Estimating the Number of Street Vendors in New York City*

Jonathan Auerbach
Department of Statistics
George Mason University
jauerba@gmu.edu

Abstract

We estimate the number of street vendors in New York City. First, we summarize the process by which vendors receive licenses and permits to operate legally in New York City. Second, we describe a survey that was administered by the Street Vendor Project while distributing Coronavirus relief aid to vendors operating in New York City both with and without a license or permit. Third, we review ratio estimation and provide a theoretical justification based on the theory of point processes. Fourth, we use ratio estimation to calculate the total number of vendors, finding approximately 23,000 street vendors operate in New York City—20,500 mobile food vendors and 2,400 general merchandise vendors—with one third located in just six ZIP Codes—11368 (16%), 11372 (3%), and 11354 (3%) in North and West Queens and 10036 (5%), 10019 (4%), and 10001 (3%) in the Chelsea and Clinton neighborhoods of Manhattan. Finally, we evaluate the accuracy of the ratio estimator when the distribution of vendors is explained by a Poisson or Yule process, and we discuss several policy implications. In particular, our estimates suggest the American Community Survey misses the majority of New York City street vendors.

1. Introduction

Street vendors are New York City's smallest businesses, selling food and merchandise from carts, stalls, and trucks throughout the five boroughs. They are an iconic part of the urban landscape and a thriving sector of the local economy, contributing millions of dollars in government revenue through taxes, fines, and fees.[1] Perhaps most importantly, street vending historically benefits underserved communities, both because vendors operate in neighborhoods with limited access to traditional stores and because vending is one of a handful of occupations in which New Yorkers of all backgrounds, immigrants in particular, are able to achieve economic mobility and a chance at the American dream (Burrows and Wallace 1998, chap. 42).

Yet despite their social and economic importance, little is known about the size and location of New York City's street vending population. This is because while local law requires street vendors to obtain licenses and permits to operate legally, the number of licenses and permits are limited, resulting in a largely unknown population of vendors that operate without a license or permit. These vendors are not easily identified from administrative datasets, such as tax records or fines, and they can be difficult to locate for government

*Thanks to Mohamed Attia, Eric Auerbach, Debipriya Chatterjee, David Kallick, Carina Kaufman-Gutierrez, Joseph Salvo, Shamier Settle, Matthew Shapiro, Emerita Torres, and Anand Vidyashankar for discussion and feedback on the work described in this paper.

[1] Fees from mobile food vending licenses and permits provide nearly a million dollars in revenue to New York City each year (Mosher and Turnquist 2024). Fines paid by vendors provide approximately $200,000 each year according to data from the New York City Office of Administrative Trials and Hearings.

surveys, such as the American Community Survey. Nevertheless, understanding the size and location of New York's street vending population is crucial for informing policy and advocacy.

In this paper, we propose an estimate for the number of street vendors in New York City, including those that operate without a license or permit. We present our work in four sections. In Section 2, we review the process by which vendors receive permits and licenses, and we describe a survey administered by the Street Vendor Project at the Urban Justice Center while distributing Coronavirus relief aid to vendors operating both with and without a license or permit.

In Section 3, we use the fact that the number of licenses and permits are limited by law to construct a ratio estimator for the number of street vendors. We then provide a theoretical justification of our estimates, assuming the spatial distribution of survey respondents is representative and well-approximated by a family of inhomogeneous Poisson processes. We find approximately 23,000 street vendors operate in New York City: 20,500 mobile food vendors and 2,400 general merchandise vendors. One third are located in just six ZIP Codes: 11368 (16%), 11372 (3%), and 11354 (3%) in North and West Queens and 10036 (5%), 10019 (4%), and 10001 (3%) in the Chelsea and Clinton neighborhoods of Manhattan.

In Section 4, we examine the validity of these assumptions. We evaluate whether the spatial distribution of respondents is representative by comparing it to the American Community Survey and administrative data, and we evaluate the Poisson process assumption by considering a more general model in which the arrival of street vendors follows a pure birth process, of which the Poisson and Yule processes are special cases.

In Section 5, we discuss some implications of our estimates. In particular, we find that though the spatial distribution of respondents matches the results of the American Community Survey, the two disagree on the size of New York City's street vending population. Our estimates suggest the American Community Survey misses the majority of New York City street vendors.

