sciendo

Journal of Official Statistics, Vol. 39, No. 1, 2023, pp. 103–121, http://dx.doi.org/10.2478/JOS-2023-0005

A Statistical Comparison of Call Volume Uniformity Due to Mailing Strategy

Andrew M. Raim[1], *Elizabeth Nichols*[1], *and Thomas Mathew*[1]

For operations such as a decennial census, the U.S. Census Bureau sends mail to potential respondents inviting a self-response. It is suspected that the mailing strategy affects the distribution of call volumes to the U.S. Census Bureau's telephone helplines. For staffing purposes, more uniform call volumes throughout the week are desirable. In this work, we formulate tests and confidence intervals to compare uniformity of call volumes resulting from competing mailing strategies. Regarding the data as multinomial observations, we compare pairs of call volume observations to determine whether one mailing strategy has multinomial cell probabilities closer to the uniform probability vector compared to another strategy. A motivating illustration is provided by call volume data recorded in three studies which were carried out in advance of the 2020 Decennial Census.

Key words: Entropy; multinomial distribution; self-response; uniform probability vector.

1. Introduction

1.1. Background

Beginning with the 1990 Census, a telephone questionnaire assistance operation has accompanied each decennial census. These operations included helpline agents representing the U.S. Census Bureau who fielded support questions and assisted the public in completing paper forms. An automated interactive voice response system was added after the 1990 Census to augment live agents. Since the 2000 Census, agents were additionally able to conduct interviews and collect census data by phone, although this has not been marketed as a response option. For the 2020 Census, the U.S. Census Bureau mailed letters and postcards to residential addresses in the country to request participation in the census, and encouraged responses on a large scale through the internet (U.S. Census Bureau 2017). Telephone helplines were highlighted in mailings, both as a means to assist with questions about the census and to serve as a mode of response in itself. From the perspective of the U.S. Census Bureau, an ideal distribution of helpline calls would be where a uniform volume of calls is received throughout the week for the duration of the

[1] U.S. Census Bureau, Washington, D.C., 20233, U.S.A. Emails: andrew.raim@census.gov, elizabeth.may.nichols@census.gov, and thomas.mathew@census.gov.

Acknowledgments: We thank Sarah Konya and Julia Coombs for providing the call data for the experiments studied in this work. Thanks to Darcy Steeg Morris, Julia Coombs, Kevin Zajac, Tommy Wright, Eric Slud, and the Decennial Research Objective and Methods (DROM) committee for their feedback on this report. Helpful comments from the reviewers and an Associate Editor are gratefully acknowledged. This article is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not those of the U.S. Census Bureau. The article has been reviewed for disclosure avoidance and approved under CBDRB-FY21-CBSM001-004.

© Statistics Sweden

operation. More uniform call volumes would provide a more constant workload for helpline staff over the course of the operation. Furthermore, a reduction of large peaks in volume may reduce the need for increased staff.

The schedule of mailings influences when calls tend to occur and is an aspect of census design which U.S. Census Bureau can control. Chesnut (2003) and Zajac (2012) discuss call volumes received during the 2000 and 2010 Decennial Census, respectively, and note how they were affected by the mailing schedule. Helplines received 7.9 million calls in the 1990 Census, six million calls in the 2000 Census, and 4.5 million calls in the 2010 Census. Over 13 million calls were received by the system used in 2020. The interactive voice response system deflected most of them, but nearly five million incoming calls were handled by customer service representatives. Although volumes of calls have changed over the decades, patterns of calls to the helpline have not changed drastically (Nichols et al. 2019). Higher call volumes occur on the expected delivery date of mailed census notification letters and postcards. The first peaks occur after the initial mailout and second mailout, and another peak occurs the week of Census Day. There is also a trend in which Mondays and Tuesdays are the highest call volume days, with a gradual decline in call volume throughout the week and a large dropoff over the weekend. This pattern is more exaggerated when mail arrives on a Monday, as was the case in the 2000 Census (Chesnut 2003). The volume of calls diminishes after both Census Day and the arrival of all mailed notifications have occurred. Expecting similar patterns with increased volume for the 2020 Census, the U.S. Census Bureau adopted staggered mailings of census notifications to spread delivery to multiple days of the week and multiple weeks of the month, anticipating that daily calls to the helpline would be more uniform and thus easier to staff efficiently (Nichols et al. 2019). Note that other aspects of a mailing schedule–such as potential impact to response rates – are important to the U.S. Census Bureau as well; however, the remainder of this article focuses on call uniformity.

The U.S. Census Bureau conducts operations throughout the decade to prepare for the decennial census. Several studies leading up to the 2020 Census featured mailings with an invitation to respond via internet where data on subsequent calls to helplines were recorded. These data provide an opportunity to compare variations between mailing strategies and the uniformity of the resulting call volumes. Assuming that the data are realizations from a multinomial distribution, in this article we compare pairs of call volume observations and develop a methodology to assess whether one mailing strategy provides multinomial cell probabilities closer to the uniform probability vector compared to another mailing strategy.

1.2. On the Kullback-Leibler Divergence

To our knowledge, inference comparing the closeness of two discrete distributions to a discrete uniform distribution is not standard. Many conventional tests are primarily designed to detect departure from equality of two distributions. However, the equality of two distributions is not the primary interest in our application; our problem is that of comparing two sets of multinomial cell probabilities in terms of their relative closeness to the uniform probability vector, and it seems appropriate to consider statistics used for assessing goodness of fit. While several empirical distribution based statistics are available

for this (for example, the Kolmogorov-Smirnov statistic; see Li et al. 2005), a statistic based on the Kullback-Leiber divergence is an option that has been investigated in the literature. As is well known, the K-L divergence measure is a directed divergence measure that is not symmetric with respect to the two distributions (see Li et al. 2005; Cover and Thomas 2006). This is appropriate for our purpose since our focus is on a measure of closeness directed towards the uniformity of multinomial cell probabilities.

In a book on information theory, Cover and Thomas (2006) discuss the K-L distance, entropy, and related concepts, and discuss fundamental applications in information theory. Dorfinger et al. (2011) use entropy as a measure of uniformity to classify in real time whether traffic in computer networks is encrypted or not. Their approach makes a decision based on the difference between the estimated entropy of an observed payload and that of a uniformly distributed random payload of the same length. Liu and Wang (2004) and Cohen et al. (2006) consider an increasing convex ordering among discrete distributions; when this ordering holds, one particular consequence is that one of the distributions has a larger entropy than the other and is therefore closer to uniform. K-L divergence and entropy have found use in many areas of the statistics literature, including: to justify information criteria in assessing model fits (Konishi and Kitagawa 2008), to obtain variational approximations to complicated distributions such as the posterior in Bayesian analysis (Ormerod and Wand 2010; Blei et al. 2017), and as a basis for statistical inference (Pardo 2006; Girardin and Lequesne 2019). Paninski (2008) proposed a method to test whether a single multinomial distribution departs from discrete uniform; this work is based on a sparse setting with many categories and relatively few observations.

From the perspective of testing goodness-of-fit, Several authors have investigated the K-L divergence measure and its properties. The article by Girardin and Lequesne (2019) gives a review of the literature on entropy-based goodness-of-fit tests. A test that is asymptotically distribution-free is investigated in Song (2002). Regarding the K-L divergence measure based test, Song (2002) notes that "With its good power properties, the method provides an extremely simple and potentially much better alternative to the classical empirical distribution function (EDF)-based test procedures". A Monte Carlo simulation study comparing several goodness of fit tests is reported in Li et al. (2005). Along with a test based on K-L divergence, the tests considered by these authors include those based on the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling statistics. The authors note that "... the tests based on the directed divergence measure give a good approximation to the given significance levels and are more powerful than other tests against the given alternative distributions". In an earlier article comparing models that are relevant in economics, Vuong (1989) used the K-L divergence criterion to measure the closeness of a model to the true model. The author has developed likelihood-ratio based statistics for testing the null hypothesis that two competing models are equally close to the true model; the alternative hypothesis being that one model is closer. Our problem is similar to what Vuong (1989) has investigated. We believe that these articles provide a strong motivation for using the K-L divergence measure to address the problem we have investigated.

