begin{aligned} \hat{p}_{i(Bayes)} &= \frac{\int_{\boldsymbol{p}} p_i \prod_{j=1}^{5} p_j^{f_j + \alpha_j - 1} (p_1 + p_2)^{f^*} d\boldsymbol{p}}{c^*} = \frac{\Gamma(f^* + 1)\Gamma(f_i + \alpha_i + 1) \prod_{j=2}^{5} \Gamma(\alpha_j + f_j)}{\Gamma(n + \sum_{j=1}^{5} \alpha_j + \alpha_1^* + 1)c^*} \\ &= \frac{f_i + \alpha_i}{n + \sum_{j=1}^{5} \alpha_j + \alpha_1^*}, \qquad i = 1, 2, 3, 4, 5 \end{aligned} \tag{3.7}$$

Further,

$$\begin{aligned} \hat{p_{12}}_{(Bayes)} &= \frac{\int_{\boldsymbol{p}} (p_1 + p_2) \prod_{j=1}^{5} p_j^{f_j + \alpha_j - 1} (p_1 + p_2)^{f^*} d\boldsymbol{p}}{c^*} = \frac{\Gamma(f^* + 2) \prod_{j=1}^{5} \Gamma(\alpha_j + f_j)}{\Gamma(n + \sum_{j=1}^{5} \alpha_j + \alpha_1^* + 1)c^*} \\ &= \frac{f_{12} + \alpha_1^*}{n + \sum_{j=1}^{5} \alpha_j + \alpha_1^*} \end{aligned} \tag{3.8}$$

4. Parametric bootstrapping and real data analysis for the estimates and variances

As an alternative to finding the MLEs and their asymptotic variances, which as we can see, gets complicated pretty quickly, one might adopt a parametric bootstrap to get the estimates and the variances of the estimates for the 3 cases discussed in Sections 2.1–2.3. These are obtained by first using the relative frequencies as the initial probabilities, and bootstrapping/simulating a large number of independent samples. As an illustration and demonstration that they provide similar results, we first present results for such bootstrapping for "Case-1", alongside the results for our Bayesian setup. Results for "Case-2" and "Case-3" can be derived similarly (see Remark at the end of Section 4.1).

Table 1
Illustrative results for Case 1 with $n = 100$.

i	1	2	3	4	5	[1, 2]
f_i	15	5	15	10	25	30
f_i/n	0.15	0.05	0.15	0.10	0.25	0.30
$\hat{p}_i^{MLE}$ $(s(\hat{p}_i^{MLE}))$	0.1875 (.0286)	0.0625 (0.0225)	0.1500 (0.0357)	0.1000 (0.0300)	0.2500 (0.0433)	0.2500 (0.0250)
$\hat{p}_i^R$ $(s(\hat{p}_i^R))$	0.1873 (0.0281)	0.0623 (0.0220)	0.1493 (0.0346)	0.1016 (0.0299)	0.2495 (0.0430)	0.2497 (0.0242)
$\hat{p}_{i(Bayes)}$	0.1451	0.0691	0.1393	0.1017	0.2499	0.2946

(The standard errors are shown within parentheses).

4.1. Illustrative results for case 1

We will consider the case as given in Section 2.1, where we allow a single "interval rating". Now let $i = 1, j = 2$ and $k = 5$. The likelihood and log-likelihood functions are as given in (2.2), (2.3) and in particular take the following forms:

$$L = p_1^{f_1} p_2^{f_2} p_3^{f_3} p_4^{f_4} p_5^{f_5} (p_1 + p_2)^{f_{12}},$$

and,

$$\log L = f_1 \log p_1 + f_2 \log p_2 + f_3 \log p_3 + f_4 \log p_4 + f_5 \log p_5 + f_{12} \log(p_1 + p_2),$$

where $p_5 = (1 - 2p_1 - 2p_2 - p_3 - p_4)$ and f_{12} denotes the only "interval rating". Now we first fix an observed vector of frequencies and assume it to come from a multinomial distribution with parameter vector $\boldsymbol{p}$, where $\boldsymbol{p} = [p_1, p_2, p_3, p_4, p_5, p_{12}]$. We intentionally fix f_{12} to be higher than both f_1 and f_2, since it is reasonable to assume that in any practical scenario when given an option, more people will likely opt for an interval rating instead of giving a single number. We then calculate the estimated probabilities using (2.4) and (2.5), which take the following forms,