2. Background and Survey Design

A street vendor is any individual who sells goods from a mobile vending unit instead of a store. We distinguish between the mobile vending unit from which the goods are sold (i.e., the vending establishment) and the individuals who own and/or operate that unit (i.e., the vendors). Note that an establishment refers to a single vending unit. There may be multiple vendors associated with any establishment, and multiple establishments may be owned by a single individual or firm.[2]

We limit our analysis to two types of street vendors: mobile food vendors—vendors who sell food items such as sandwiches, drinks, and cut fruit—and general merchandise vendors—vendors who sell merchandise items such as electronics, clothing, and accessories. We refer to these vendors as food vendors and merchandise vendors, respectively. In Section 2.1, we review the process by which vendors receive permits and licenses

[2] For example, suppose Sally owns the business Sally's Salads, which consists of two food trucks, each staffed by three employees. Then Sally's Salads counts as one business, two establishments, and seven vendors.

as required by law. In Section 2.2, we describe a survey administered by the Street Vendor Project while distributing Coronavirus relief aid to vendors operating in New York City both with and without a license or permit.

2.1 Background

A long list of rules determines the individuals, locations, and time periods during which vendors can legally sell their goods in New York City. Most relevant is the fact that a merchandise vendor requires a license to operate legally, and a food vendor requires both a license and a permit. The number of food permits and merchandise licenses are limited by law: 5,100 permits and 853 general merchandise licenses are available. Mobile food vending licenses are not limited. Note that a mobile food vending permit is issued to an individual or business to allow for the sale of food from a specific mobile food vending unit, such as a cart or truck. A mobile food vending license authorizes an individual to prepare or serve food from a permitted mobile food vending unit.[3]

There are several types of mobile food vendor permits with varying restrictions. Of the 5,100 permits available, 200 are borough permits that limit vendors to one of the five boroughs; 100 are reserved for veterans or vendors with a disability; 1,000 are seasonal and valid only from April to October; 1,000 are green cart permits that limit vendors to selling fruit, vegetables, plain nuts, and water; and 2,800 are unrestricted. Multiple individuals with food vendor licenses can legally operate from one permitted food vending unit. Merchandise licenses are renewed annually, while food licenses and permits are renewed biennially.

There are three relevant exceptions to these rules. The first exception is that merchandise vendors who are veterans are not subject to the limit of 853 licenses. According to data obtained by Mosher and Turnquist (2024), there are approximately 1,000 licensed merchandise vendors who are veterans in New York City. We assume the number of veterans selling merchandise without a license is negligible.

The second exception is First Amendment vendors. A First Amendment vendor sells expressive merchandise such as newspapers, books, and art. Expressive merchandise is considered free speech and protected under the First Amendment of the U.S. Constitution so that the number of First Amendment vendors cannot be restricted by law.

The final exception are concessionaires that operate on New York City parkland through an NYC Parks permit or license. According to the New York City Department of Parks and Recreation, there are approximately 400 concessions within New York City parks. Many offer food services, ranging from food carts to restaurants. We do not consider First Amendment vendors or park concessionaires in this paper. Our estimates are limited to food and merchandise vendors operating outside of parks.

[3] Note that our estimate reflects the number of street vendors as of 2021, prior to the implementation of Local Law 18 of 2021, which was enacted concurrently. Local Law 18 made several changes to the license and permit process. These changes do not impact our estimates and therefore are not discussed in this paper.

2.2 Survey Design

The Street Vendor Project (SVP), part of the Urban Justice Center, is a non-profit organization that advocates on behalf of New York City street vendors. SVP administered a survey to approximately 2,000 street vendors while distributing Coronavirus relief aid in 2021. The aid was a one-time payment of $1,000, available to any individual that owned, operated, or was otherwise employed by a street vending business in New York City between 2020 and 2021. First Amendment vendors were also eligible. There were no limits based on the size of the business, the number of sales, or whether licensed and/or permitted. All individuals eligible for aid were invited to complete a survey.

The population of street vendors estimated in this paper is the population eligible for aid (as determined by SVP) that self-identifies as either a food or merchandise vendor. SVP found aid-eligible individuals through membership lists, referrals, and canvassing operations in which SVP affiliates visited neighborhoods. Survey operations continued until $2,415,000 in funds were distributed, yielding 2,060 responses.

The survey included 100 questions and was conducted in eight different languages: Arabic, Bangla, Cantonese, English, French, Mandarin, Spanish, Tibetan, and Wolof.[4] The survey items solicited a variety of information from vendors, such as logistical information (e.g., vending location, residential location, and frequency of operation), economic information (e.g., items sold, income, and expenses), and demographic information (e.g., age, race, ethnicity, and immigration status). Most relevant is the fact that respondents classified themselves by the goods they sold (e.g., food vendors, merchandise vendors, First Amendment vendors, etc.), and respondents indicated whether they had the relevant licenses and permits to vend. Of the 2,060 responses, 1,400 identified as food vendors and 559 as merchandise vendors. The remaining 101 respondents were predominantly First Amendment Vendors, which we exclude from our analysis.