The rest of the article proceeds as follows. Section 2 introduces the call volume data and provides some exploratory analysis. Section 3 discusses test and confidence interval procedures based on the K-L divergence criterion in order to compare the closeness of two discrete

distributions to uniformity. Section 4 gives results of the data analysis. Finally, Section 5 concludes the article. Additional technical details of the statistical procedures are given in the appendices. Simulation studies to validate the procedures are provided as a online supplement.

2. Call Volume Data

A standard mailing strategy would be to send mail to all intended recipients on a common schedule. Alternatively, consider a staggered strategy where respondents are randomly partitioned into several groups which are sent mailings on different schedules. It is thought that compared to an unstaggered strategy, a staggered strategy leads to a more uniform distribution of calls, in the sense that the probability of a mail recipient calling the helpline on any day of the week is closer to being uniform (i.e., equal to 1/7). To explore this theory, we make use of data sets from three studies in which up to four mailings were sent to a target population and subsequent calls to census helplines were recorded. Each of these operations is referred to as a National Census Bureau Survey (NCBS) in mailing materials. The materials in these studies mimicked typical U.S. Census Bureau mailings for production surveys at that time. The first and fourth mailings included a one page letter with the help phone line and the survey URL. These mailings also included a thicker stock card with an authentication code for the web survey and the mailing address which showed through the envelope window. The second and third mailings were postcards which also included the phone line, URL, and authentication codes. The 2016 September NCBS (Eggleston and Coombs 2017) and 2016 June NCBS (Coombs 2017) are two studies where an unstaggered mailing strategy was utilized. A staggered mailing schedule was used in the 2017 March NCBS (Nichols et al. 2019); mail recipients were randomly assigned into either a Monday Mailout group, to whom three out of four mailings were initiated on Mondays, or a Thursday Mailout, where three out of four mailings were initiated on Thursdays.Table 1 displays the schedules for each mailing in the three studies. In all three operations, no live agents were present to answer the helpline and callers instead received a prerecorded message. Callers' identities were not recorded so that the data do not distinguish whether multiple calls were made by the same caller.

For each study, we designate day 1 as the day of the first mailing. For the 2017 March NCBS, where there are two mailing schedules, day 1 is the day of the very first mailing, Monday May 6. We then designate week 1 as days 1–7, week 2 as days 8–14, and so forth. We consider weeks 1–5 in each study, and disregard calls which occurred in week 6 or later because call activity becomes sparse. For each pair of studies, we compare the uniformity for week j of the first study to week j of the second study, for $j = 1, \ldots, 5$. It is possible to consider other methods of designating weeks, such as counting each Sunday as the start of a new week; however, our main interest is in call behavior relative to the mailing schedule. We also considered dropping weekends or consolidating Saturday and Sunday into a single "weekend" category, but decided to keep weekends intact. Changing designations of weeks could substantially change results, and such a choice should ideally not be based on the observed data. Table 2 reports the weekly call counts for each study. It is also possible to compare different weeks from pairs of studies, but this yields a large number of possible comparisons. According to our definition of weeks, there is very little opportunity to receive calls in week 1 of the 2017 March NCBS for the Thursday Mailout

Table 1. The schedule of mailings for each study. The 2017 March NCBS (a) was targeted to 8,000 recipients. Of these, half were assigned to the Monday mailout group, and half were assigned to the Thursday mailout group. The 2016 September NCBS (b) was targeted to 9,000 recipients. The 2016 June NCBS (c) was targeted to 8,000 recipients.

(a) 2017 March NCBS.

	Monday Mailout Group		Thursday Mailout Group	
Mailing	Date	Day of Week	Date	Day of Week
1	March 6, 2017	Monday	March 9, 2017	Thursday
2	March 9, 2017	Thursday	March 13, 2017	Monday
3	March 20, 2017	Monday	March 23, 2017	Thursday
4	March 27, 2017	Monday	March 30, 2017	Thursday

(b) 2016 September NCBS.

Mailing	Date	Day of Week
1	August 25, 2016	Thursday
2	September 1, 2016	Thursday
3	September 8, 2016	Thursday
4	September 15, 2016	Thursday

(c) 2016 June NCBS.

Mailing	Date	Day of Week
1	June 13, 2016	Monday
2	June 15, 2016	Wednesday
3	June 24, 2016	Friday
4	July 5, 2016	Tuesday

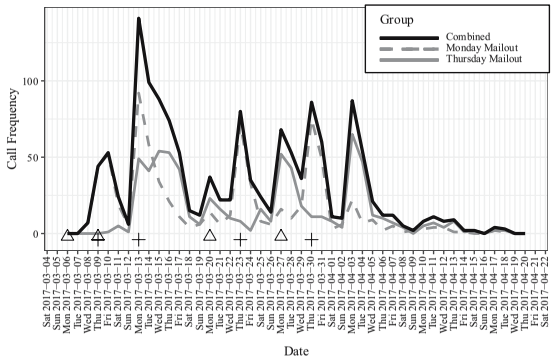
Table 2. Call counts by designated week of study.

			2017 March			
Week	2016 June	2016 Sept	Mon	Thu	Total	$\hat{\pi}_j$
1	353	490	127	7	134	0.9478
2	747	689	226	256	482	0.4414
3	757	970	151	83	234	0.6453
4	383	528	177	147	324	0.5463
5	273	129	48	145	193	0.2487
Total	2513	2806	729	638	1367	

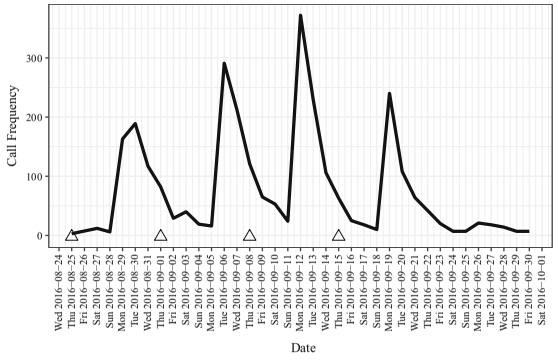
Note: The quantity $\hat{\pi}_j$ in the last column, defined in the next section, is the ratio of the entries under the Mon column to those in the Total column.

group (seven calls). Therefore, if increased uniformity is observed in week 1 for the 2017 March NCBS, it is likely due to some factor other than the staggered mailing schedule.

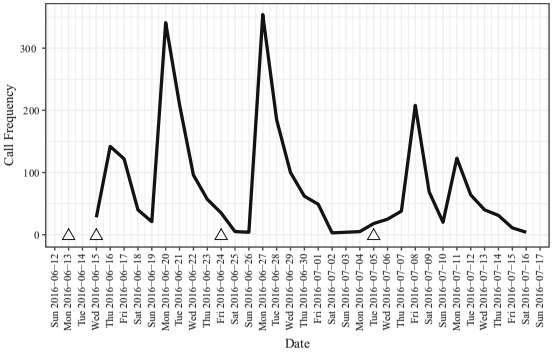
To compare uniformity of call volumes in the three studies, we first examine plots of call frequencies. Figure 1 presents daily call volumes for the three studies. Mailing dates are marked in each plot. Receipt times of the mailings were not known precisely; however, spikes in call volumes can be observed about three days after each mailing, or on the following workday if the



(a) 2017 March NCBS.



(b) 2016 September NCBS.



(c) 2016 June NCBS.

Fig. 1. Daily call volumes during the three NCBS studies. In (a), triangles (Δ) and pluses (+) along the x-axis represent mailing dates for Monday and Thursday Mailout groups, respectively. In (b) and (c), triangles (Δ) represent the mailing dates.

third day happened to fall on a weekend. Even with a staggered mailout, Figure 1a exhibits spikes on Mondays for both mailing schedules, as well as on the expected delivery date of Thursday for the Monday Mailout group. Because the expected delivery day is Monday for the Thursday Mailout group, call volumes primarily spike on Mondays and decrease throughout the rest of the week. For the 2016 June NCBS, spikes can again be seen either on Mondays or three days after a mailing if that day fell on a weekday. A similar pattern can be seen in the 2016 September NCBS; note that September 5, 2016 was a Labor Day holiday so that the expected spike in call volume shifted from Monday to the next day (September 6). Figure 2 presents call frequencies summed by day of week for each study. It appears that the distribution for the 2017 March NCBS (Monday and Thursday mailing groups combined) is flatter than either the distributions for the 2016 September NCBS or 2016 June NCBS. Figure 2d exhibits a large spike on Monday, and therefore appears to be the furthest from being uniform.

