RESEARCH REPORT SERIES
(Statistics #2018-09)

A Statistical Comparison of Call Volume Uniformity Due to Mailing Strategy

Andrew M. Raim
Elizabeth Nichols
Thomas Mathew [1]

[1] Department of Mathematics and Statistics,
University of Maryland, Baltimore County

Center for Statistical Research and Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: September 13, 2018

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

A Statistical Comparison of Call Volume Uniformity Due to Mailing Strategy

Andrew M. Raim[a,*], Elizabeth Nichols[b], and Thomas Mathew[a,c]

[a]Center for Statistical Research and Methodology, U.S. Census Bureau

[b]Center for Survey Measurement, U.S. Census Bureau

[c]Department of Mathematics and Statistics,
University of Maryland, Baltimore County

Abstract

The U.S. Census Bureau conducts a variety of experiments to prepare for a full-scale decennial census and investigate possible improvements to operations. The Census Bureau is investigating possible strategies for sending mail to motivate recipients to self-respond. It is suspected that the choice of mailing strategy affects the distribution of call volumes to the Census Bureau's telephone helplines. For staffing purposes, more uniform call volumes throughout the week are desirable. In this work, we consider formal statistical methods to compare call volumes resulting from several recent experiments, and determine whether one mailing strategy yields a more uniform call distribution than others.

1 Introduction

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 it has not been marketed as a response option. For the 2020 Census, the Census Bureau will mail letters and postcards to each household in the country to request participation in the census, and will encourage responses on a large scale through the internet (U.S. Census Bureau, 2017). Telephone helplines will be highlighted in mailings, both as a means to assist with internet response and to serve as a mode of response themselves. From the perspective of the 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 operation. One reason to

Disclaimer: This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

[*]Contact: Andrew M. Raim (andrew.raim@census.gov), Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, 20233, U.S.A.

prefer a more uniform distribution is that fewer helpline staff could potentially be hired and be given a more constant workload over the course of the operation.

The schedule of mailings influences when calls tend to occur and is an aspect of census design which 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, 6 million calls in the 2000 Census, and 4.5 million calls in the 2010 Census. The increased emphasis on data collection by internet and phone suggests that an increased volume of calls could be expected in the 2020 Census. Although volumes of calls have changed over the decades, patterns of calls to the helpline have not changed drastically (Nichols et al., 2018). 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 these patterns to continue with an increased volume for the 2020 Census, the Census Bureau is considering plans to stagger mailing of census notifications so that they are delivered on different days of the week, anticipating that calls to the helpline will be more uniform throughout the week and thus easier to staff efficiently (Nichols et al., 2018). Note that other aspects of a mailing schedule—such as potential impact to response rates—are important to the Census Bureau as well; however, the remainder of this report focuses on call uniformity.

The Census Bureau conducts experiments throughout the decade to prepare for the decennial census. Several experiments for the 2020 Census have sent mailings to invite an internet response, and have recorded instances of subsequent calls to the helplines. These experiments employed various mailing schedules, providing an opportunity to compare strategies and see if any have led to call volumes which are statistically closer to a uniform distribution.

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; examples include chi-square tests for equality of proportions and Kolmogorov-Smirnov tests for equality of continuous distributions. However, the equality of two distributions is not the primary interest in our application. We consider use of Kullback-Leibler (K-L) divergence, which is seen to be equivalent to comparing the entropy of one distribution to the other. Procedures to test statistical hypotheses and compute related confidence intervals are presented, making use of basic results from large sample theory. These procedures are applied to call volume data from three census experiments.

Cover and Thomas (2006) introduces K-L distance, entropy, and related concepts, and discusses fundamental applications in information theory. 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, 2017). 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. 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.

The rest of the paper proceeds as follows. Section 2 discusses testing and confidence interval procedures to compare the closeness of two discrete distributions to uniformity. Section 3 presents basic simulation studies to validate the methods. Section 4 introduces the call volume data and gives results of the data analysis. Finally, Section 5 concludes the paper. Tables and figures are given at the end of the paper. This paper is a companion to Nichols et al. (2018), which presents an analysis of a recent census experiment that includes our findings.

2 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, $(1, \ldots, k)$ represent days of the week $(\mathrm{Sun}, \mathrm{Mon}, \ldots, \mathrm{Sat})$ with $k = 7$, and $\boldsymbol{p}$ and $\boldsymbol{q}$ are probabilities of a census participant 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 $D(\boldsymbol{p}, \boldsymbol{q}) = \sum_{j=1}^{k} p_j \log(p_j/q_j)$ denote the Kullback-Leibler (K-L) divergence, which is often used to measure the distance between two probability distributions. Also, let $\bar{\boldsymbol{e}} = (1/k \ldots, 1/k)$ denote the probabilities for the discrete uniform distribution. We will say that $\boldsymbol{q}$ is a “more uniform” distribution than $\boldsymbol{p}$ if $\boldsymbol{q}$ is closer to the discrete uniform distribution than $\boldsymbol{p}$; in other words, if

$$\begin{aligned}
&D(\boldsymbol{p}, \bar{\boldsymbol{e}}) > D(\boldsymbol{q}, \bar{\boldsymbol{e}}) \\
&\quad \iff \left[\sum_{j=1}^{k} p_j \log p_j - \sum_{j=1}^{k} p_j \log(1/k)\right] > \left[\sum_{j=1}^{k} q_j \log q_j - \sum_{j=1}^{k} q_j \log(1/k)\right] \\
&\quad \iff \sum_{j=1}^{k} p_j \log p_j > \sum_{j=1}^{k} q_j \log q_j \\
&\quad \iff \mathcal{E}(\boldsymbol{p}) < \mathcal{E}(\boldsymbol{q}). \qquad (2.1)
\end{aligned}$$

Here, $\mathcal{E}(\boldsymbol{p}) = -\sum_{j=1}^{k} p_j \log p_j$ is the entropy of the distribution with probabilities $\boldsymbol{p}$. Let $\boldsymbol{e}_j$ denote the jth column of the $k \times k$ identity matrix. For any distribution $\boldsymbol{p}$ on $(1, \ldots, k)$, it is well known that

$$\begin{aligned}
&\mathcal{E}(\boldsymbol{p}) \leq \mathcal{E}(\bar{\boldsymbol{e}}) = \log k, \\
&\mathcal{E}(\boldsymbol{p}) \geq \mathcal{E}(\boldsymbol{e}_j) = 0, \quad \text{for any } j = 1, \ldots, k;
\end{aligned}$$

so that entropy is minimized by a point mass and maximized by the discrete uniform distribution. Suppose $\boldsymbol{p}$ and $\boldsymbol{q}$ are parameterized by $\boldsymbol{\theta}$ which depends on the choice of model, to be discussed later in this section. Let $g(\boldsymbol{\theta}) = \mathcal{E}(\boldsymbol{q}) - \mathcal{E}(\boldsymbol{p})$ represent the difference in entropy. Motivated by (2.1), we will consider testing hypotheses of the form

$$H_0 : g(\boldsymbol{\theta}) = 0 \quad \text{vs.} \quad H_1 : g(\boldsymbol{\theta}) \neq 0, \tag{2.2}$$

$$H_0 : g(\boldsymbol{\theta}) \leq 0 \quad \text{vs.} \quad H_1 : g(\boldsymbol{\theta}) > 0, \tag{2.3}$$

$$H_0 : g(\boldsymbol{\theta}) \geq 0 \quad \text{vs.} \quad H_1 : g(\boldsymbol{\theta}) < 0. \tag{2.4}$$

If H_0 is rejected in (2.3), for example, we conclude that $\boldsymbol{q}$ is a distribution with higher entropy, or equivalently that $\boldsymbol{q}$ is closer to the discrete uniform distribution. Note that these hypotheses are invariant to the order of elements in both $\boldsymbol{p}$ and $\boldsymbol{q}$; this is desirable for our call volume application 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 $g(\boldsymbol{\theta})$.

Remark 2.1. We make note of several points before proceeding.