Of the 1,400 food vendors, 349 (25%) indicated they had a permit to vend. Of the 559 merchandise vendors, 505 were not veterans, of which 308 (61%) indicated they had a license to vend. The number of respondents is listed by neighborhood in Section 7.1, Table 1. (Neighborhoods with few respondents are grouped together.) A map of the number of respondents by ZIP Code Tabulation Area (ZIP Code) is shown in the top left panel of Section 7.2, Figure 1. Approximately five percent of vendors did not identify their vending location. These vendors are included in the New York City total, but they are excluded from the neighborhood estimates in Table 1 and Figure 1. (For this reason, the Respondent and Population columns do not sum exactly to the New York City total.) Veteran merchandise vendors are excluded entirely from Table 1 and Figure 1, although these vendors are reflected in our overall estimate of 23,000 vendors.

[4] Vendors who spoke other languages, such as Dari, Farsi, Russian, and Turkish, typically spoke English with a high level of fluency, and those surveys were conducted in English. Similarly, vendors who spoke indigenous languages from Latin America, such as Nahuatl or Quechua, often spoke Spanish with a high level of fluency.

3. Methodology and Data Analysis

We estimate approximately 23,000 street vendors operate in New York City: 20,500 mobile food vendors and 2,400 general merchandise vendors. We arrive at the estimate by using ratio estimation, which leverages the fact that the number of licenses and permits are limited by local law.

In Section 3.1, we provide a simple explanation of ratio estimation, which is intended to be accessible to a general readership. In Section 3.2, we provide a theoretical justification of our estimates based on the theory of point processes, which, while more technical, is intended to provide a somewhat general argument for ratio estimation with spatial data when a design-based approach cannot be justified. Inference is conducted using increasing domain asymptotics, and the details are outlined in Section 7.3. In Section 3.3, we apply the methodology to the survey data discussed in Section 2.2.

3.1 Ratio Estimation

Ratio estimation is a common approach for estimating the size of a population. See Cochran (1978) and Hald (1998, chap. 16) for a historical discussion and Lohr (2021, chap. 4) for an introduction. We provide a simple explanation based on cross-multiplication, also called the rule of three. The purpose is to highlight the main assumption.

Consider a fixed region A, and let $N_i(A)$ denote the number of individuals in region A of type i. To fix ideas, let A be New York City, let $i = 0$ indicate mobile food vendors with a permit, and let $i = 1$ indicate mobile food vendors without a permit. Also let $n_i(A)$ denote the number of respondents of type i.

We observe $N_1(A)$, $n_1(A)$, and $n_0(A)$ from which we estimate $\theta(A) = \mathbb{E}\left[N_0(A)\right]$, the expected number of vendors without a permit. The expected number of mobile food vendors with or without a permit is then $\tau(A) = \theta(A) + N_1(A)$.

The ratio estimator of $\theta(A)$ is

$$\hat{\theta}(A) = \frac{N_1(A)\, n_0(A)}{n_1(A)} \quad .$$

It is called the ratio estimator because it depends on the ratio of random variables $n_0(A) \,/\, n_1(A)$. (The total number of permitted vendors, $N_1(A)$, is considered fixed.) The estimator is accurate when the response rate, p, is approximately the same for each type, i.e.,

$$p \approx \frac{n_0(A)}{N_0(A)} \approx \frac{n_1(A)}{N_1(A)} \quad ,$$

in which case $\hat{\theta}(A)$ is the solution for $N_0(A)$ in terms of $N_1(A)$, $n_1(A)$, and $n_0(A)$, obtained by cross-multiplication.

The main assumption of ratio estimation is that the survey data are representative in the sense that the response rate does not depend on whether a mobile food vendor has a permit. If we further assume that the response rate does not vary by subregion, we can also estimate the total number of mobile food vendors by subregion. For example, suppose the response rate in subregion $B \subset A$ is also approximately equal to p, e.g.,

$$p \approx \frac{n_1(A)}{N_1(A)} \approx \frac{n_0(B)}{N_0(B)} \qquad .$$

Then we can estimate $\theta(B)$ by the subregion estimator

$$\tilde{\theta}(B) = \frac{N_1(A)\, n_0(B)}{n_1(A)} \qquad ,$$

where $\tilde{\theta}(B)$ is the solution for $N_0(B)$ in terms of $N_1(A)$, $n_1(A)$, and $n_0(B)$.