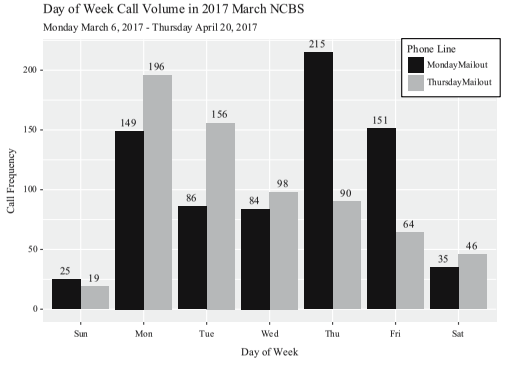
Section 3 describes a framework to carry out more formal statistical analyses. We then return to analysis of the data in Section 4 and report the results. It should be emphasized that the three data sets were collected in three separate NCBS studies, and not as the result of a controlled experiment. Therefore, significant differences in uniformity might be suspected to be caused by factors other than staggering which were not controlled. Such factors include timing of the mailings and timing of the study within the years 2016–2017. Although conclusions which can be drawn from the data are limited, the data provide a concrete example that illustrates the methodology.

3. Methodology

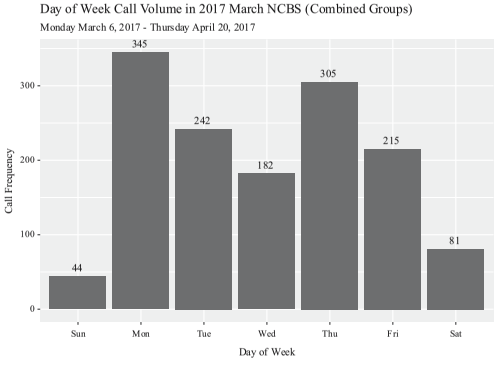
Suppose $\boldsymbol{p} = (p_1, \ldots, p_K)$ and $\boldsymbol{q} = (q_1, \ldots, q_K)$ are probability distributions on categories labeled $1, \ldots, K$. In our application, categories $1, \ldots, K$ represent the $K =$ seven days of the week: Sun, Mon, $\ldots$, Sat, and $\boldsymbol{p}$ and $\boldsymbol{q}$ are probabilities of a mail recipient calling the helpline on those days, given that the call will occur during that week. In general, we can consider $\boldsymbol{p}$ and $\boldsymbol{q}$ to be probability vectors on any K categories. Let $D(\boldsymbol{p}, \boldsymbol{q}) = \sum_{k=1}^{K} p_k \log\left(p_k/q_k\right)$ denote the K-L divergence between distributions with probabilities $\boldsymbol{p}$ and $\boldsymbol{q}$. Let $\varepsilon(p) = -\sum_{k=1}^{K} p_k \log p_k$ denote the entropy of a distribution with probabilities $\boldsymbol{p}$. Write $\bar{e} = (1/K, \ldots, 1/K)$ and let $\boldsymbol{e}_k$ be a point mass on the kth category; that is, a vector with a 1 in the kth position and zeros in the remaining $K-1$ positions.

As noted in the introduction, the methodology developed in this section is based on the K-L divergence measure. We will say that $\boldsymbol{q}$ is "more uniform" than $\boldsymbol{p}$ if $\boldsymbol{q}$ is closer to $\bar{e}$ than $\boldsymbol{p}$ in terms of K-L divergence; in other words, if

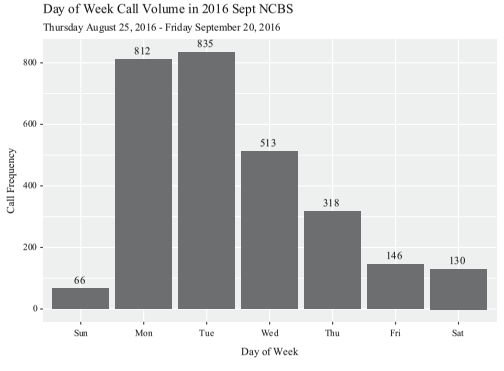
$$
\begin{aligned}
& D(\boldsymbol{p}, \bar{e}) > D(\boldsymbol{q}, \bar{e}) \\
& \Leftrightarrow \left[\sum_{k=1}^{K} p_k \log p_k - \sum_{k=1}^{K} p_k \log (1/K)\right] > \left[\sum_{k=1}^{K} q_k \log q_k - \sum_{k=1}^{K} q_k \log (1/K)\right] \\
& \Leftrightarrow \sum_{k=1}^{K} p_k \log p_k > \sum_{k=1}^{K} q_k \log q_k \\
& \Leftrightarrow \varepsilon(p) < \varepsilon(q). \qquad (1)
\end{aligned}
$$



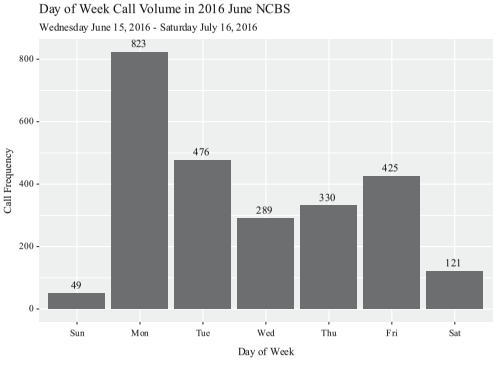
(a) 2017 March NCBS; Monday and Thursday Mailout groups separated.



(b) 2017 March NCBS; Monday and Thursday Mailout groups combined.



(c) 2016 September NCBS.



(d) 2016 June NCBS.

Fig. 2. Day-of-week call volumes for three NCBS studies. The 2017 March NCBS is displayed with Monday and Thursday Mailout groups separated in (a) and combined in (b).

It is well known that entropy is maximum at $\bar{e}$,

$$\varepsilon(p) \leq \varepsilon(\bar{e}) = \log K,$$

over any $\boldsymbol{p} = (p_1, \ldots, p_K)$, and minimized by the point masses,

$$\varepsilon(p) \geq \varepsilon(e_k) = 0, \text{ for any } k = 1, \ldots, K.$$

Suppose $\boldsymbol{p}$ and $\boldsymbol{q}$ are parameterized by unknown $\boldsymbol{\theta}$ to be discussed later in this section. Let $g(\boldsymbol{\theta}) = \varepsilon(\boldsymbol{q}) - \varepsilon(\boldsymbol{p})$ represent the difference in entropy. Motivated Equation (1), we will consider testing hypotheses of the form

$$H_0 : g(\boldsymbol{\theta}) = 0 \text{ vs. } H_1 : g(\boldsymbol{\theta}) \neq 0, \tag{2}$$

$$H_0 : g(\boldsymbol{\theta}) \leq 0 \text{ vs. } H_1 : g(\boldsymbol{\theta}) > 0, \tag{3}$$

$$H_0 : g(\boldsymbol{\theta}) \geq 0 \text{ vs. } H_1 : g(\boldsymbol{\theta}) < 0. \tag{4}$$

If H_0 is rejected in Equation (3), for example, we conclude that $\boldsymbol{q}$ has a higher entropy, or equivalently that $\boldsymbol{q}$ is closer to the uniform probability vector. Note that these hypotheses are invariant to the order of elements in both $\boldsymbol{p}$ and $\boldsymbol{q}$; this is desirable for call volume analysis because we are primarily interested in comparing flatness of distributions, and not whether volumes have simply shifted to different days of the week. In addition to hypothesis testing, we consider point estimates and confidence intervals for the quantity $g(\boldsymbol{\theta})$. We provide additional information on the statistical properties in Appendix (Section 6).

We make note of several points before proceeding:

(A) As a guide to interpret the size of the effect $g(\boldsymbol{\theta})$, recall that $0 \leq \varepsilon(\boldsymbol{p}) \leq \log K$ for any $\boldsymbol{p}$, so that $-\log K \leq g(\boldsymbol{\theta}) \leq \log K$.