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

b. Let $\mathcal{E}_a(\cdot)$ denote the entropy function where logarithms are taken under base a. Here, $\mathcal{E}_a(\boldsymbol{q}) - \mathcal{E}_a(\boldsymbol{p}) = [\log a]^{-1} g(\boldsymbol{\theta})$, so that the change of base only serves to scale our quantity of interest by a constant. Then, without loss of generality, we will consider natural logarithms for the remainder of the paper.

c. It is possible to compare the entropy of two discrete distributions with different numbers of support points. If $\boldsymbol{p} = (p_1, \ldots, p_{k_1})$, $\boldsymbol{q} = (q_1, \ldots, q_{k_2})$, and $\boldsymbol{e}_k = (1/k, \ldots, 1/k)$, we obtain the analog to (2.1) that $D(\boldsymbol{p}, \bar{\boldsymbol{e}}_{k_1}) > D(\boldsymbol{q}, \bar{\boldsymbol{e}}_{k_2}) \iff \mathcal{E}(\boldsymbol{p}) - \log k_1 < \mathcal{E}(\boldsymbol{q}) - \log k_2$.

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

$$\begin{aligned} \mathrm{P}(\boldsymbol{X} = \boldsymbol{x}) &= \frac{m!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, \quad \text{where} \\ \boldsymbol{x} &\in \Big\{ (z_1, \ldots, z_k) : z_j \in \{0, 1, \ldots, m\}, z_1 + \cdots + z_k = m \Big\}. \end{aligned}$$

Consider the comparison of two census experiments where a total of I mailing schedules were attempted among the two experiments. For the ith mailing schedule, let J_i denote the total number of weeks of the experiment. In our application, 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

$$\boldsymbol{X}_{ij} \overset{\text{ind}}{\sim} \text{Mult}_k(m_{ij}, \boldsymbol{p}_{ij}), \tag{2.5}$$

where $\boldsymbol{p}_{ij} = (p_{ij1}, \ldots, p_{ijk})$ is the day-of-week distribution and $m_{ij} = \sum_{\ell=1}^{k} X_{ij\ell}$ is the total call count on the jth week for the ith mailing schedule. Note that model (2.5) regards total call counts for week j of mailing schedule i as fixed, but the day of week in which calls occur as independent multinomial (random) trials. We will specifically consider two scenarios:

S1. Two census experiments with $I = 2$ and one mailing schedule used in each experiment.

S2. Two census experiments with $I = 3$; one mailing schedule was used in the first experiment and two were used in the second experiment. Here we assume that $i = 1$ corresponds to the first experiment and $i \in \{2, 3\}$ corresponds to the second.

All experiments under consideration consist of $J = 5$ weeks of data, and we will compare experiments on a week-by-week basis. Under Scenario S1, we are interested in

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

For Scenario S2, 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 experiment. To combine data from the two mailing schedules, let π_j be the probability of receiving a call from a respondent in the first mailing schedule, so that $1 - \pi_j$ is the probability of receiving a call from a respondent in the second mailing schedule. By the law of total probability,

$$\begin{aligned} q_{j\ell} &= \mathrm{P}\{\text{Call occurs on day-of-week } \ell\} \\ &= \mathrm{P}\{\text{Call occurs on day-of-week } \ell \mid \text{Caller is from Mailing Schedule 1}\} \times \\ &\quad \mathrm{P}\{\text{Caller is from Mailing Schedule 1}\} + \\ &\quad \mathrm{P}\{\text{Call occurs on day-of-week } \ell \mid \text{Caller is from Mailing Schedule 2}\} \times \\ &\quad \mathrm{P}\{\text{Caller is from Mailing Schedule 2}\} \\ &= \pi_j p_{2j\ell} + (1 - \pi_j) p_{3j\ell}. \end{aligned}$$

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

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

The π_j are unknown and therefore will be replaced by an estimate $\hat{\pi}_j = m_{2j}/(m_{2j} + m_{3j})$. 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 therefore may be interesting to consider.

In order to discuss statistical procedures, let us generally write

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

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 census experiments 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 $\sum_{\ell=1}^{k} X_{ij\ell} = m_{ij}$ and $\sum_{\ell=1}^{k} p_{ij\ell} = 1$. Without loss of generality, we will consider the first category as the redundant one, and write $\boldsymbol{X}_{ij}^{-} = (X_{ij2}, \ldots, X_{ijk})$ and $\boldsymbol{p}_{ij}^{-} = (p_{ij2}, \ldots, p_{ijk})$. Under model (2.5), the unknown parameter may be written as $\boldsymbol{\theta} = (\boldsymbol{p}_{11}^{-}, \ldots, \boldsymbol{p}_{1J}^{-}, \ldots, \boldsymbol{p}_{I1}^{-}, \ldots, \boldsymbol{p}_{IJ}^{-})$, and its maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ replaces each $\boldsymbol{p}_{ij}^{-}$ with $\hat{\boldsymbol{p}}_{ij}^{-} = \boldsymbol{X}_{ij}^{-}/m_{ij}$. As discussed in Lehmann (2004, p. 314), we have a large sample normal distribution $\hat{\boldsymbol{\theta}} \stackrel{\cdot}{\sim} \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, where

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

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

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

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

$$\frac{\partial g_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{e}_j \otimes \left[c_1 \nabla \mathcal{E}\left(\sum_{i=1}^I c_i \boldsymbol{p}_{ij}^-\right) - d_1 \nabla \mathcal{E}\left(\sum_{i=1}^I d_i \boldsymbol{p}_{ij}^-\right)\right] \\ \vdots \\ \boldsymbol{e}_j \otimes \left[c_I \nabla \mathcal{E}\left(\sum_{i=1}^I c_i \boldsymbol{p}_{ij}^-\right) - d_I \nabla \mathcal{E}\left(\sum_{i=1}^I d_i \boldsymbol{p}_{ij}^-\right)\right] \end{pmatrix},$$

where $\nabla \mathcal{E}(\boldsymbol{p}^-) = (-\log(p_2/p_1), \ldots, -\log(p_k/p_1))^\top$, $\boldsymbol{e}_j$ is the jth column of a $J \times J$ identity matrix, and $\otimes$ denotes the matrix Kronecker product.

Remark 2.2. Some insight into the behavior of $g(\boldsymbol{\theta})$ can be seen from its gradient. Consider Scenario S1 with $J = 1$ and suppress the j index; we have

$$\begin{aligned} \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \begin{pmatrix} -\nabla \mathcal{E}(\boldsymbol{p}_1^-) \\ \nabla \mathcal{E}(\boldsymbol{p}_2^-) \end{pmatrix} \\ &= \Big(\log(p_{12}/p_{11}), \ldots, \log(p_{1k}/p_{11}), -\log(p_{22}/p_{21}), \ldots, -\log(p_{2k}/p_{21})\Big)^\top. \end{aligned}$$

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 $-\nabla \mathcal{E}(\boldsymbol{p}_1^-) \approx \boldsymbol{0}$; therefore, $g(\boldsymbol{\theta})$ 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 $-\nabla \mathcal{E}(\boldsymbol{p}_1^-) \to (0, \infty, \ldots, \infty)$; 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 $g(\boldsymbol{\theta})$.

Under the null hypothesis, the restriction $g_j(\boldsymbol{\theta}) = 0$ gives

$$\mathcal{Z} = \frac{g_j(\hat{\boldsymbol{\theta}})}{\sqrt{\sigma^2_{g_j(\hat{\boldsymbol{\theta}})}}} \stackrel{.}{\sim} \mathrm{N}(0,1).$$

Denoting α as the desired significance level for a test and z_α as the α quantile of the $\mathrm{N}(0,1)$ distribution, we obtain the usual tests for $g(\boldsymbol{\theta})$ based on normality: reject H_0 in hypothesis (2.2) if $|\mathcal{Z}| > z_{\alpha/2}$, reject H_0 in hypothesis (2.3) if $\mathcal{Z} > z_\alpha$, or reject H_0 in hypothesis (2.4) if $\mathcal{Z} < z_\alpha$. We can also obtain the usual level $1-\alpha$ confidence limits for $g(\boldsymbol{\theta})$: the confidence interval $g(\hat{\boldsymbol{\theta}}) \pm z_{\alpha/2}\sigma_{g(\hat{\boldsymbol{\theta}})}$, the lower confidence limit $g(\hat{\boldsymbol{\theta}}) - z_\alpha \sigma_{g(\hat{\boldsymbol{\theta}})}$, and the upper confidence limit $g(\hat{\boldsymbol{\theta}}) + z_\alpha \sigma_{g(\hat{\boldsymbol{\theta}})}$. Code for all procedures has been implemented in the `R` programming language (R Core Team, 2018).