The subregion estimator is necessary when $N_1(B)$ is not observed and $\hat{\theta}(B)$ cannot be calculated directly. This is the case with the survey data described in Section 2.2. We observe $N_1(A)$, the number of permits city-wide (i.e., within region A), but not $N_1(B)$, the number of permits within subregion B. (We also observe $n_0(B)$ and $n_1(B)$.)

The subregion estimator $\tilde{\theta}(B)$ works by replacing the term $N_1(B)$ in the ratio estimator $\hat{\theta}(B)$ with the estimate

$$\tilde{N}_1(B) = \frac{N_1(A)\, n_1(B)}{n_1(A)} \qquad .$$

The estimate $\tilde{N}_1(B)$ is also used to estimate the total number of mobile food vendors expected in subregion B. That is, we estimate $\tau(B) = \theta(B) + N_1(B)$ with

$$\tilde{\tau}(B) = \tilde{\theta}(B) + \tilde{N}_1(B) = \frac{N_1(A)\ (n_0(B) + n_1(B))}{n_1(A)} \qquad .$$

3.2 Model

We provide a point process justification of the ratio estimation procedure described in Section 3.1. The purpose is to more closely examine the underlying assumptions and derive the standard errors of the estimates.

Let $\{\Pi_i\}$ denote a family of inhomogeneous spatial Poisson processes referencing the location $x \in A \subset \mathbb{R}^2$ of each street vendor at the time the survey was conducted. The index i denotes the status of the vendor. As in Section 3.1, we fix ideas by letting $i = 0$ indicate mobile food vendors with a permit and $i = 1$ indicate mobile food vendors without a permit.

The Poisson process assumption may be justified by the law of rare events. New York City can be partitioned into a large number of theoretically vendable locations. Whether a vendor is located within a partition has a vanishingly small probability, such that the number of vendors in any area is well approximated by a Poisson distribution. This approximation is accurate even if these probabilities are weakly dependent. For example, see Freedman (1974), Chen (1975), and Serfling (1975). Other justifications are possible. For example, see

Section 4.2 for a justification of the Poisson process based on the theory of birth processes.

Let $\{\lambda_i\}$ denote the mean measures of $\{\Pi_i\}$ so that the number of individuals in any subregion $B \subseteq A$, $N_i(B)$, is distributed Poisson with mean $\int_B \lambda_i(x)\,dx$, i.e.,

$$N_i(B) \sim \text{Poisson}\left(\int_B \lambda_i(x)\,dx\right) \qquad .$$

We assume each individual responds independently to the survey with probability $p_i(x)$ such that by Campbell's Theorem (Kingman 1992), the number of respondents $n_i(B)$ is distributed Poisson with mean $\int_B \lambda_i(x)\,p_i(x)\,dx$. i.e.,

$$n_i(B) \sim \text{Poisson}\left(\int_B \lambda_i(x)\,p_i(x)\,dx\right) \qquad .$$

To determine the number of street vendors in region $B \subseteq A$, we estimate $\theta(B) = \mathbb{E}\left[N_0(B)\right] = \int_B \lambda_0(x)\,dx$, the expected number of vendors without permits. The total expected in region B is then $\tau(B) = N_1(B) + \theta(B)$.

In Section 3.1, we described the main assumption of ratio estimation: that the status and spatial distribution of the respondents is representative. We now make this statement precise. We assume that $p_i(x)$ can be decomposed into a constant that does not depend on i or x plus a spatially varying error term, $p_i(x) = p + \epsilon_i(x)$, and the error term is orthogonal to the corresponding mean measure. i.e.,

$$\langle \lambda_i, \epsilon_i \rangle = \int_B \lambda_i(x)\,\epsilon_i(x)\,dx = 0 \qquad .$$

It follows that

$$n_i(B) \sim \text{Poisson}\left(p \int_B \lambda_i(x)\,dx\right)$$

so that by conditioning on $N_1(B)$, we arrive at the following probability model for the number of respondents with and without permits

$$\begin{aligned} n_0(B) &\sim \text{Poisson}\big(p\,\theta(B)\big) \\ n_1(B) \mid N_1(B) &\sim \text{Binomial}\big(p,\, N_1(B)\big) \end{aligned} \qquad .$$