(B) Let $\varepsilon_a(\cdot)$ denote the entropy function where logarithms are taken under base a. Here, $\varepsilon_a(\boldsymbol{q}) - \varepsilon_a(\boldsymbol{p}) = [\log a]^{-1} g(\boldsymbol{\theta})$, so that the change of base serves only to scale our quantity of interest by a constant. Therefore, we will consider natural logarithms for the remainder of the article.

(C) It is possible to compare the entropies of $\boldsymbol{p}$ and $\boldsymbol{q}$ with different numbers of support points. If $\boldsymbol{p} = (p_1, \ldots, p_{K_1})$, $\boldsymbol{q} = (q_1, \ldots, q_{K_2})$, and $\bar{e}^{(K)} = (1/K, \ldots, 1/K)$, we obtain the analog to Equation (1) that $D(\boldsymbol{p}, \bar{e}^{(K_1)}) > D(\boldsymbol{q}, \bar{e}^{(K_2)}) \Leftrightarrow \varepsilon(\boldsymbol{p}) - \log K_1 < \varepsilon(\boldsymbol{q}) - \log K_2$. For example, multinomial outcomes from a month with $K_1 = 31$ days may be compared to a month with $K_2 = 28$ days.

(D) As noted in the introduction, and as is well known, the K-L divergence measure is a directed divergence measure that is not symmetric with respect to the two distributions. This is appropriate for our purpose since our focus is on a measure of closeness directed towards uniformity of the probability vectors.

Let $\boldsymbol{X} \sim \text{Mult}_K(m, \boldsymbol{p})$ denote that random variable $\boldsymbol{X}$ has a multinomial distribution

$$\text{P}(X = x) = \frac{m!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K}, \; x \in \{0, 1, \ldots, m\}^K, \; \sum_{k=1}^{K} x_k = m.$$

Consider the comparison of two studies where a total of I mailing schedules were attempted among the two. For the ith mailing schedule, let J_i denote the total number of

weeks of the study. In our data, all J_i are equal and represented by a common J. Define $\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijK})$ as the call counts observed on (Sun, Mon,. . ., Sat) on the jth week for the ith mailing schedule, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. We will assume that

$$X_{ij} \overset{\text{ind}}{\sim} \text{Mult}_K\left(m_{ij}, p_{ij}\right), \tag{5}$$

where $\boldsymbol{p}_{ij} = (p_{ij1}, \ldots, p_{ijK})$ is the day-of-week distribution and $m_{ij} = \sum_{k=1}^{K} X_{ijk}$ is the total call count on the jth week for the ith mailing schedule. With no further assumptions about probabilities $\boldsymbol{p}_{ij}$, the unknown parameter is $\theta = \left(p_{11}^{-}, \ldots, p_{1J}^{-}, \ldots, p_{I1}^{-}, \ldots, p_{IJ}^{-}\right)$ where $p_{ij}^{-} = \left(p_{ij2}, \ldots, p_{ijK}\right)$. It is possible to extend the model to support regression; see Appendix (Subsection 6.2). Note that model (5) regards the totals m_{ij} as fixed, but regards the day of week of each call as an independent multinomial (random) trial. We will specifically consider two scenarios:

1. Two studies with $I = 2$ and one mailing schedule used in each study.
2. Two studies with $I = 3$; one mailing schedule was used in the first study and two were used in the second. Here we assume that $i = 1$ corresponds to the first and $i \in \{2, 3\}$ corresponds to the second.

All studies under consideration consist of $J =$ five weeks of data, and we will compare them on a week-by-week basis. Under Scenario 1, we are interested in

$$g_j(\boldsymbol{\theta}) = \varepsilon(\boldsymbol{p}_{2j}) - \varepsilon(\boldsymbol{p}_{1j}), \quad j = 1, \ldots, J.$$

We will write $g(\boldsymbol{\theta})$ when referring to the entropy difference in a general sense, and $g_j(\boldsymbol{\theta})$ when a particular week is emphasized. For Scenario 2, let $\boldsymbol{q}_j = (q_{j1}, \ldots, q_{jK})$ be the overall day-of-week distribution for calls from the jth week of the second study. To combine data from the two mailing schedules, define random variables A and B as follows. For a given call, let $A \in \{1, \ldots, K\}$ indicate the day-of-week in which the call occurs and let $B \in \{2, 3\}$ indicate the mailing schedule to which the caller was assigned. Suppose $\text{P}(B = 2) = \pi_j$ and $\text{P}(B = 3) = 1 - \pi_j$. The overall probability of receiving a call on day of week $k \in \{1, \ldots, K\}$ during week j is

$$\begin{aligned} q_{jk} &= \text{P}(A = k) \\ &= \text{P}(A = k \mid B = 2)\text{P}(B = 2) + \text{P}(A = k \mid B = 3)\text{P}(B = 3) \\ &= \pi_j p_{2jk} + (1 - \pi_j) p_{3jk}. \end{aligned}$$

We may then write $\boldsymbol{q}_j = \pi_j \boldsymbol{p}_{2j} + (1 - \pi_j)\boldsymbol{p}_{3j}$, and our quantities of interest are

$$g_j(\boldsymbol{\theta}) = \varepsilon(\boldsymbol{q}_j) - \varepsilon(\boldsymbol{p}_{1j}), \; j = 1, \ldots, J.$$

The π_j are unknown and therefore will be replaced by estimates $\hat{\pi}_j = m_{2j}/(m_{2j} + m_{3j})$. Our analysis will be carried out conditionally on the m_{ij} for the sake of tractability; however, modeling the m_{ij} as observed quantities would likely express additional variability in the results and may be considered in future work.

Appendix (Subsection 6.1) describes a statistic Z based on an estimate $g(\hat{\boldsymbol{\theta}})$ of $g(\boldsymbol{\theta})$. Denoting α as the desired significance level for a test and z_α as the $1-\alpha$ quantile of the

standard normal distribution, we obtain familiar tests for $g(\boldsymbol{\theta})$: reject H_0 in hypothesis (2) if $|Z| > z_{\alpha/2}$, reject H_0 in hypothesis (3) if $Z > z_\alpha$, or reject H_0 in hypothesis (4) if $Z < z_\alpha$. We also obtain familiar confidence limits for $g(\boldsymbol{\theta})$ with level $1-\alpha$: the two-sided confidence interval $g(\hat{\theta}) \pm z_{\alpha/2} \cdot \widehat{\mathrm{SE}}$, the lower confidence limit $g(\hat{\theta}) - z_\alpha \cdot \widehat{\mathrm{SE}}$, and the upper confidence limit $g(\hat{\theta}) + z_\alpha \cdot \widehat{\mathrm{SE}}$. Properties of the test and confidence bounds based on Z are investigated by simulation in the supplementary material. The tables in the supplementary material give the estimated type I error probabilities and powers of the one-sided test for $g(\boldsymbol{\theta})$ based on Z, empirical coverage probabilities of the lower confidence limit for $g(\boldsymbol{\theta})$, and the empirical widths associated with the lower confidence limits. The numerical results correspond to a 5% nominal level for the test, and a 95% nominal level for the confidence limit. The numerical results show that the estimated type I error probabilities and the estimated coverage probabilities approach their respective nominal levels as the sample size becomes large. Furthermore, the power of the test increases as $g(\boldsymbol{\theta})$ approaches its maximum value. These results are to be anticipated.

It should be noted that in addition to the K-L divergence criterion, one can think of numerous other criteria that are appropriate for assessing the relative closeness of two sets of multinomial cell probabilities to the uniform probability vector; for example, the power divergence family of criteria investigated in Cressie and Read (1984). An extensive numerical investigation will be necessary before recommendations can be made on specific criteria for practical use. We have not undertaken such a simulation study since there is already some precedence in using the K-L divergence measure, as noted in Subsection 1.2. Furthermore, the focus of our work is on an application relevant to the U.S. Census Bureau, rather than the comparison of various criteria. Thus we are not making a definite recommendation in favor of tests based on the K-L divergence criterion, compared to other tests; for example, Anderson-Darling and other empirical distribution function based tests.