$$\hat{p}_1 = \frac{f_1}{2n}\left(1 + \frac{f_{12}}{f_1 + f_2}\right), \hat{p}_2 = \frac{f_2}{2n}\left(1 + \frac{f_{12}}{f_1 + f_2}\right), \hat{p}_i = \frac{f_i}{n},$$

for $i = 3, 4, 5$. Now assuming these estimates are the actual probabilities ($\hat{\boldsymbol{p}} \equiv \boldsymbol{p}$), we bootstrap a large number of samples (say $R = 10^3$) from a $Mult(\boldsymbol{p})$ distribution, and recalculate the probability estimates using (2.4) and (2.5). We then observe the pattern of the estimates by looking at the mean and standard errors of these estimates over the R bootstraps. Let these be denoted by $\hat{\boldsymbol{p}}^R$ and $s(\hat{\boldsymbol{p}}^R)$ respectively. The asymptotic variances of $\hat{\boldsymbol{p}}$ take the following forms:

$$V(\hat{p}_1) = \frac{p_1[p_1(1 - 2p_1) - 2p_1p_2 + 2p_2]}{2n(p_1 + p_2)};\ V(\hat{p}_2) = \frac{p_2[p_2(1 - 2p_2) - 2p_1p_2 + 2p_1]}{2n(p_1 + p_2)};$$
$$V(\hat{p}_{12}) = \frac{1}{2n}(p_1 + p_2)(p_3 + p_4 + p_5);\ V(\hat{p}_i) = \frac{p_i(1 - p_i)}{n},\ i = 3, 4, 5.$$

The above expressions for the asymptotic variances are derived in Appendix A.1 (found in the supplement).

We also provide the Bayes estimates, using Dirichlet priors for the given data sets. Tables 1 and 2 outline the results for two different values of n, for a given set of observed frequencies. For our Bayesian setup, as our first instance we fix $\boldsymbol{\alpha} = (2, 3, 1, 2, 4, 4)$, whereas as a second instance we fix $\boldsymbol{\alpha} = (4, 6, 1, 1, 2, 4)$. Now if one compares values of $\hat{p}_i^R$ and f_i/n across all values of i, the interval rating [1, 2] puts an additional mass on both $\hat{p}_1^R$ and $\hat{p}_2^R$ while all others stay comparable. Intuitively, the jumps from f_1/n to $\hat{p}_1^R$ and f_2/n to $\hat{p}_2^R$ are proportional to f_1 and f_2.

Remark 3. One can obtain bootstrapping results for "Case-2" and "Case-3" in a similar manner. The likelihood and log-likelihood functions for "Case-2" can be written down along the lines of (2.6) and (2.7), whereas the estimated probabilities are as per (2.8) and (2.9). Expressions for the asymptotic variances of the estimated probabilities for "Case-2" are derived in Appendix A.2 (found in the supplement). Further, the likelihood and log-likelihood functions for "Case-3" can be written down along the lines of (2.10) and (2.11). However the estimated probabilities do not have closed-form analytical solutions and require either a numerical maximization or parametric bootstrapping, as demonstrated in this section for "Case 1".

4.2. Analysis for a real data set

We now present a real data example which demonstrates the applicability of estimators discussed here. As the entire world suffers from the current COVID-19 pandemic, there has been an ever increasing demand for a software where a group is able to conduct online meetings and live sessions. Many competitors have cropped up in the market. The following survey was conducted at the University of California, Santa Barbara by one of the authors recently, which asked a class of students about their overall experience with regard to one such widely used software. They could give ratings of 1 through 5 (5 being the highest rating) along with a couple of "interval ratings" consisting of [1, 2] and [4, 5]. Data is

Table 2
Illustrative results for Case 1 with $n = 500$.

i	1	2	3	4	5	[1, 2]
f_i	90	70	40	80	20	200
f_i/n	0.18	0.14	0.08	0.16	0.04	0.40
$\hat{p}_i^{MLE}$ $(s(\hat{p}_i MLE))$	0.2025 (0.0144)	0.1575 (0.0140)	0.08 (0.0121)	0.16 (0.0163)	0.04 (0.0087)	0.36 (0.0100)
$\hat{p}_i^{R}$ $(s(\hat{p}_i^{R}))$	0.2020 (0.0145)	0.1580 (0.0146)	0.0800 (0.0121)	0.1596 (0.0166)	0.0401 (0.0089)	0.3600 (0.0102)
$\hat{p}_{i(Bayes)}$	0.1814	0.1467	0.0791	0.1563	0.0424	0.3938

(The standard errors are shown within parentheses).