The maximum likelihood estimates for p and $\theta(A)$ are

$$\hat{p} = \frac{n_1(A)}{N_1(A)} \qquad \text{and} \qquad \hat{\theta}(A) = \frac{N_1(A)\,n_0(A)}{n_1(A)} \qquad .$$

Using increasing domain asymptotics, the estimator is asymptotically normal with mean $\theta(A)$ and standard error

$$\text{SE}[\hat{\theta}(A)] = \theta(A)\sqrt{\frac{1}{p\,\theta(A)} + \frac{1-p}{p\,N_1(A)}} \qquad .$$

Substituting $\hat{p}$ and $\hat{\theta}(A)$ for p and $\theta(A)$ yields the following plug-in estimator for the standard error,

$$\widehat{\mathrm{SE}}[\hat{\theta}(A)] = \frac{N_1(A)\, n_0(A)}{n_1(A)} \sqrt{\frac{1}{n_0(A)} + \frac{1}{n_1(A)} - \frac{1}{N_1(A)}} \quad .$$

See Section 7.3.2 for details.

The maximum likelihood estimate for $\theta(B)$ is

$$\tilde{\theta}(B) = \frac{N_1(A)\, n_0(B)}{n_1(A)} \quad ,$$

which is asymptotically normal with mean $\theta(B)$ and standard error

$$\mathrm{SE}[\tilde{\theta}(B)] = \theta(B) \sqrt{\frac{1}{p\,\theta(B)} + \frac{1-p}{p\,N_1(A)}} \quad .$$

The plug-in estimator for the standard error is

$$\widetilde{\mathrm{SE}}[\tilde{\theta}(B)] = \frac{N_1(A)\, n_0(B)}{n_1(A)} \sqrt{\frac{1}{n_0(B)} + \frac{1}{n_1(A)} - \frac{1}{N_1(A)}} \quad .$$

To determine the subregion total $\tau(B) = \theta(B) + N_1(B)$, we estimate $\theta(B)$ and $N_1(B)$ with $\tilde{\theta}(B)$ and $\tilde{N}_1(B) = N_1(A)\, n_1(B) \,/\, n_1(A)$, yielding

$$\tilde{\tau}(B) = \tilde{\theta}(B) + \tilde{N}_1(B) = \frac{N_1(A)\ (n_0(B) + n_1(B))}{n_1(A)} = \frac{N_1(A)\ (n_0(B) + n_1(B))}{n_1(B) + n_1(A \backslash B)}$$

where $A \backslash B$ denotes the compliment of subregion B.

The asymptotic standard error is

$$\mathrm{SE}[\tilde{\tau}(B)] = \theta(B) \sqrt{\frac{1}{p\,\theta(B)} + \frac{1-p}{p\,N_1(A)} \left(1 + \frac{N_1(B)\, N_1(A \backslash B)}{\theta(B)^2}\right)} \quad ,$$

and the plug-in estimator is

$$\widetilde{\mathrm{SE}}[\tilde{\theta}(B)] = \frac{N_1(A)\, n_0(B)}{n_1(A)} \sqrt{\frac{1}{n_0(B)} + \left(\frac{1}{n_1(A)} - \frac{1}{N_1(A)}\right) \left(1 + \frac{n_1(B)\, n_1(A \backslash B)}{n_0(B)^2}\right)} \quad .$$

See Section 7.3.3 for details.

3.3 Data Analysis

We use the methodology described in Section 3.1 and 3.2 to estimate the number of street vendors in New York City. According to the survey data described in Section 2.2, $n_1(A) = 349$ mobile food vendor respondents indicated they had a permit, and $n_0(A) = 1,051$ indicated they did not. Since there are $N_1(A) = 5,100$ permits available, it follows that the number of vendors without permits is $15,350 \approx 5,100 \times 1,051\,/\,349$. Combined with the $5,100$ permitted vendors, the total number of mobile food vendors is approximately $20,500$.

By similar reasoning—i.e., exploiting the fact that the number of general merchandise licenses is limited for non-veterans—the number of non-veteran merchandise vendors is estimated to be approximately $1,400$. When combined with the estimated $1,000$ veteran merchandise vendors (see Section 2.1) and $20,500$ food vendors, the total number of street vendors is approximately $23,000$.

Estimates for select subregions are listed in Section 7.1, Table 1. We find that approximately one quarter of street vendors are located in the Manhattan neighborhoods of Chelsea, Clinton, and Lower Manhattan. Another quarter is located in West and North Queens. We also report the margin of error (i.e., two standard errors, half of the width of a 95% confidence interval).