4. Data Analysis

Subsections 4.1, 4.2, and 4.3 compare the three possible pairs among the three NCBS studies. In each section, the given pair is compared using tests and confidence intervals based on the Z statistic from Section 3. Code for all procedures has been implemented in the R programming language (R Core Team 2022).

Although each dataset has been used twice among the three analyses, we have not adjusted results for multiple comparisons (e.g.,Westfall et al. 2011). It should however be noted that a number of p-values in our analysis are quite small and such an adjustment will not matter.

4.1. 2016 September NCBS versus 2017 March NCBS

Our first comparison is between the 2016 September NCBS and 2017 March NCBS call volumes, which falls into Scenario 2. Take $\boldsymbol{X}_{1j}$ to be the call frequencies observed (Sunday, Monday,. . ., Saturday) on the jth week of the 2016 September NCBS. Accordingly, $\boldsymbol{X}_{2j}$ and $\boldsymbol{X}_{3j}$ are call frequencies on the jth week of the 2017 March NCBS for the Monday Mailout and Thursday Mailout groups, respectively. Estimates for π_j are given in Table 2. We test hypothesis (3) for each week $j = 1,. . ., 5$, which can be written as follows.

[Test 1] H_0: "The day-of-week distribution in week j resulting from the 2016 September NCBS mailing schedule has larger or equal entropy for the call probabilities than the day-of-week distribution resulting from the 2017 March NCBS mailing schedule" versus H_1: "Not".

Therefore, rejection of H_0 for week j suggests that the 2017 March NCBS mailing strategy leads to call probabilities closer to the uniform probability vector during that week.

Table 3a gives results of testing this hypothesis for weeks 1–5. Recall that quantity $g(\boldsymbol{\theta})$ is bounded, so that $-1.946 \leq g(\boldsymbol{\theta}) \leq 1.946$ for any $\boldsymbol{\theta}$. The U.S. Census Bureau uses $\alpha = 0.10$ as its standard significance level for hypothesis testing. H_0 can be rejected for weeks 1–4, but there is insufficient evidence to reject during week 5. The Z-statistic is a rather large negative value in week 5, suggesting that there is evidence that the 2016 September NCBS call distribution had higher entropy during that time. Table 4a displays the estimated probabilities $\hat{p}_{1j}$ and $\hat{q}_j$ for weeks $j = 1, \ldots, 5$. We notice in week 1 that the 2017 March NCBS had a higher estimated entropy despite very low call probabilities on Monday and Tuesday; recall that it is very unlikely for Monday Mailout group respondents to receive the first mailing by this Monday or Tuesday, and impossible for the Thursday Mailout group. In week 5, large peaks are observed for the 2017 March NCBS on Monday

Table 3. Results for inference on the quantity g(θ). Estimates, standard errors (SEs), and Z-statistics are displayed in the first three columns. (a) and (b) give a p-value for hypothesis (4) and a level 0.90 lower confidence limit for g(θ). (c) gives a p-value for hypothesis (2) and a level 0.90 (two-sided) confidence interval for g(θ).

(a) 2016 September NCBS to 2017 March NCBS.

Week	Estimate	SE	Z-statistic	p-value	CI Lo
1	0.0823	0.0602	1.3681	0.0856	0.0052
2	0.2605	0.0402	6.4731	<0.0001	0.2090
3	0.1480	0.0411	3.6026	0.0002	0.0953
4	0.2273	0.0453	5.0166	<0.0001	0.1693
5	-0.3376	0.0775	-4.3563	1.0000	-0.4369

(b) 2016 June NCBS to 2017 March NCBS.

Week	Estimate	SE	Z-statistic	p-value	CI Lo
1	-0.0153	0.0631	-0.2426	0.5958	-0.0962
2	0.3463	0.0379	9.1467	<0.0001	0.2977
3	0.3905	0.0442	8.8356	<0.0001	0.3338
4	0.3523	0.0565	6.2409	<0.0001	0.2800
5	0.0253	0.0640	0.3956	0.3462	-0.0567

(c) 2016 June NCBS to 2016 September NCBS.

Week	Estimate	SE	Z-statistic	p-value	CI Lo	CI Hi
1	-0.0976	0.0451	-2.1630	0.0305	-0.1719	-0.0230
2	0.0857	0.0433	1.9782	0.0479	0.0144	0.1570
3	0.2425	0.0365	6.6499	<0.0001	0.1825	0.3025
4	0.1250	0.0607	2.0602	0.0394	0.0252	0.2247
5	0.3629	0.0535	6.7823	<0.0001	0.2749	0.4509

Table 4. Estimated probabilities for data analyses. Estimates are ordered from largest to smallest within each week. The corresponding day of week is shown to the right of each probability. The column labeled $\hat{\varepsilon}$ displays the entropy of the estimated probabilities.

(a) 2016 September NCBS vs. 2017 March NCBS.

Week	$\hat{p}$ (Week)							$\hat{\varepsilon}$
1	0.3857 Tue	0.3327 Mon	0.2388 Wed	0.0245 Sat	0.0122 Sun	0.0061 Thu	0.0000 Fri	1.2515
2	0.4224 Tue	0.3077 Wed	0.1190 Thu	0.0581 Sat	0.0421 Fri	0.0276 Sun	0.0232 Mon	1.4650
3	0.3835 Mon	0.2361 Tue	0.1247 Thu	0.1093 Wed	0.0670 Fri	0.0546 Sat	0.0247 Sun	1.6414
4	0.4545 Mon	0.2045 Tue	0.1212 Wed	0.1193 Thu	0.0473 Fri	0.0341 Sat	0.0189 Sun	1.5272
5	0.3256 Thu	0.1628 Mon	0.1550 Fri	0.1395 Tue	0.1085 Wed	0.0543 Sun	0.0543 Sat	1.7819
Week	$\hat{q}$ (Week)							$\hat{\varepsilon}$
1	0.3955 Fri	0.3284 Thu	0.1791 Sat	0.0522 Wed	0.0448 Sun	0.0000 Mon	0.0000 Tue	1.3338
2	0.2925 Mon	0.2054 Tue	0.1826 Wed	0.1535 Thu	0.1100 Fri	0.0311 Sat	0.0249 Sun	1.7255
3	0.3419 Thu	0.1581 Mon	0.1496 Fri	0.1026 Sat	0.0940 Tue	0.0940 Wed	0.0598 Sun	1.7894
4	0.2654 Thu	0.2099 Mon	0.1852 Fri	0.1636 Tue	0.1111 Wed	0.0340 Sat	0.0309 Sun	1.7545
5	0.4508 Mon	0.2798 Tue	0.1088 Wed	0.0622 Thu	0.0622 Fri	0.0259 Sat	0.0104 Sun	1.4443

(b) 2016 June NCBS vs. 2017 March NCBS.

Week	$\hat{p}$(Week)							$\hat{\varepsilon}$
1	0.4023 Thu	0.3456 Fri	0.1133 Sat	0.0793 Wed	0.0595 Sun	0.0000 Mon	0.0000 Tue	1.3492
2	0.4565 Mon	0.2798 Tue	0.1285 Wed	0.0763 Thu	0.0469 Fri	0.0067 Sat	0.0054 Sun	1.3793
3	0.4676 Mon	0.2444 Tue	0.1321 Wed	0.0819 Thu	0.0647 Fri	0.0053 Sun	0.0040 Sat	1.3989
4	0.5431 Fri	0.1802 Sat	0.0992 Thu	0.0653 Wed	0.0522 Sun	0.0470 Tue	0.0131 Mon	1.4022
5	0.4505 Mon	0.2344 Tue	0.1465 Wed	0.1136 Thu	0.0403 Fri	0.0147 Sat	0.0000 Sun	1.4190
Week	$\hat{p}$(Week)							$\hat{\varepsilon}$
1	0.3955 Fri	0.3284 Thu	0.1791 Sat	0.0522 Wed	0.0448 Sun	0.0000 Mon	0.0000 Tue	1.3338
2	0.2925 Mon	0.2054 Tue	0.1826 Wed	0.1535 Thu	0.1100 Fri	0.0311 Sat	0.0249 Sun	1.7255
3	0.3419 Thu	0.1581 Mon	0.1496 Fri	0.1026 Sat	0.0940 Tue	0.0940 Wed	0.0598 Sun	1.7894
4	0.2654 Thu	0.2099 Mon	0.1852 Fri	0.1636 Tue	0.1111 Wed	0.0340 Sat	0.0309 Sun	1.7545
5	0.4508 Mon	0.2798 Tue	0.1088 Wed	0.0622 Thu	0.0622 Fri	0.0259 Sat	0.0104 Sun	1.4443

(c) 2016 June NCBS vs. 2016 September NCBS.