3 Simulations

We present several simulations to study properties of the procedures discussed in Section 2. Here, we consider a setting based on Scenario S1 and a setting based on S2, taking $J = 1$ in both. We consider empirical rejection rates of hypothesis (2.3) with significance level $\alpha = 0.05$, as well as the empirical coverage and "width" for level $1 - \alpha = 0.95$ lower confidence limits for $g(\boldsymbol{\theta})$.

3.1 Scenario S1

Consider the setting of Scenario S1 and suppose

$$\boldsymbol{p}_1 := \begin{pmatrix} p_{11} \\ p_{12} \\ p_{13} \\ p_{14} \\ p_{15} \\ p_{16} \\ p_{17} \end{pmatrix} = \frac{1}{28} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{pmatrix} \approx \begin{pmatrix} 0.0357 \\ 0.0714 \\ 0.1071 \\ 0.1429 \\ 0.1786 \\ 0.2143 \\ 0.2500 \end{pmatrix} \quad \text{and} \quad \boldsymbol{p}_2 = \begin{pmatrix} p_{14} \\ p_{15} \\ p_{16} \\ p_{17} - \delta \\ p_{11} + \delta \\ p_{12} \\ p_{13} \end{pmatrix} \approx \begin{pmatrix} 0.1429 \\ 0.1786 \\ 0.2143 \\ 0.2500 - \delta \\ 0.0357 + \delta \\ 0.0714 \\ 0.1071 \end{pmatrix},$$

where $1/28$ is the normalizing constant needed to transform the vector $(1, 2, \ldots, 7)$ to a probability distribution. Here we consider the effect on $g(\boldsymbol{\theta})$ when $\boldsymbol{p}_2$ is a permutation of $\boldsymbol{p}_1$ except that the most frequent category donates some of its probability mass to the least frequent category. Restricting our attention to hypothesis (2.3), H_0 is true when $\delta = 0$ because $\boldsymbol{p}_1$ is exactly a permutation of $\boldsymbol{p}_2$. H_0 is also true for $\delta < 0$, where $\boldsymbol{p}_2$ becomes a more peaked distribution than $\boldsymbol{p}_1$. The value of $g(\boldsymbol{\theta}) = \mathcal{E}(\boldsymbol{p}_2) - \mathcal{E}(\boldsymbol{p}_1)$ increases to its maximum as δ increases from zero to 0.1071428, but decreases as δ is increased further; see Figure 1a. The following steps are repeated 1,000 times:

1. Draw $\boldsymbol{X}_1 \sim \text{Mult}_k(m, \boldsymbol{p}_1)$ and $\boldsymbol{X}_2 \sim \text{Mult}_k(m, \boldsymbol{p}_2)$.
2. Compute the $\mathcal{Z}$-statistic and lower confidence limit $\mathcal{L}$ based on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

The empirical rejection rate is obtained from the proportion of rejections of the test $\mathcal{Z} > z_\alpha$. The empirical coverage of the lower confidence limit is obtained by the proportion of instances where $\mathcal{L} \leq g(\boldsymbol{\theta})$. The empirical width of the confidence limit is computed by averaging the individual widths $g(\boldsymbol{\theta}) - \mathcal{L}$. This was repeated for each

$$\begin{aligned} &\delta \in \{-0.02, -0.01, -0.005, -0.002, -0.001, -0.0005, -0.0002, -0.0001, 0, \\ &\qquad 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.14\}, \\ \text{and} \quad &m \in \{10, 20, 50, 100, 200, 500, 1000, 2000, 5000\}. \end{aligned}$$

Tables 1, 2, and 3 present results for empirical rejection rate, empirical coverage, and empirical width respectively.

Results are mostly as expected, but some interesting features can be noted. Table 1 shows that the rejection rate reaches the nominal level of 0.05 as the sample size increases and δ approaches 0 from below. The power of the test increases when δ approaches the value 0.1071428 (which maximizes $g(\boldsymbol{\theta})$), with the increase being faster as the sample size is taken larger. We notice some oscillations in the power; for example, when $\delta = 0$, the rejection rate reduces from 0.0490 at $m = 2000$ to 0.0460 at 5000. Table 2 shows that coverage probability for the confidence limit approaches the nominal 0.95 level as sample size becomes large, for all values of δ. Oscillations in coverage probability can be seen, for example, when $\delta = 0.0002$ and m increases from 500 to 5000. Empirical width shown in Table 3 appears to be decreasing for all δ as sample size is increased. However, for fixed δ widths appear to become smaller as $g(\boldsymbol{\theta})$ is taken larger.

Regarding the oscillations, our $\mathcal{Z}$-statistic is closely related to the Wald statistic used for inference on p in the situation $X \sim \text{Binomial}(m, p)$. Brown et al. (2001) discuss at length how coverage probabilities for confidence intervals based on the Wald statistic exhibit an oscillating behavior as m and p vary. For example, with p fixed, coverage probability will meet the nominal level for some m and fail to meet the nominal level for some larger m. Brown et al. (2001) suggest several alternative intervals based on the binomial data, and more recently Franco et al. (2014) and Kott (2017) compare binomial intervals in the setting of complex surveys. Alternative intervals could also be considered in our setting.

3.2 Scenario S2

Now consider the setting of Scenario S2 and let

$$\boldsymbol{p}_1 = \frac{1}{28}\begin{pmatrix}1\\2\\3\\4\\5\\6\\7\end{pmatrix} \approx \begin{pmatrix}0.0357\\0.0714\\0.1071\\0.1429\\0.1786\\0.2143\\0.2500\end{pmatrix},$$

$\boldsymbol{q} = \pi\boldsymbol{p}_2 + (1-\pi)\boldsymbol{p}_3$ with $\pi = 1/2$, $\boldsymbol{p}_2 = \boldsymbol{p}_1$, and

$$\boldsymbol{p}_3 = (1-\delta)\boldsymbol{p}_1 + \delta\frac{1}{28}\begin{pmatrix}7\\6\\5\\4\\3\\2\\1\end{pmatrix} = \frac{1}{28}\begin{pmatrix}(1-\delta)1+\delta 7\\(1-\delta)2+\delta 6\\(1-\delta)3+\delta 5\\(1-\delta)4+\delta 4\\(1-\delta)5+\delta 3\\(1-\delta)6+\delta 2\\(1-\delta)7+\delta 1\end{pmatrix}.$$

Focusing on hypothesis (2.3), H_0 is true when $\delta = 0$ so that $\boldsymbol{p}_1 = \boldsymbol{q}$. The value of $g(\boldsymbol{\theta}) = \mathcal{E}(\boldsymbol{q}) - \mathcal{E}(\boldsymbol{p})$ increases to its maximum when $\delta = 1$, where

$$\boldsymbol{q} = \frac{1}{28}\Big[\pi(1,\ldots,7) + (1-\pi)(7,\ldots,1)\Big] = \frac{1}{7}(1,\ldots,1)$$

is the discrete uniform distribution. Figure 1b displays a plot of $g(\boldsymbol{\theta})$ for $\delta \in (0,1)$. The following steps are repeated 1,000 times:

1. Draw $m_2 \sim \text{Binomial}(m,\pi)$ and let $m_3 = m - m_2$.
2. Draw $\boldsymbol{X}_1 \sim \text{Mult}_k(m,\boldsymbol{p}_1)$, $\boldsymbol{X}_2 \sim \text{Mult}_k(m_2,\boldsymbol{p}_2)$, and $\boldsymbol{X}_3 \sim \text{Mult}_k(m_3,\boldsymbol{p}_3)$.
3. Compute the $\mathcal{Z}$-statistic and lower confidence limit $\mathcal{L}$ based on $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\hat{\pi} = m_2/(m_2+m_3)$.

Scenario 2 is otherwise similar to Scenario 1, except that we consider

$$\begin{aligned}&\delta \in \{0, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}\\ \text{and}\quad &m \in \{50, 100, 200, 500, 1000, 2000, 5000\}.\end{aligned}$$

We avoided the smallest sample sizes from Scenario 1 which occasionally resulted in $m_2 = 0$ or $m_3 = 0$.