Table 3
Real data example for Case 2 with $n = 90$.

i	1	2	3	4	5	[1, 2]	[4, 5]
f_i	2	8	13	21	6	11	29
f_i/n	0.022	0.088	0.144	0.233	0.066	0.122	0.322
$\hat{p}_i^{MLE}$	0.023	0.093	0.144	0.241	0.069	0.116	0.31

collected from a group of $n = 90$ students from the class. This falls in the paradigm of our current problem, in particular, "Case-2", given in Section 2.2, and the likelihood and log-likelihood functions take the following forms:

$$L = p_1^{f_1} p_2^{f_2} p_3^{f_3} p_4^{f_4} p_5^{f_5} (p_1 + p_2)^{f_{12}} (p_4 + p_5)^{f_{45}},$$

and,

$$\log L = f_1 \log p_1 + \cdots + f_5 \log p_5 + f_{12} \log(p_1 + p_2) + f_{45} \log(p_4 + p_5),$$

where $p_3 = (1 - 2p_1 - 2p_2 - 2p_4 - 2p_5)$. We then obtain the estimated probabilities (MLEs) from (2.8) and (2.9). Table 3 outlines results for these students' observed frequencies. Note that all these results are from a single run (single question in the survey).

As can be seen from the above table, the intervals [1, 2] and [4, 5] put a slight additional mass on $\hat{p}_1^{MLE}, \hat{p}_2^{MLE}$ as well as $\hat{p}_4^{MLE}, \hat{p}_5^{MLE}$ compared to the original values of f_i/n, $i = 1, 2, 4, 5$.

5. Conclusions

In this paper we develop a middle-censoring scheme under a multinomial setup which allows outcomes to fall within intervals besides individual categories. Although the general framework has been presented for estimating the individual probabilities, analytical solutions become onerous pretty quickly and may need numerical solutions. To illustrate the ideas, we consider special cases and demonstrate how the Maximum Likelihood Estimators work out, as well as under a Bayesian setup. Also provided are the asymptotic variances of the multinomial probability vector under these cases. Parametric bootstrap has been suggested for getting the estimates and their variances when the MLEs get complicated. A real data analysis is carried out illustrating the results derived.

CRediT authorship contribution statement

S. Rao Jammalamadaka: Building methodology, Providing different examples, Reviewing, Editing of the draft. **Sudeep R. Bapat:** Building methodology, Providing different examples, Reviewing, Editing of the draft.

Acknowledgments

We wish to thank Professor Yonathan Arbel of the University of Alabama Law School who presented one of the authors with this problem, and the Referee for a careful reading and constructive suggestions.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2020.108916.

References

Ahmadi, K., Rezaei, M., Yousefzadeh, F., 2017. Statistical analysis of middle-censored competing risks data with exponential distribution. J. Statist. Comput. Simul. 16, 3082–3110.
Alam, K., 1979. Estimation of multinomial probabilities. Ann. Stat. 7, 282–283.
Davarzani, N., Parsian, A., 2011. Statistical inference for discrete middle-censored data. J. Statist. Plan. Inference 141, 1455–1462.
Ferrie, C., Blume-Kohout, R., 2016. Bayes estimator for multinomial parameters and Bhattacharyya distances. https://arxiv.org/abs/1612.07946.
Iyer, S.K., Jammalamadaka, S.R., Kundu, D., 2008. Analysis of middle censored data with exponential lifetime distributions. J. Statist. Plan. Inference 138, 3550–3560.
Jammalamadaka, S.R., Iyer, S.K., 2004. Approximate self consistency for middle censored data. J. Statist. Plan. Inference 124, 75–86.
Jammalamadaka, S.R., Leong, E., 2015. Analysis of discrete lifetime data under middle-censoring and in the presence of covariates. J. Appl. Stat. 42, 905–913.
Jammalamadaka, S.R., Mangalam, V., 2003. Non-parametric estimation for middle-censored data. J. Nonparametr. Stat. 15, 253–265.
Kunte, S., Upadhya, K.S., 1996. Estimating multinomial probabilities. Am. Stat. 50, 214–216.
Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation. Springer.