Week	$\hat{p}$(Week)							$\hat{\varepsilon}$
1	0.4023 Thu	0.3456 Fri	0.1133 Sat	0.0793 Wed	0.0595 Sun	0.0000 Mon	0.0000 Tue	1.3492
2	0.4565 Mon	0.2798 Tue	0.1285 Wed	0.0763 Thu	0.0469 Fri	0.0067 Sat	0.0054 Sun	1.3793
3	0.4676 Mon	0.2444 Tue	0.1321 Wed	0.0819 Thu	0.0647 Fri	0.0053 Sun	0.0040 Sat	1.3989
4	0.5431 Fri	0.1802 Sat	0.0992 Thu	0.0653 Wed	0.0522 Sun	0.0470 Tue	0.0131 Mon	1.4022
5	0.4505 Mon	0.2344 Tue	0.1465 Wed	0.1136 Thu	0.0403 Fri	0.0147 Sat	0.0000 Sun	1.4190
Week	$\hat{q}$ (Week)							$\hat{\varepsilon}$
1	0.3857 Tue	0.3327 Mon	0.2388 Wed	0.0245 Sat	0.0122 Sun	0.0061 Thu	0.0000 Fri	1.2515
2	0.4224 Tue	0.3077 Wed	0.1190 Thu	0.0581 Sat	0.0421 Fri	0.0276 Sun	0.0232 Mon	1.4650
3	0.3835 Mon	0.2361 Tue	0.1247 Thu	0.1093 Wed	0.0670 Fri	0.0546 Sat	0.0247 Sun	1.6414
4	0.4545 Mon	0.2045 Tue	0.1212 Wed	0.1193 Thu	0.0473 Fri	0.0341 Sat	0.0189 Sun	1.5272
5	0.3256 Thu	0.1628 Mon	0.1550 Fri	0.1395 Tue	0.1085 Wed	0.0543 Sun	0.0543 Sat	1.7819

and Tuesday. Many calls occurring on these peak days are from the Thursday Mailout group, as seen in Figure 1a, whose final mailing was initiated the previous Thursday (March 30). However, calls are also contributed from the Monday Mailout group, whose final mailing was the previous Monday (March 27).

4.2. 2016 June NCBS versus 2017 March NCBS

Our second comparison is between the 2016 June NCBS versus 2017 March NCBS call volumes, which falls into Scenario 2. Take $\boldsymbol{X}_{1j}$ to be the frequencies of the 2016 June

NCBS calls observed on (Sunday, Monday,. . ., Saturday). Take $\boldsymbol{X}_{2j}$, $\boldsymbol{X}_{3j}$, and π_j to be defined as in Subsection 4.1. We test hypothesis (3) for each week $j = 1, \ldots, 5$, which can be written as follows.

> [Test 2] H_0: "The day-of-week distribution in week j resulting from the 2016 June NCBS mailing schedule has larger or equal entropy for the call probabilities than the day-of-week distribution resulting from the 2017 March NCBS mailing schedule" versus H_1: "Not".

Here, rejection of H_0 for week j suggests that the 2017 March NCBS mailing strategy leads to call probabilities closer to the uniform probability vector during that week.

Table 3b gives results of testing this hypothesis for weeks 1–5. Table 4b displays the estimated probabilities $\hat{p}_{1j}$ and $\hat{q}_j$ for weeks $j = 1, \ldots, 5$. There is strong evidence to reject H_0 for weeks 2, 3, and 4, but insufficient evidence to reject during weeks 1 and 5 at significance level $\alpha = 0.10$. Recall that the Thursday Mailout group had little opportunity to call during week 1, so the results for this week are primarily based on the Monday Mailout group.

4.3. 2016 June NCBS versus 2016 September NCBS

Our third comparison is between the 2016 September NCBS and 2016 June NCBS call volumes, which matches Scenario 1. Take $\boldsymbol{X}_{1j}$ to be the frequencies of 2016 June NCBS as in Subsection 4.2 and $\boldsymbol{X}_{2j}$ to be the frequencies of 2016 September NCBS calls as in Subsection 4.2. Because both studies used a single mailing strategy for all respondents, we assume a null hypothesis that neither strategy leads to a significantly more uniform call distribution. Therefore, we test hypothesis (2) for each week $j = 1, \ldots, 5$, which can be written as follows.

> [Test 3] H_0: "The day-of-week distribution in week j resulting from the 2016 June NCBS mailing schedule and that resulting from the 2016 September NCBS mailing schedule have equal entropies for the call probabilities" versus H_1: "Not".

Here, rejection of H_0 for week j suggests that the two mailing strategies do not lead to call probabilities that are equally close to the uniform probability vector during that week.

Table 3c gives results of testing this hypothesis for weeks 1–5, and Table 4c displays the estimated probabilities $\hat{p}_{1j}$ and $\hat{p}_{2j}$ for weeks $j = 1, \ldots, 5$. The test can be rejected at significance level $\alpha = 0.10$ for all five weeks, although the evidence is much stronger in weeks 3 and 5. The Z-statistics are positive for weeks 2–5, indicating a larger entropy for the 2016 September NCBS except during week 1.

5. Discussion and Conclusions

In this work, we analyzed call volumes observed in several census studies to determine whether one mailing strategy produced more uniform call probabilities than another. To do this, we made use of statistical tests and confidence intervals determined from basic large sample theory. Our analysis found that a staggered strategy–as used in the 2017 March NCBS–yielded a significantly higher entropy than the two unstaggered strategies toward the middle of the study period, after both the Monday and Thursday Mailout groups

received the first mailing. However, the two unstaggered strategies–the 2016 September NCBS and 2016 June NCBS–also yielded significantly different entropies when compared to each other; this demonstrates that other aspects of mailing schedule design, aside from staggering, affect uniformity of calls from week to week. After the final mailing is sent, the choice of mailing schedule is expected to have a diminishing effect on call uniformity as the overall volume of calls diminishes. One way to attenuate starting and ending differences among the operations would be, say, to combine weeks 1 and 2 into a "beginning period", label week 3 as a "middle period", and combine weeks 4+ into an "ending period"; the methodology could be applied to the three periods instead of the individual weeks without any changes.

Although some evidence was found in the data that staggering increases call uniformity, the three featured studies were carried out separately and may have been subject to subtle variations other than mailing strategy. A controlled experiment would help to isolate this effect from other factors. Procedures such as the ones presented in this article could help to guide design and sample size determination for such an experiment.

The report by Nichols et al. (2019) represents the analysis-of-record based on the methodology in the present article. Marasteanu (2019) describes a staggered mailing strategy for the 2020 decennial census which makes use of findings by Nichols et al. (2019) and others, while also taking into account a number of important practical concerns. An unspecified model is used to express volumes to call centers and to illustrate the effects of staggered mailings on them over time.

While our model was based on independent multinomial observations, we could consider a regression model with appropriate covariates as described in Appendix (Subsection 6.2). Here, it becomes necessary to check model adequacy–e.g., via goodness-of-fit testing–before proceeding with inference on $g(\boldsymbol{\theta})$. However, were a sufficiently good predictive model available, it could be used to optimize over a class of mailing strategies and identify which one(s) achieved an optimal uniformity. This could be an objective in future analyses of U.S. Census Bureau operations, as a step beyond statistical inference.