Tables 4, 5, and 6 present results for empirical rejection rate, empirical coverage, and empirical width respectively. As in Scenario 1, results here are mostly as expected, with similar oscillations in rejection rate and coverage. The effect size $g(\boldsymbol{\theta})$ becomes substantially larger here than in Scenario 1 for the larger choices of δ; see Figure 1b.

4 Data Analysis

Our objective is now to determine the effect of mailing strategy on the uniformity of volumes of calls to the helpline. Namely, we consider a staggered strategy where respondents are randomly partitioned into several groups which are sent mailings on different schedules, as well as more traditional strategies where all respondents are sent mailings on the same schedule. It is thought that a staggered strategy leads to a more uniform distribution of calls than an unstaggered strategy. To explore this theory, we make use of datasets from three census experiments in which 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 2016 September NCBS (Eggleston and Coombs, 2017) and 2016 June NCBS (Coombs, 2017) are two experiments where an unstaggered mailing strategy was utilized. A staggered mailing schedule was used in the 2017 March NCBS (Nichols et al., 2018); study participants 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 7 displays the schedules for each mailing in the three experiments. In these studies, no live agents were present to answer the helpline and callers instead received a prerecorded message. Callers' identities were not recorded, so the data do not distinguish whether multiple calls were made by the same caller.

To compare uniformity of call volumes between the three experiments, we first examine plots of call frequencies. Figures 2, 3, and 4 present 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 third day happened to fall on a weekend. Even with a staggered mailout, Figure 2 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 the Labor Day holiday was observed on September 5, 2016, so the expected Monday spike in call volume shifted to the next day (September 6). Figures 5, 6, 7, and 8 present call frequencies summed by day of week. Figure 5 shows Monday and Thursday Mailout groups separately, while Figure 6 combines them. It appears that the combined distribution in Figure 6 is flatter than either Figure 7 or Figure 8. Figure 8 has a large spike occurring on Monday, and therefore appears to be the lowest entropy—or furthest from uniform—distribution.

For each census experiment, 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 census experiment, and disregard calls which occurred in week 6 or later because call activity becomes sparse. For each pair of census experiments, we compare the entropy for week j of the first experiment to week j of the second experiment, for $j = 1, \dots, 5$. 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, but such a choice should not be based on the observed data. Table 8 reports the weekly call counts for each experiment. It is also possible to compare different weeks from pairs of experiments, 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 group (7 calls). Therefore, if increased entropy is observed in week 1 for the 2017 March NCBS, it is likely due to some factor other than the staggered mailing schedule.

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 S2. 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 8. We test hypothesis (2.3) for each week $j = 1, \dots, 5$, 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 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 means that the 2017 March NCBS mailing strategy leads to a more uniform distribution of calls during that week.

Table 9a gives results of testing this hypothesis for weeks 1–5. Recall that quantity $g(\boldsymbol{\theta})$ is bounded, so that $-1.94591 \leq g(\boldsymbol{\theta}) \leq 1.94591$ for any $\boldsymbol{\theta}$. The 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 $\mathcal{Z}$-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 10a displays the estimated probabilities $\hat{\boldsymbol{p}}_{1j}$ and $\hat{\boldsymbol{q}}_j$ for weeks $j = 1, \dots, 5$. 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 and Tuesday. Many calls occurring on these peak days are from the Thursday Mailout group, as seen in Figure 2, 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 S2. Take $\boldsymbol{X}_{1j}$ to be the frequencies of the 2016 June NCBS calls observed on (Sunday, Monday, ..., Saturday). Take $\boldsymbol{X}_{2j}$ and $\boldsymbol{X}_{3j}$ to be the day-of-week frequencies of 2017 March NCBS calls for the Monday Mailout and Thursday Mailout treatments, respectively. Estimates for π_j are given in Table 8. We test hypothesis (2.3) for each week $j = 1, \dots, 5$, 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 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 means that the 2017 March NCBS mailing strategy leads to a more uniform distribution of calls during that week.

Table 9b gives results of testing this hypothesis for weeks 1–5. Table 10b displays the estimated probabilities $\hat{\boldsymbol{p}}_{1j}$ and $\hat{\boldsymbol{q}}_j$ 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 S1. Take $\boldsymbol{X}_{1j}$ to be the frequencies of 2016 June NCBS calls observed on (Sunday, Monday, ..., Saturday) and $\boldsymbol{X}_{2j}$ to be the frequencies of 2016 September NCBS calls. Because both experiments 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 (2.2) 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 has equal entropy to the day-of-week distribution resulting from the 2016 September NCBS mailing schedule" versus H_1: "Not".

Here, rejection of H_0 for week j means that the two mailing strategies do not lead to an equally uniform distribution of calls during that week.

Table 9c gives results of testing this hypothesis for weeks 1–5, and Table 10c displays the estimated probabilities $\hat{\boldsymbol{p}}_{1j}$ and $\hat{\boldsymbol{p}}_{2j}$ 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 $\mathcal{Z}$-statistics are positive for weeks 2–5, indicating a larger entropy for the 2016 September NCBS except during week 1.

5 Discussion

In this work, we compared pairs of discrete distributions to infer which is closer to a discrete uniform distribution. We developed procedures using basic large sample theory and applied them to call volume data from several census experiments. Our analysis found that the staggered strategy—the one used in the 2017 March NCBS—yielded a significantly higher entropy than the two unstaggered experiments 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 the 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 as well. One way to attenuate starting and ending differences among census experiments 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 there appears to be evidence that staggering increases call uniformity, a designed experiment would help to distinguish this apart from other factors. The three available experiments were carried out at different time periods on populations which may also be considered different. Mailing strategies could be compared more reliably on a common population and time period. Furthermore, variations of mailing strategy treatments could be more carefully controlled to study their individual effects.

We noted in Remark 2.2 that changes in the quantity $g(\boldsymbol{\theta})$ are smaller when the component distributions are closer to uniform. Differences in this setting may therefore be difficult to detect with the $\mathcal{Z}$-statistic, which may warrant investigation of alternative distance measures. Alternative distance measures might also be considered if one is thought to better express the cost of departure from uniformity than K-L divergence.

While our model was based on independent multinomial observations, we could consider a regression model with appropriate covariates. 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 Census Bureau experiments, as a step beyond usual statistical inference.

Acknowledgements

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.

References

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–117, 2001.

John Chesnut. Telephone questionnaire assistance. Census 2000 Evaluation A.1.a, U.S. Census Bureau, 2003. URL https://www.census.gov/pred/www/rpts/A.1.a.pdf.

Arthur Cohen, John Kolassa, and Harold Sackrowitz. *A test for equality of multinomial distributions vs increasing convex order*, volume 50 of *Lecture Notes–Monograph Series*, pages 156–163. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2006.

Julia Coombs. 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. In *2020 Research and Testing*. 2017. (Internal Report).

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Peter Dorfinger, Georg Panholzer, and Wolfgang John. Entropy estimation for real-time encrypted traffic identification (short paper). In Jordi Domingo-Pascual, Yuval Shavitt, and Steve Uhlig, editors, *Traffic Monitoring and Analysis*, pages 164–171, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

Casey Eggleston and Julia Coombs. Effect of data use statements and postcard format on login rates for a mandatory online Census Bureau survey. In *2020 Research and Testing*. 2017. (Internal Report).

Carolina Franco, Roderick J .A. Little, Thomas A. Louis, and Eric V. Slud. Coverage properties of confidence intervals for proportions in complex sample surveys. In *JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association*, pages 1799–1813, 2014.

Valérie Girardin and Justine Lequesne. Entropy-based goodness-of-fit tests-a unifying framework: Application to dna replication. *Communications in Statistics - Theory and Methods*, 2017. doi: 10.1080/03610926.2017.1401084.

Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer, 2008.

Phillip S. Kott. A note on Wilson coverage intervals for proportions estimated from complex samples. *Survey Methodology*, 43(2):235–240, 2017.

E. L. Lehmann. *Elements of Large-Sample Theory*. Springer, 2004.

Xinsheng Liu and Jinde Wang. Testing the equality of multinomial populations ordered by increasing convexity under the alternative. *Canadian Journal of Statistics*, 32(2):159–168, 2004.