We check these estimates by comparing them to a second, independent assessment conducted by the Street Vendor Project in the Bronx on May 13 and 15, 2022, in which the number of street vendors was documented through on-the-ground observations. The independent assessment identified 188 total street vendors near Fordham Road (Fordham Road BID and Street Vendor Project 2024). According to the survey data, 17 respondents indicated that they operate near Fordham Road, yielding an estimate of 152 street vendors total with a standard error of 54. We conclude this estimate is consistent with the 188 street vendors counted independently by Fordham Road BID and Street Vendor Project (2024).

4. Model Validation

We make two assumptions in Section 3. The main assumption is that the status and spatial distribution of the respondents is representative. This assumption is necessary to ensure the ratio estimator is accurate in large samples (i.e., consistent). A secondary assumption is that the spatial distribution of vendors is well described by a family of inhomogeneous spatial Poisson processes. This assumption is necessary to derive the standard errors.

In this section, we examine each assumption. Since neither can be evaluated from the survey data alone, we consider additional information. In Section 4.1, we compare the spatial distribution of the respondents to data from the New York City Office of Administrative Trials and Hearings and the American Community Survey. In Section 4.2, we model the arrival of street vendors as a pure birth process, and we derive a more general formula for the standard errors in which the amount of extra-Poisson variation can be determined by the number of vendors that cluster within markets.[5]

4.1 Representativeness

The main assumption in Section 3 is that the status and spatial distribution of the respondents is representative (i.e., the response probability is spatially uncorrelated with the mean measure). If this assumption holds, then the ratio estimator is consistent for a wider class of point processes in which Campbell's Theorem applies, not just the Poisson process. See Daley and Vere-Jones (2007) for a general statement of Campbell's Theorem.

If this assumption is violated, then $\hat{\theta}(A)$ may be inconsistent. For example, if respondents are more likely to come from locations in which a higher proportion of vendors have permits, then the ratio $n_0(A) \,/\, n_1(A)$ would not be representative, and $\hat{\theta}(A)$ would likely underestimate $\theta(A)$ even in large samples.

We think this assumption is reasonable because the Street Vendor Project canvassed New York City to distribute a large amount of relief aid. We argue that trust in SVP and its mission, along with the aid, removed many barriers that typically prevent survey operations from reaching hard-to-reach populations. Instead, nonresponse reflects chance variation in the personal circumstances of a vendor that are largely unrelated to location.

Nevertheless, it is possible that canvassing operations systematically missed locations, and these missed locations may be revealed by examining other data sources. We thus compare the spatial distribution of vending locations reported by respondents to the spatial distribution of vendors who allegedly violated street vending laws during the year 2021. We consider the location of an individual's first violation in 2021, reported in administrative data from the NYC Office of Administrative Trials and Hearings (OATH) (accessed 2024-04-03). We also compare the residencies reported by the respondents to the residencies of Door-To-Door Sales Workers, News And Street Vendors, And Related Workers (Standard Occupational Classification 41-9091) estimated in the 2017-2021 American Community Survey (ACS) Public Use Micro Sample from the U.S. Census Bureau (accessed 2022-12-22).

These comparisons are visualized in Section 7.2, Figure 1. We find the spatial distribution of vending locations largely agree (top panels) as do the spatial distribution of the residencies (bottom panels), although there are several small discrepancies. For example, the OATH data suggest a larger percentage of vendors work in South Brooklyn (Coney Island and Brighton Beach) than captured by the Street Vendor Project survey, and the ACS data suggest a larger percentage of street vendors live in South Queens (Rockaway). However, both the OATH and ACS data suggest these locations reflect only a small percentage of New York City vendors, and thus any resulting undercount is likely to have little impact on our overall estimate.

We also note that even if this discrepancy suggests a location was missed by the Street Vendor Project,

[5] The results of this section can also be derived from a spatial Negative Binomial Lévy Process in a manner that mirrors Section 3.2. The advantage of the birth process approach taken in Section 4.2 is that it provides a plausible explanation for why the Poisson assumption made in Section 3.2 may not hold in practice.

our main assumption is not necessarily violated. That is because the undercount in some areas may be compensated by an overcount in others (e.g., the Upper East Side of Manhattan or East Brooklyn) such that our overall estimate is still accurate.