6. Appendix

6.1. Inference on Entropy Difference

In order to discuss statistical procedures, let us generally write

$$g_j(\boldsymbol{\theta}) = \varepsilon(c_1\boldsymbol{p}_{1j} + \cdots + c_I\boldsymbol{p}_{Ij}) - \varepsilon(d_1\boldsymbol{p}_{1j} + \cdots + d_I\boldsymbol{p}_{Ij}), \quad j = 1, \ldots, J, \tag{6}$$

for given coefficients $\boldsymbol{c} = (c_1, \ldots, c_I)$ and $\boldsymbol{d} = (d_1, \ldots, d_I)$ which are distributions on $\{1, \ldots, I\}$. We do not encounter a situation where two studies use data from a common mailing schedule; therefore, we will have $c_i d_i = 0$ for $i = 1, \ldots, I$. In a multinomial analysis, one of our day-of-week categories is redundant because $\sum_{k=1}^{K} X_{ijk} = m_{ij}$ and $\sum_{k=1}^{K} p_{ijk} = 1$. Without loss of generality, we will consider the first category as the redundant one, and write $X_{ij}^{-} = (X_{ij2}, \ldots, X_{ijK})$ and $p_{ij}^{-} = (p_{ij2}, \ldots, p_{ijK})$. Denote $p^{-} = \left(p_{11}^{-}, \ldots, p_{1J}^{-}, \ldots, p_{I1}^{-}, \ldots, p_{IJ}^{-}\right)$, and its maximum likelihood estimator as $\hat{p}^{-}$, which replaces each p_{ij}^{-} with $\hat{p}_{ij}^{-} = X_{ij}^{-}/m_{ij}$. Under model (5), the unknown $\boldsymbol{\theta}$ is taken

assumed to be $\boldsymbol{p}^-$. Write $\boldsymbol{T} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote that random variable $\boldsymbol{T} \in \mathbb{R}^n$ follows a normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and variance $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. As discussed in Lehmann (2004, 314), we have a large sample normal distribution $\hat{p}^- \sim \mathrm{N}(\boldsymbol{p}^-, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \mathrm{Blockdiag}\left(m_{11}^{-1}\left[\mathrm{Diag}\left(p_{11}^-\right) - p_{11}^- p_{11}^{-\top}\right], \ldots, m_{IJ}^{-1}\left[\mathrm{Diag}\left(p_{IJ}^-\right) - p_{IJ}^- p_{1J}^{-\top}\right]\right) \quad (7)$$

is an $IJ(K-1) \times IJ(K-1)$ covariance matrix. Furthermore, the delta method (Lehmann 2004, 315) gives the large sample distribution

$$g_j\left(\hat{p}^-\right) \sim \mathrm{N}\left(g_j\left(p^-\right), \sigma^2_{g_j(p-)}\right), \quad \sigma^2_{g_j(p-)} = \left[\frac{\partial g_j(p^-)}{\partial p^-}\right] \Sigma \left[\frac{\partial g_j(p^-)}{\partial p^-}\right]^\top,$$

for $j = 1, \ldots, J$. After some algebra, we obtain the $1 \times IJ(K-1)$ gradient

$$\frac{\partial g_j(p^-)}{\partial p^-} = \begin{pmatrix} e_j \otimes \left[c_1 \nabla \varepsilon\left(\Sigma_{i=1}^I c_i p_{ij}^-\right) - d_1 \nabla \varepsilon\left(\Sigma_{i=1}^I d_i p_{ij}^-\right)\right] \\ e_j \otimes \left[c_I \nabla \varepsilon\left(\Sigma_{i=1}^I c_i p_{ij}^-\right) - d_1 \nabla \varepsilon\left(\Sigma_{i=1}^I d_i p_{ij}^-\right)\right] \end{pmatrix}^\top, \quad (8)$$

where $\nabla \varepsilon(\boldsymbol{x}) = (-\log(x_2/x_1), \ldots, -\log(x_K/x_1))^\top$, $\boldsymbol{e}_j$ is the jth column of a $J \times J$ identity matrix, and $\otimes$ denotes the matrix Kronecker product.

Let $\hat{p}^-$ be a maximum likelihood estimator of $\boldsymbol{p}^-$ and define the statistic $Z = g_j(\hat{p}^-)/\sigma_{g_j(\hat{p}^-)}$ with the expression for $\sigma^2_{g_j(p^-)}$. Now, $Z \sim \mathrm{N}(0,1)$ approximately under the restriction $g_j(\boldsymbol{p}^-) = 0$, which occurs at the boundary of the null hypotheses (2), (3), and (4). Therefore, we obtain the test and confidence interval procedures discussed at the end of Section 3 with $\widehat{\mathrm{SE}} = \sigma_{g(\hat{p}^-)}$.

Some insight into the behavior of $g(\boldsymbol{p}^-)$ can be seen from its gradient. Consider Scenario S1 with $J = 1$ and suppress the j index; we have

$$\begin{aligned} \frac{\partial g\left(p^-\right)}{\partial p^-} &= \begin{pmatrix} -\nabla \varepsilon\left(p_1^-\right)^\top & \nabla \varepsilon\left(p_2^-\right)^\top \end{pmatrix} \\ &= \log(p_{12}/p_{11}) \quad \cdots \quad \log(p_{1K}/p_{11}) \quad -\log(p_{22}/p_{21}) \quad \cdots \quad -\log(p_{2K}/p_{21}). \end{aligned}$$

Because the gradient separates into a component involving only $\boldsymbol{p}_1$ and similar one involving only $\boldsymbol{p}_2$, it suffices to comment only on the former. When $\boldsymbol{p}_1 \approx (1/K, \ldots, 1/K)$, it is seen that $-\nabla \varepsilon(\boldsymbol{p}_1^-) \approx \boldsymbol{0}$; therefore, $g(\boldsymbol{p}^-)$ increases very slowly to its maximum of log K as $\boldsymbol{p}_1$ approaches a discrete uniform distribution. On the other hand, when $\boldsymbol{p}_1 \to \boldsymbol{e}_2$, then $-\nabla \varepsilon(\boldsymbol{p}_1^-) \to (0, \infty, \ldots, \infty)$; therefore, when $\boldsymbol{p}_1$ is close to a point mass, small changes in $\boldsymbol{p}_1$ result in very large changes in some of the components of $g(\boldsymbol{p}^-)$.

This remark notes that changes in the quantity $g(\boldsymbol{p}^-)$ are smaller when the component distributions are closer to uniform. Differences in this setting may therefore be difficult to detect with the Z-statistic, which may warrant consideration of alternative distance measures in future work. Alternative distance measures might also be considered if one is thought to better express the cost of departure from uniformity than K-L divergence.

6.2. Regression Model

The framework can be extended to a more general multinomial regression model for additional flexibility. Suppose we wish to compare two mailing strategies based on I mailing schedules where calls are recorded in J weeks. Let $\boldsymbol{Y}_{ij} = (Y_{i11}, \ldots, Y_{ijK})$ be a vector of call counts received on the kth day in the jth week from the ith mailing schedule, for $i = 1, \ldots, I, j = 1, \ldots, J$, and $k = 1, \ldots, K$, with $m_{ij} = \sum_{k=1}^{K} Y_{ijk}$ total calls received. Let $w_{ijk} \in \mathbb{R}^D$ be a corresponding fixed covariate. The discrete choice model described by Agresti (2013, Sec. 8.5),

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Mult}_K\left(m_{ij}, p_{ij}\right), \quad p_{ijk} = \frac{\exp\left\{w_{ijk}^{\top}\beta\right\}}{\Sigma_{\ell=1}^{k} \exp\left\{w_{ij\ell}^{\top}\beta\right\}}, \tag{9}$$

is a form of multinomial regression which permits the covariate to vary by category. The independent variable w_{ijk} may contain indicators such as day of the week, week of the mailing schedule, holidays and other features of interest which vary by day. In this model, $\log\left(p_{ijk}/p_{ij\ell}\right) = \left(w_{ijk} - w_{ij\ell}\right)^{\top}\beta$ for any $k,\ell \in \{1, \ldots, K\}$. Regression model (9) allows a number of alternative formations; for example, rather than partitioning the data into J weeks (or other non-overlapping time periods) as desired in our application, one may also consider taking $J = 1$, the K multinomial categories to correspond to all days relevant to the analysis, and covariates w_{ijk} to include adjustments for day-of-week and/or day-of-study effects.