Elizabeth Nichols, Sarah Konya, Rachel Horwitz, and Andrew Raim. The effect of the mail delivery date on survey login rates and helpline call rates. In *2020 Research and Testing Report*. 2018. (Forthcoming).

J. T. Ormerod and M. P. Wand. Explaining variational approximations. *The American Statistician*, 64(2): 140–153, 2010.

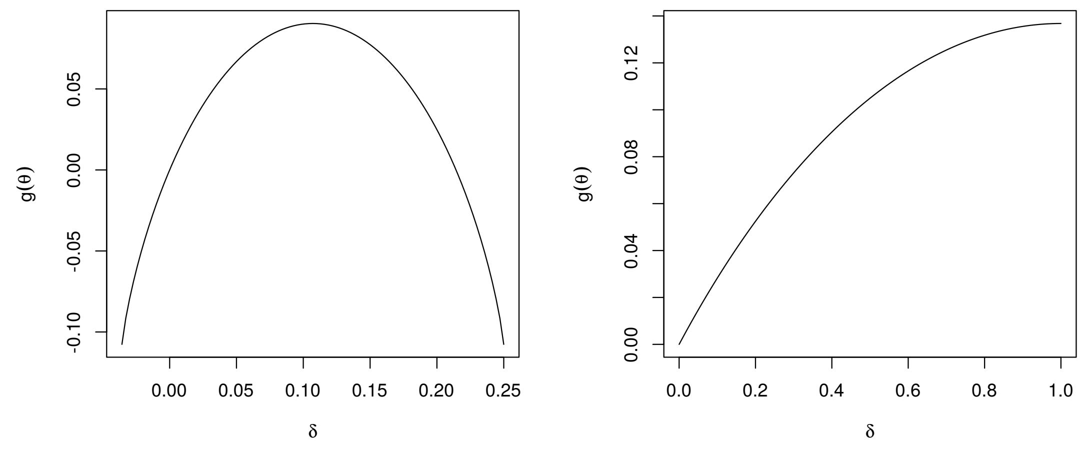
Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

Leandro Pardo. *Statistical inference based on divergence measures*. Chapman & Hall/CRC, 2006.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

U.S. Census Bureau. 2020 Census Operational Plan, 2017. URL https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/planning-docs/operational-plan.html. Version 3.0.

Kevin Zajac. 2010 census telephone questionnaire assistance assessment report. 2010 Census Planning Memoranda Series No. 231, U.S. Census Bureau, 2012. URL https://www.census.gov/2010census/pdf/2010_Census_TQA_Assessment.pdf.



(a)

(b)

Figure 1: The value of $g(\boldsymbol{\theta})$ for: (a) $\delta \in (-0.03571429, 0.25)$, when $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are as prescribed in Section 3.1, and (b) $\delta \in (0,1)$, when $\boldsymbol{p}_1$ and $\boldsymbol{q}$ are as prescribed in Section 3.2.

Table 1: Empirical rejection rate of one-sided test from simulation study for Scenario 1.

δ	$g(\boldsymbol{\theta})$	$m = 10$	20	50	100	200	500	1000	2000	5000
-0.0200	-0.0468	0.1340	0.0850	0.0460	0.0150	0.0050	0.0000	0.0000	0.0000	0.0000
-0.0100	-0.0212	0.1600	0.1070	0.0480	0.0290	0.0120	0.0070	0.0060	0.0010	0.0000
-0.0050	-0.0101	0.1590	0.1180	0.0690	0.0380	0.0310	0.0190	0.0160	0.0100	0.0020
-0.0020	-0.0040	0.1930	0.1270	0.0760	0.0550	0.0390	0.0350	0.0280	0.0330	0.0180
-0.0010	-0.0020	0.1620	0.1350	0.0780	0.0530	0.0470	0.0380	0.0370	0.0380	0.0360
-0.0005	-0.0010	0.1920	0.1140	0.0760	0.0400	0.0420	0.0450	0.0430	0.0450	0.0470
-0.0002	-0.0004	0.1620	0.1080	0.0640	0.0600	0.0320	0.0370	0.0550	0.0450	0.0390
-0.0001	-0.0002	0.1410	0.1160	0.0800	0.0450	0.0490	0.0380	0.0440	0.0530	0.0440
0.0000	0.0000	0.1640	0.1100	0.0800	0.0510	0.0450	0.0550	0.0430	0.0490	0.0460
0.0001	0.0002	0.1810	0.1250	0.0890	0.0600	0.0490	0.0570	0.0430	0.0490	0.0670
0.0002	0.0004	0.1730	0.1070	0.0680	0.0440	0.0490	0.0430	0.0370	0.0560	0.0640
0.0005	0.0010	0.1740	0.1140	0.0740	0.0660	0.0680	0.0520	0.0510	0.0520	0.0630
0.0010	0.0019	0.1620	0.1450	0.0870	0.0560	0.0510	0.0550	0.0530	0.0600	0.0710
0.0020	0.0038	0.1930	0.1450	0.0820	0.0650	0.0630	0.0660	0.0680	0.0840	0.1080
0.0050	0.0093	0.1690	0.1290	0.0860	0.0720	0.0680	0.0810	0.1300	0.1650	0.2810
0.0100	0.0180	0.1900	0.1460	0.0950	0.0820	0.0810	0.1630	0.2160	0.3580	0.6180
0.0200	0.0333	0.1840	0.1580	0.1390	0.1390	0.1730	0.3260	0.5200	0.7870	0.9850
0.0500	0.0669	0.2340	0.1910	0.2030	0.2490	0.4810	0.8170	0.9780	0.9990	1.0000
0.1000	0.0900	0.2600	0.2460	0.2520	0.4050	0.7230	0.9770	0.9990	1.0000	1.0000
0.1400	0.0828	0.2540	0.2120	0.2560	0.3800	0.6110	0.9530	0.9980	1.0000	1.0000

Table 2: Empirical coverage rate of lower confidence limit from simulation study for Scenario 1.

δ	$g(\boldsymbol{\theta})$	$m = 10$	20	50	100	200	500	1000	2000	5000
-0.0200	-0.0468	0.8220	0.8760	0.9170	0.9440	0.9530	0.9500	0.9430	0.9430	0.9450
-0.0100	-0.0212	0.8280	0.8770	0.9340	0.9330	0.9440	0.9580	0.9520	0.9450	0.9530
-0.0050	-0.0101	0.8350	0.8730	0.9180	0.9450	0.9360	0.9580	0.9510	0.9520	0.9440
-0.0020	-0.0040	0.8050	0.8700	0.9220	0.9410	0.9520	0.9440	0.9570	0.9500	0.9540
-0.0010	-0.0020	0.8380	0.8630	0.9180	0.9430	0.9470	0.9550	0.9570	0.9540	0.9480
-0.0005	-0.0010	0.8080	0.8840	0.9230	0.9580	0.9550	0.9520	0.9560	0.9470	0.9440
-0.0002	-0.0004	0.8380	0.8920	0.9350	0.9390	0.9680	0.9620	0.9420	0.9530	0.9600
-0.0001	-0.0002	0.8590	0.8840	0.9200	0.9550	0.9510	0.9620	0.9540	0.9470	0.9550
0.0000	0.0000	0.8360	0.8900	0.9200	0.9490	0.9550	0.9450	0.9570	0.9510	0.9540
0.0001	0.0002	0.8190	0.8750	0.9120	0.9400	0.9510	0.9440	0.9580	0.9540	0.9380
0.0002	0.0004	0.8270	0.8930	0.9320	0.9570	0.9520	0.9580	0.9660	0.9480	0.9400
0.0005	0.0010	0.8260	0.8890	0.9270	0.9350	0.9330	0.9500	0.9540	0.9540	0.9500
0.0010	0.0019	0.8380	0.8560	0.9140	0.9460	0.9550	0.9530	0.9540	0.9520	0.9520
0.0020	0.0038	0.8090	0.8580	0.9210	0.9360	0.9460	0.9500	0.9530	0.9500	0.9500
0.0050	0.0093	0.8410	0.8790	0.9230	0.9450	0.9590	0.9580	0.9470	0.9490	0.9460
0.0100	0.0180	0.8330	0.8730	0.9270	0.9450	0.9610	0.9420	0.9520	0.9480	0.9490
0.0200	0.0333	0.8430	0.8780	0.9080	0.9430	0.9530	0.9460	0.9480	0.9580	0.9540
0.0500	0.0669	0.8320	0.8860	0.9080	0.9400	0.9530	0.9550	0.9660	0.9510	0.9580
0.1000	0.0900	0.8310	0.8680	0.9190	0.9580	0.9630	0.9490	0.9450	0.9490	0.9480
0.1400	0.0828	0.8330	0.8830	0.9180	0.9480	0.9410	0.9510	0.9530	0.9480	0.9500

Table 3: Empirical width of lower confidence limit from simulation study for Scenario 1.