4.2 Overdispersion

The second assumption in Section 3.2 is that the spatial distribution of vendors is well approximated by a family of inhomogeneous Poisson processes. We consider this assumption secondary because the consistency of $\hat{\theta}(A)$ holds under a wide class of point processes described in Section 4.1. That said, the standard errors derived in Section 3.2 do not necessarily hold when the Poisson assumption is violated.[6]

We think the Poisson assumption is reasonable because a wide class of data generating processes are well-approximated by the Poisson process under the law of rare events as discussed in Section 3.2. Nevertheless, to evaluate how the standard errors might change when this assumption is violated, we consider a general model that describes how street vendors arrived at their present locations. Specifically, we consider two families of pure birth processes: one with a constant arrival rate (or "birth rate") in which the Poisson assumption holds and one with a linear arrival rate in which the Poisson assumption does not hold.

Suppose vendors serve customers who congregate at fixed locations, which we call markets. Let $M(B)$ denote the number of markets in any subregion $B \subseteq A \subset \mathbb{R}^2$. If every market is in a competitive equilibrium such that the expected profit for a new vendor is the same at each market, then the time between vendor arrivals at each market may be well-approximated by an exponential distribution with a constant arrival rate λ_i.[7] Further suppose each market starts with one vendor at $t = 0$. Then at the time the survey was conducted ($t = 1$), the number of vendors, $N_i(B)$, would follow a Poisson distribution with rate $\int_B \lambda_i(x)\, dx = \lambda_i\, M(B)$, and the results of Section 3.2 hold.

Now suppose instead that larger markets grow faster than smaller markets, an empirical phenomenon known as preferential attachment. Then the time between vendor arrivals may be better approximated by an exponential distribution with a linear arrival rate, or Yule process (Yule 1925). That is, let $N_i^t(B)$ denote the number of vendors in subregion B at time t so that the time between vendor arrivals at each market follows an exponential distribution with rate $\lambda_i\, N_i^t(B)$. If we assume further that each market started with one vendor at $t = 0$, then the number of vendors at the time the survey was conducted ($t = 1$), $N_i(B)$, would follow a negative binomial distribution with mean $\mu_i(B) = \exp(\lambda_i)\, M(B)$ and variance $\mu_i(B)\big(\mu_i(B) - M(B)\big) \,/\, M(B)$. i.e.,

$$N_i(B) \sim \text{ Negative Binomial}\left(\mu_i(B),\ \mu_i(B)\, \frac{\mu_i(B) - M(B)}{M(B)}\right)$$

where we parameterize the negative binomial distribution by its mean and variance. In this case, $N_i(B)$

[6] The ratio estimator may no longer be asymptotically efficient as well. In Section 7.3, we show that under the Poisson process assumption, the ratio estimator maximizes the likelihood, which implies the ratio estimator is asymptotically efficient. However, the ratio estimator may not maximize the likelihood when the Poisson process assumption is violated.

[7] Arriving vendors may form a new vending establishment or sell their labor to an existing establishment.

exhibits extra-Poisson variation, and the results of Section 3.2 do not hold. See Ross (2014, chap. 6) for details.[8]

To derive the standard errors, suppose a portion of the markets, p, were selected for the survey so that

$$n_i(B) \sim \text{ Negative Binomial}\left(p\,\mu_i(B),\ p\,\mu_i(B)\,\frac{\mu_i(B) - M(B)}{M(B)}\right) \qquad .$$

The conditional distribution of $n_1(B)$ given $N_1(B)$ is then

$$n_1(B) \mid N_1(B) \sim \text{ Negative Hypergeometric}\big(N_1(B) - 1,\ M(B) - 1,\ p\,M(B)\big)$$

with conditional mean and variance

$$\mathbb{E}\left[n_1(B) \mid N_1(B)\right] = p\,N_1(B) \qquad \text{and} \qquad \mathbb{V}\big(n_1(B) \mid N_1(B)\big) = w_1\,p\,(1-p)\,N_1(B)$$

so that under increasing domain asymptotics, the ratio estimator

$$\hat{\theta}(A) = \frac{N_1(A)\ n_0(A)}{n_1(A)}$$

is asymptotically normal with mean $\theta(A) = \mu_0(A)$ and standard error

$$\text{SE}[\hat{\theta}(A)] = \theta(A)\sqrt{w_0\left(\frac{1}{p\,\theta(A)}\right) + w_1\left(\frac{1-p}{p\,N_1(A)}\right)}$$

where

$$w_0 = \frac{\theta(A) - M(A)}{M(A)} \qquad \text{and} \qquad w_1 = \frac{N_1(A) - M(A)}{M(A) + 1} \qquad .$$

See Section 7.3.4 for details. Note that the weights w_0 and w_1 depend on A, although this dependence is not reflected in the notation.