Under model (9), the unknown parameter $\boldsymbol{\theta}$ is taken to be the coefficient vector $\boldsymbol{\beta}$. To perform inference on (9) in the regression case, the delta method applied to the maximum likelihood estimator $\hat{\beta}$ yields large sample distribution

$$g\left(\boldsymbol{p}^{-}(\hat{\beta})\right) \sim \text{N}\left(g\left(p^{-}(\beta)\right), \sigma^2_{g\left(p^{-}(\beta)\right)}\right), \quad \sigma^2_{g\left(p^{-}(\beta)\right)} = \left[\frac{\partial g\left(p^{-}(\beta)\right)}{\partial \beta}\right] \mathcal{I}^{-1}(\beta) \left[\frac{\partial g\left(p^{-}(\beta)\right)}{\partial \beta}\right]^{\top}$$

where $\mathcal{I}(\beta)$ is the Fisher information matrix with respect to β under the outcomes $\boldsymbol{Y}_{ij}$. An estimate of $\mathcal{I}^{-1}(\beta)$ based on $\hat{\beta}$ may be produced by statistical software such as the `mlogit` package (Croissant 2020) in R, or may be computed directly via $\mathcal{I}(\beta) = \left[\frac{\partial p^{-}(\beta)}{\partial \beta}\right]^{\top} \sum^{-1} \left[\frac{\partial p^{-}(\beta)}{\partial \beta}\right]$ using $\boldsymbol{\Sigma}$ given in Equation (7). To compute the remaining terms of $\sigma^2_{g\left(p^{-}(\beta)\right)}$, we have

$$\frac{\partial g\left(p^{-}(\beta)\right)}{\partial \beta} = \frac{\partial g\left(p^{-}\right)}{\partial p^{-}} \frac{\partial p^{-}(\beta)}{\partial \beta}$$

with $\frac{\partial g\left(p^{-}\right)}{\partial p^{-}}$ given by Equation (8) and

$$\frac{\partial p_{ijk}(\beta)}{\partial \beta_d} = \left\{\sum_{\ell=1}^{K} \exp\left\{\left(w_{ij\ell} - w_{ijk}\right)^{\top}\beta\right\}\right\}^{-2} \sum_{\ell=1}^{K} \exp\left\{\left(w_{ij\ell} - w_{ijk}\right)^{\top}\beta\right\}\left(w_{ij\ell d} - w_{ijkd}\right).$$

7. References

Agresti, A. 2013. *Categorical Data Analysis*. Hoboken, NJ: Wiley, 3rd edition.

Blei, D.M., A. Kucukelbir, and J.D. McAuliffe 2017. "Variational Inference: A Review for Statisticians". *Journal of the American Statistical Association* 112: 859–877. DOI: https://dx.doi.org/10.1080/01621459.2017.1285773.

Chesnut, J. 2003. "Telephone Questionnaire Assistance". Census 2000 Evaluation A.1.a, U.S. Census Bureau. Available at: https://www.census.gov/pred/www/rpts/A.1.a.pdf (accessed June 2022).

Cohen, A., J. Kolassa, and H. Sackrowitz. 2006. "A Test for Equality of Multinomial Distributions vs Increasing Convex Order". In *Recent Developments in Nonparametric Inference and Probability*, edited by J. Sun, A. DasGupta, V. Melfi, and C. Page. *Lecture Notes–Monograph Series*, Volume 50: 156–163. Institute of Mathematical Statistics, Beachwood, OH. DOI: https://dx.doi.org/10.1214/074921706000000662.

Coombs, J. 2017. *Analysis Report for the Small-Scale Mailout Testing Program June 2016 Test on the Placement and Length of the User ID for an Online Census Bureau Survey*. Census Program Internal Memorandum Series 2017.2.i, U.S. Census Bureau.

Cover, T.M., and J.A. Thomas. 2006. *Elements of Information Theory*. Hoboken, NJ: Wiley, 2nd edition.

Cressie, N., and T.R. Read. 1984. "Multinomial Goodness-of-Fit Tests". *Journal of the Royal Statistical Society: Series B* 46: 440–464. DOI: https://dx.doi.org/10.1111/j.2517-6161.1984.tb01318.x.

Croissant, Y. 2020. "Estimation of Random Utility Models in R: The mlogit Package". *Journal of Statistical Software* 95: 1–41. DOI: https://dx.doi.org/10.18637/jss.v095.i11.

Dorfinger, P., G. Panholzer, and W. John. 2011. "Entropy Estimation for Real-Time Encrypted Traffic Identification (Short Paper)". In *Traffic Monitoring and Analysis*, edited by J. Domingo-Pascual, Y. Shavitt, and S. Uhlig: 164–171. Berlin, Heidelberg. Springer. DOI: https://dx.doi.org/10.1007/978-3-642-20305-3_14.

Eggleston, C., and J. Coombs. 2018. *Effect of Data Use Statements and Postcard Format on Login Rates for a Mandatory Online Census Bureau Survey*. U.S. Census Program Internal Memorandum Series 2018.11.i, U.S. Census Bureau.

Girardin, V., and J. Lequesne. 2019. "Entropy-Based Goodness-of-Fit Tests–A Unifying Framework: Application to DNA Replication". *Communications in Statistics–Theory and Methods* 48: 62–74. DOI: https://dx.doi.org/10.1080/03610926.2017.1401084.

Konishi, S., and G. Kitagawa. 2008. *Information Criteria and Statistical Modeling*. New York, NY: Springer.

Lehmann, E.L. 2004. *Elements of Large-Sample Theory*. New York, NY: Springer.

Li, Y., R.S. Singh, and Y. Sun. 2005. "Goodness-of-Fit Tests of a Parametric Density Functions: Monte Carlo Simulation Studies". *Journal of Statistical Research* 39: 103–125. Available at: http://jsr.isrt.ac.bd/article/goodness-fit-tests-parametric-density-functions-monte-carlo-simulation-studies/.

Liu, X., and J. Wang. 2004. "Testing the Equality of Multinomial Populations Ordered by Increasing Convexity Under the Alternative". *Canadian Journal of Statistics* 32: 159–168. DOI: https://dx.doi.org/10.2307/3315939.

Marasteanu, I. 2019. "The Research and Methodology on Staggering the 2020 Census Mailings". In JSM Proceedings, Government Statistics Section: 2457–2467. Alexandria, VA: American Statistical Association.

Nichols, E., S. Konya, R. Horwitz, and A. Raim. 2019. *2020 Census Research and Testing Report: The Effect of the Mail Delivery Date on Survey Login Rates and Helpline Call Rates*. U.S. Census Bureau, Research and Methodology Directorate, Center for Behavioral Science Methods Research Report Series (Survey Methodology): 2019–01,

U.S. Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2019/adrm/rsm2019-01.pdf (accessed June 2022).

Ormerod, J.T., and M.P. Wand. 2010. "Explaining Variational Approximations". *The American Statistician* 64: 140–153. DOI: https://dx.doi.org/10.1198/tast.2010.09058.

Paninski, L. 2008. "A Coincidence-Based Test for Uniformity Given Very Sparsely Sampled Discrete Data". *IEEE Transactions on Information Theory* 54: 4750–4755. DOI: https://dx.doi.org/10.1109/TIT.2008.928987.

Pardo, L. 2006. *Statistical Inference Based on Divergence Measures*. Boca Raton, FL: Chapman and Hall/CRC.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Song, K.S. 2002. "Goodness-of-Fit Tests Based on Kullback-Leibler Discrimination Information". *IEEE Transactions on Information Theory* 48: 1103–1117. DOI: https://dx.doi.org/10.1109/18.995548.

U.S. Census Bureau. 2017. *2020 Census Operational Plan*. Available at: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/planning-docs/operational-plan.html. Version 3.0. (accessed June 2022).

Vuong, Q.H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses". *Econometrica* 57: 307–333. DOI: https://dx.doi.org/10.2307/1912557.

Westfall, P.H., R.D. Tobias, and R.D. Wolfinger. 2011. *Multiple Comparisons and Multiple Tests Using SAS*. Cary, NC: SAS Institute, Inc., 2nd edition.

Zajac, K. 2012. *2010 Census Telephone Questionnaire Assistance Assessment Report*. 2010 Census Planning Memoranda Series No. 231, U.S. Census Bureau. Available at: https://www.census.gov/2010census/pdf/2010_Census_TQA_Assessment.pdf (accessed June 2022).

Received June 2021
Revised March 2022
Accepted July 2022