δ	$g(\boldsymbol{\theta})$	$m = 10$	20	50	100	200	500	1000	2000	5000
-0.0200	-0.0468	0.2639	0.2319	0.1647	0.1142	0.0817	0.0508	0.0358	0.0246	0.0156
-0.0100	-0.0212	0.2634	0.2417	0.1711	0.1152	0.0786	0.0503	0.0357	0.0242	0.0156
-0.0050	-0.0101	0.2706	0.2345	0.1590	0.1175	0.0795	0.0499	0.0353	0.0248	0.0151
-0.0020	-0.0040	0.2697	0.2302	0.1669	0.1166	0.0808	0.0484	0.0349	0.0242	0.0157
-0.0010	-0.0020	0.2773	0.2291	0.1629	0.1152	0.0787	0.0490	0.0347	0.0245	0.0153
-0.0005	-0.0010	0.2490	0.2516	0.1635	0.1196	0.0790	0.0512	0.0345	0.0247	0.0148
-0.0002	-0.0004	0.2841	0.2428	0.1655	0.1124	0.0809	0.0492	0.0357	0.0246	0.0158
-0.0001	-0.0002	0.2898	0.2485	0.1670	0.1116	0.0793	0.0491	0.0341	0.0233	0.0149
0.0000	0.0000	0.2903	0.2380	0.1629	0.1132	0.0806	0.0498	0.0345	0.0243	0.0155
0.0001	0.0002	0.2737	0.2438	0.1613	0.1106	0.0771	0.0475	0.0350	0.0235	0.0149
0.0002	0.0004	0.2757	0.2449	0.1684	0.1137	0.0802	0.0498	0.0337	0.0237	0.0152
0.0005	0.0010	0.2748	0.2416	0.1660	0.1110	0.0787	0.0483	0.0350	0.0239	0.0155
0.0010	0.0019	0.2928	0.2269	0.1629	0.1129	0.0812	0.0482	0.0345	0.0244	0.0151
0.0020	0.0038	0.2613	0.2370	0.1625	0.1152	0.0791	0.0480	0.0337	0.0243	0.0157
0.0050	0.0093	0.2782	0.2428	0.1625	0.1115	0.0789	0.0496	0.0336	0.0238	0.0150
0.0100	0.0180	0.2733	0.2334	0.1630	0.1136	0.0788	0.0471	0.0331	0.0239	0.0152
0.0200	0.0333	0.2945	0.2372	0.1513	0.1110	0.0752	0.0461	0.0334	0.0226	0.0145
0.0500	0.0669	0.2868	0.2383	0.1527	0.1079	0.0706	0.0442	0.0312	0.0217	0.0136
0.1000	0.0900	0.2776	0.2253	0.1468	0.1021	0.0676	0.0421	0.0286	0.0208	0.0128
0.1400	0.0828	0.2787	0.2392	0.1480	0.1025	0.0690	0.0422	0.0301	0.0208	0.0133

Table 4: Empirical rejection rate of one-sided test from simulation study for Scenario 2.

δ	$g(\boldsymbol{\theta})$	$m = 50$	100	200	500	1000	2000	5000
0.0000	0.0000	0.0800	0.0540	0.0430	0.0520	0.0390	0.0470	0.0480
0.0001	0.0000	0.0700	0.0620	0.0380	0.0620	0.0580	0.0430	0.0590
0.0002	0.0001	0.0780	0.0610	0.0470	0.0550	0.0390	0.0590	0.0590
0.0005	0.0002	0.0800	0.0680	0.0540	0.0560	0.0660	0.0570	0.0520
0.0010	0.0003	0.0880	0.0470	0.0430	0.0520	0.0410	0.0410	0.0630
0.0020	0.0006	0.0680	0.0540	0.0520	0.0520	0.0560	0.0590	0.0480
0.0050	0.0015	0.0710	0.0560	0.0550	0.0710	0.0620	0.0590	0.0690
0.0100	0.0030	0.0880	0.0690	0.0650	0.0680	0.0650	0.0710	0.0880
0.0200	0.0060	0.0790	0.0530	0.0750	0.0750	0.0880	0.1120	0.1630
0.0500	0.0147	0.0930	0.0760	0.0940	0.1250	0.2020	0.2430	0.5090
0.1000	0.0283	0.1200	0.1140	0.1490	0.2560	0.4370	0.6140	0.9410
0.2000	0.0525	0.1680	0.2200	0.2920	0.6340	0.8620	0.9910	1.0000
0.5000	0.1049	0.3200	0.5500	0.8560	0.9970	1.0000	1.0000	1.0000
1.0000	0.1368	0.4520	0.8350	0.9970	1.0000	1.0000	1.0000	1.0000

Table 5: Empirical coverage of lower confidence limit from simulation study for Scenario 2.

δ	$g(\boldsymbol{\theta})$	$m = 50$	100	200	500	1000	2000	5000
0.0000	0.0000	0.9200	0.9460	0.9570	0.9480	0.9610	0.9530	0.9520
0.0001	0.0000	0.9300	0.9380	0.9620	0.9380	0.9420	0.9580	0.9410
0.0002	0.0001	0.9220	0.9390	0.9540	0.9450	0.9620	0.9410	0.9420
0.0005	0.0002	0.9200	0.9320	0.9460	0.9460	0.9350	0.9460	0.9500
0.0010	0.0003	0.9130	0.9530	0.9580	0.9480	0.9600	0.9600	0.9400
0.0020	0.0006	0.9320	0.9460	0.9510	0.9500	0.9450	0.9450	0.9590
0.0050	0.0015	0.9310	0.9460	0.9490	0.9370	0.9480	0.9530	0.9460
0.0100	0.0030	0.9150	0.9370	0.9480	0.9440	0.9520	0.9600	0.9500
0.0200	0.0060	0.9290	0.9570	0.9360	0.9470	0.9440	0.9490	0.9490
0.0500	0.0147	0.9240	0.9500	0.9440	0.9580	0.9490	0.9500	0.9360
0.1000	0.0283	0.9200	0.9550	0.9450	0.9580	0.9430	0.9570	0.9430
0.2000	0.0525	0.9130	0.9430	0.9510	0.9490	0.9580	0.9520	0.9390
0.5000	0.1049	0.9180	0.9470	0.9640	0.9550	0.9560	0.9450	0.9370
1.0000	0.1368	0.9120	0.9560	0.9550	0.9610	0.9560	0.9650	0.9500

Table 6: Empirical width of lower confidence limit from simulation study for Scenario 2.

δ	$g(\boldsymbol{\theta})$	$m = 50$	100	200	500	1000	2000	5000
0.0000	0.0000	0.1556	0.1133	0.0791	0.0471	0.0353	0.0244	0.0155
0.0001	0.0000	0.1615	0.1129	0.0794	0.0472	0.0337	0.0249	0.0153
0.0002	0.0001	0.1572	0.1098	0.0782	0.0487	0.0360	0.0234	0.0153
0.0005	0.0002	0.1592	0.1160	0.0783	0.0473	0.0341	0.0245	0.0152
0.0010	0.0003	0.1562	0.1134	0.0786	0.0488	0.0354	0.0251	0.0153
0.0020	0.0006	0.1591	0.1135	0.0788	0.0493	0.0340	0.0240	0.0158
0.0050	0.0015	0.1615	0.1159	0.0807	0.0474	0.0341	0.0245	0.0150
0.0100	0.0030	0.1595	0.1122	0.0793	0.0494	0.0346	0.0244	0.0151
0.0200	0.0060	0.1621	0.1164	0.0773	0.0489	0.0344	0.0248	0.0150
0.0500	0.0147	0.1597	0.1115	0.0784	0.0487	0.0325	0.0242	0.0146
0.1000	0.0283	0.1547	0.1137	0.0769	0.0470	0.0321	0.0238	0.0145
0.2000	0.0525	0.1499	0.1043	0.0761	0.0441	0.0316	0.0215	0.0141
0.5000	0.1049	0.1401	0.0961	0.0656	0.0393	0.0271	0.0185	0.0119
1.0000	0.1368	0.1341	0.0915	0.0580	0.0358	0.0250	0.0171	0.0111

Table 7: Schedule of mailings.

(a) The 2017 March NCBS 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.

	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) The 2016 September NCBS was targeted to 9,000 recipients.

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) The 2016 June NCBS was targeted to 8,000 recipients.

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 8: 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	512	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	2739	721	627	1348	

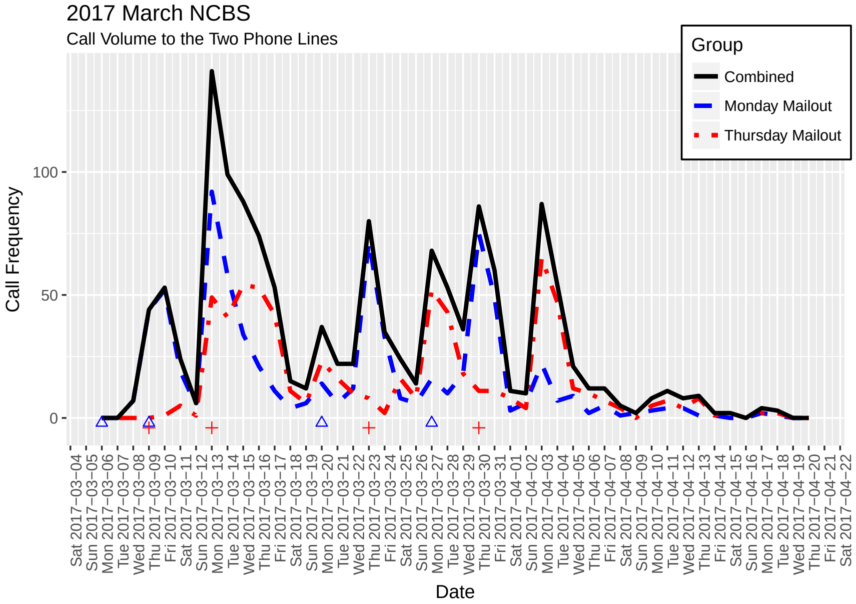


Figure 2: Daily call volumes during 2017 March NCBS Test. Blue triangles and red pluses along the x-axis represent mailing dates for Monday and Thursday Mailout groups, respectively.

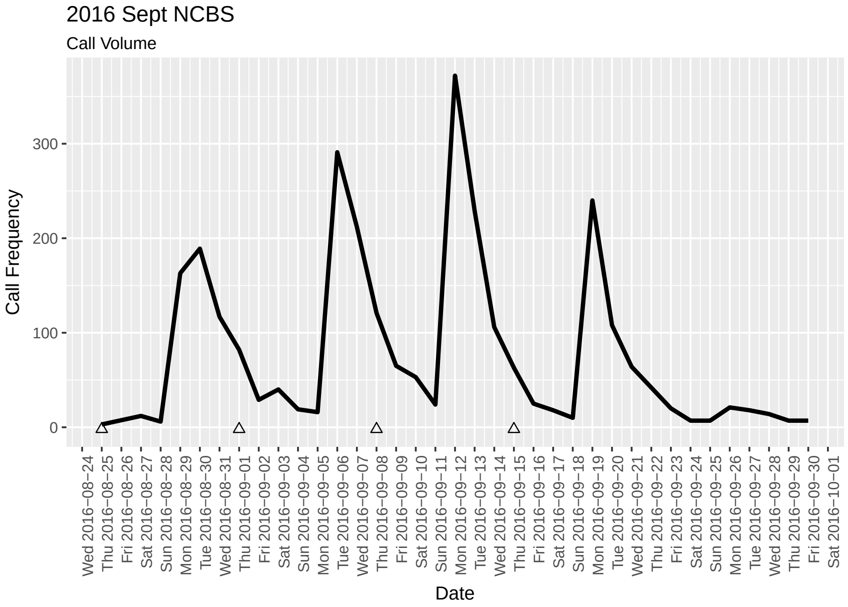


Figure 3: Daily call volumes during 2016 September NCBS. Black triangles along the x-axis represent mailing dates.

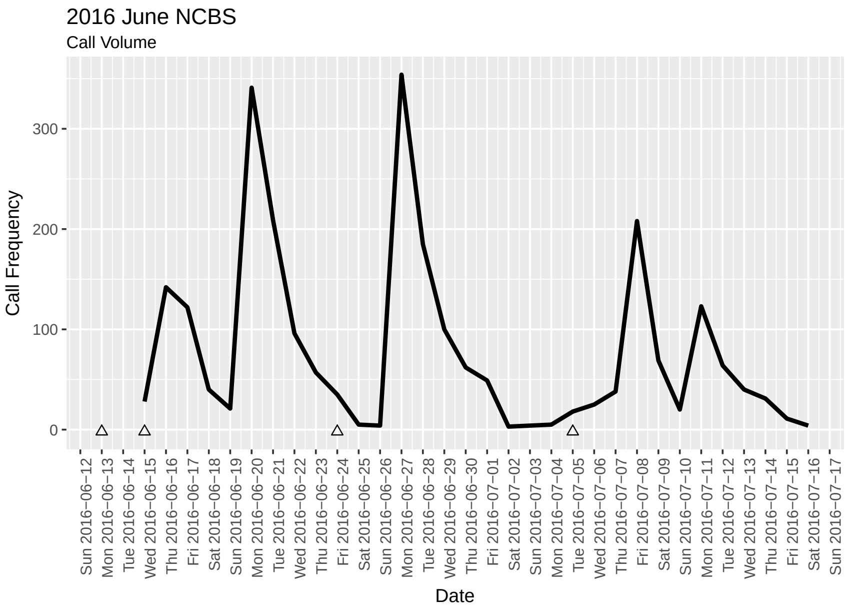


Figure 4: Daily call volumes during 2016 June NCBS. Black triangles along the x-axis represent mailing dates.

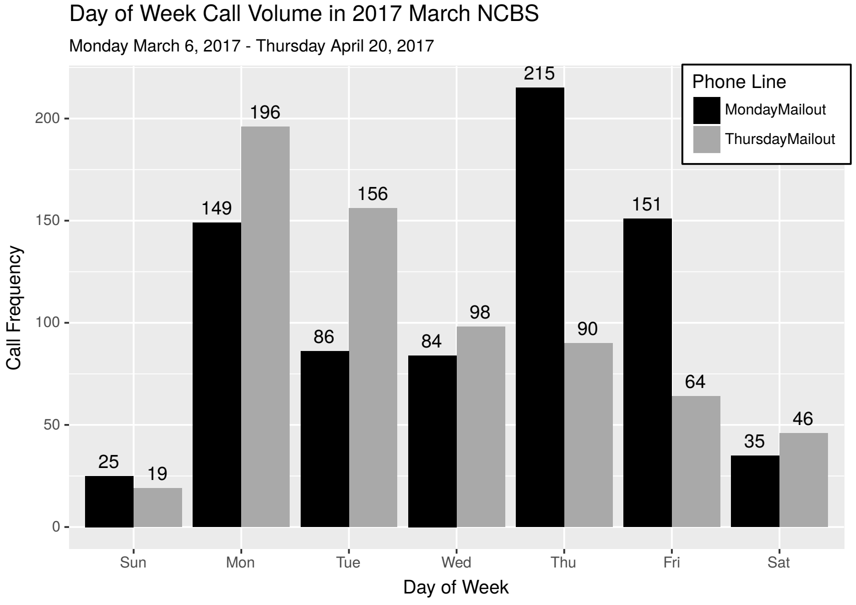


Figure 5: Day-of-week call volumes for 2017 March NCBS Test.

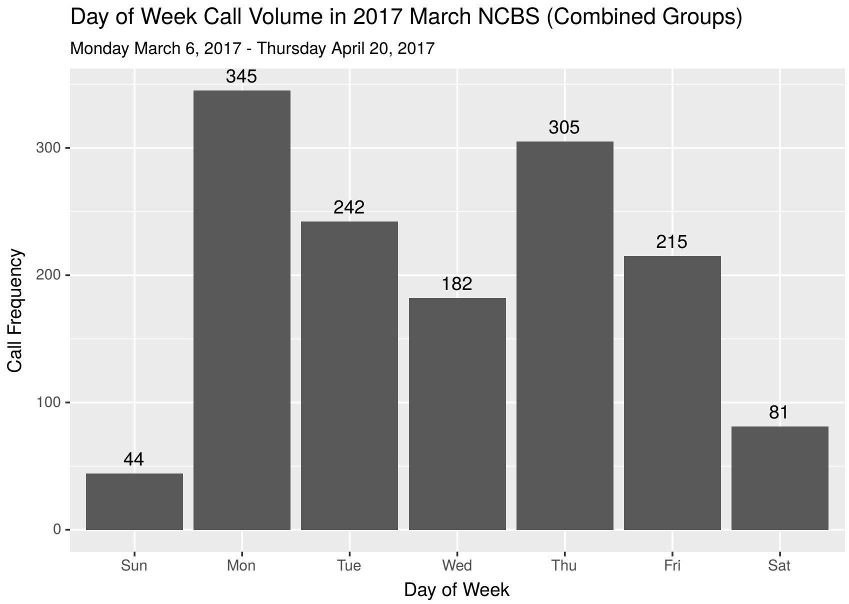


Figure 6: Day-of-week call volumes for 2017 March NCBS Test; Monday and Thursday Mailout groups combined.

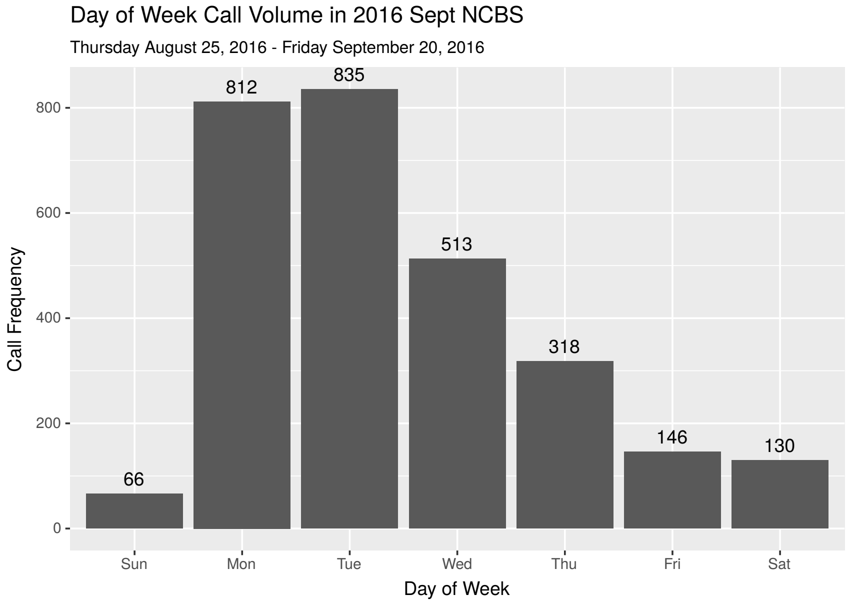


Figure 7: Day-of-week call volumes for 2016 September NCBS.

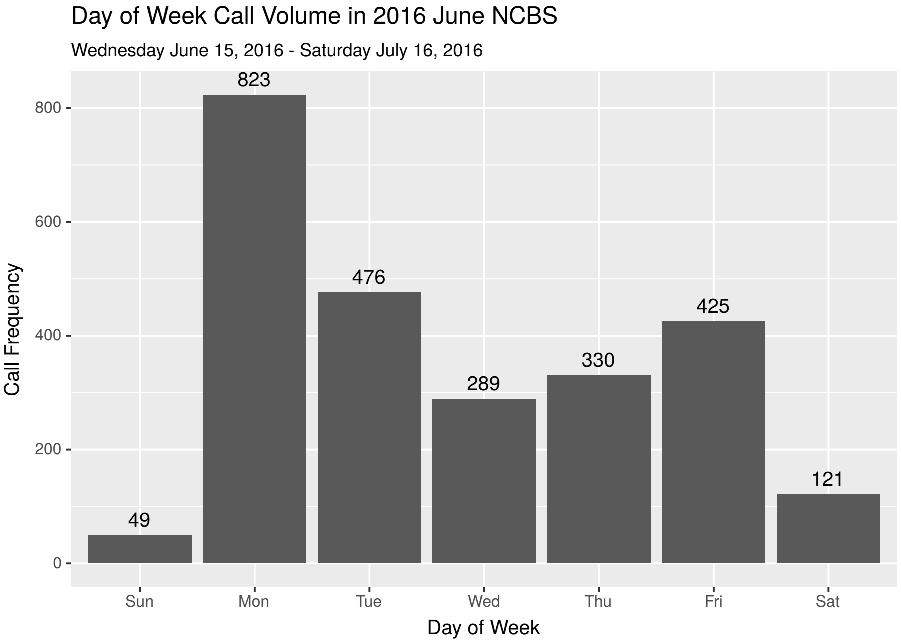


Figure 8: Day-of-week call volumes for 2016 June NCBS.

Table 9: Results for inference on the quantity $g(\boldsymbol{\theta})$. Estimates, standard errors (SEs), and $\mathcal{Z}$-statistics are displayed in the first three columns. (a) and (c) give a p-value for hypothesis (2.3) and a level 0.90 lower confidence limit for $g(\boldsymbol{\theta})$. (b) gives a p-value for hypothesis (2.2) and a level 0.90 (two-sided) confidence interval for $g(\boldsymbol{\theta})$.

(a) 2016 September NCBS to 2017 March NCBS.

Week	Estimate	SE	$\mathcal{Z}$-statistic	p-value	CI Lo
1	0.0823	0.0602	1.3681	0.0856	0.0052
2	0.2605	0.0402	6.4731	4.802e-11	0.2090
3	0.1480	0.0411	3.6026	0.0002	0.0953
4	0.2273	0.0453	5.0166	2.629e-07	0.1693
5	-0.3376	0.0775	-4.3563	1.0000	-0.4369

(b) 2016 June NCBS to 2017 March NCBS.

Week	Estimate	SE	$\mathcal{Z}$-statistic	p-value	CI Lo
1	-0.0153	0.0631	-0.2426	0.5958	-0.0962
2	0.3463	0.0379	9.1467	2.935e-20	0.2977
3	0.3905	0.0442	8.8356	4.980e-19	0.3338
4	0.3523	0.0565	6.2409	2.175e-10	0.2800
5	0.0253	0.0640	0.3956	0.3462	-0.0567

(c) 2016 June NCBS to 2016 September NCBS.

Week	Estimate	SE	$\mathcal{Z}$-statistic	p-value	CI Lo	CI Hi
1	-0.0976	0.0451	-2.1630	0.0305	-0.1719	-0.0234
2	0.0857	0.0433	1.9782	0.0479	0.0144	0.1570
3	0.2425	0.0365	6.6499	2.934e-11	0.1825	0.3025
4	0.1250	0.0607	2.0602	0.0394	0.0252	0.2247
5	0.3629	0.0535	6.7823	1.183e-11	0.2749	0.4509

Table 10: 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 $\hat{\mathcal{E}}$ displays the entropy of the estimated probabilities.

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

Week	$\hat{\boldsymbol{p}}$(Week)							$\hat{\mathcal{E}}$
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	$\hat{\boldsymbol{q}}$(Week)							$\hat{\mathcal{E}}$
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	$\hat{\boldsymbol{p}}$(Week)							$\hat{\mathcal{E}}$
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	$\hat{\boldsymbol{q}}$(Week)							$\hat{\mathcal{E}}$
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	$\hat{\boldsymbol{p}}$(Week)							$\hat{\mathcal{E}}$
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	$\hat{\boldsymbol{q}}$(Week)							$\hat{\mathcal{E}}$
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