The standard error derived in this section for $\hat{\theta}(A)$ is identical to the standard error for $\hat{\theta}(A)$ from Section 3.2 except for the weights w_0 and w_1. These weights are the result of the extra-Poisson variation induced by preferential attachment—that is, the linear "birth" rate in which the arrival of new vendors is proportional to the number of vendors in the market.

The weights w_0 and w_1 correspond to the relative number of vendors without permits and with permits per market respectively. These weights cannot be estimated from the survey data alone. However, we believe that $\theta(A)$ and $N_1(A)$ are no more than 10 times larger than $M(A)$. It follows that w_0 and w_1 are less than

[8] Note that we define the negative binomial and negative hypergeometric distributions as the total number of trials required to get a predetermined number of successes. The negative binomial assumes sampling with replacement, while the negative hypergeometric distribution assumes sampling without replacement.

9, and thus the standard errors from Section 3.2 and the margins of error from Table 1 are at most be 3 times larger if the Poisson assumption is violated and the Yule process holds. We suspect this bound is conservative, however, and the typical market only contains one or two vendors.

The same behavior is exhibited by the subregion estimator of $\theta(B)$,

$$\tilde{\theta}(B) = \frac{N_1(A)\; n_0(B)}{n_1(A)} ,$$

which is asymptotically normal with mean $\theta(B) = \mu_0(B)$ and standard error

$$\mathrm{SE}[\tilde{\theta}(B)] = \theta(B) \sqrt{v_0 \left(\frac{1}{p\, \theta(B)} \right) + w_1 \left(\frac{1-p}{p\, N_1(A)} \right)} .$$

where

$$v_0 = \frac{\theta(B) - M(B)}{M(B)} .$$

The estimator for the subregion total, $\tilde{\tau}(B) = \tilde{\theta}(B) + \tilde{N}_1(B)$, is also asymptotically normal with mean $\tau(B)$ and standard error

$$\mathrm{SE}[\tilde{\tau}(B)] = \theta(B) \sqrt{v_0 \left(\frac{1}{p\, \theta(B)} \right) + \left(\frac{v_1\, N_1(B)\, (N_1(A \backslash B) - \theta(B))^2 + v_2\, N_1(A \backslash B)\, (N_1(B) + \theta(B))^2}{N_1(A)\, \theta(B)^2} \right) \left(\frac{1-p}{p\, N_1(A)} \right)}$$

where

$$v_1 = \frac{N_1(B) - M(B)}{M(B) + 1} \qquad \text{and} \qquad v_2 = \frac{N_1(A \backslash B) - M(A \backslash B)}{M(A \backslash B) + 1} .$$

See Section 7.3.4 for details.

5. Discussion

We conclude by discussing the implications of our estimates. In particular, we assess the completeness of the New York City Office of Administrative Trials and Hearings (OATH) and American Community Survey (ACS) data. Our estimates suggest that while both data sets are spatially representative, they miss the majority of street vendors. We provide several explanations for this discrepancy.

We estimate 23,000 vendors work in New York City. In comparison, the 2017-2021 ACS estimates 4,634 individuals work the occupation of Door-To-Door Sales Workers, News And Street Vendors, And Related Workers in New York City. This suggests that according to the ACS, there are at most 4,634 street vendors in New York City, and thus our finding indicates that the ACS estimate misses 80% of street vendors or more.

A fairer comparison might account for the fact that the standard error of our estimate is approximately 1,000, and the standard error of the ACS estimate is approximately 488, which we obtained from the replicate weights. Using the lower limit of a 95% confidence interval for our estimate (21,000) and the upper 95%

confidence limit for the ACS estimate (5,610), our findings suggest the ACS misses roughly three quarters of street vendors or more.

Note the importance of the assumptions stated in Section 3.2 since other derivations of ratio estimation may imply smaller or larger standard errors, from which we might conclude the ACS misses a smaller or larger portion of the street vending population. For example, if the standard errors from Section 4.2 hold, then the confidence intervals would overlap if there were more than 300 vendors in the average market. We believe this is highly unlikely, suggesting the two estimates are inconsistent even under overdispersion.

Additional evidence comes from the OATH data, which indicate approximately 2,500 (unique) vendors are summoned to court each year for violating vending laws. If the ACS estimate holds, then more than half of vendors are summoned to court each year. If our estimate holds, then approximately one in ten vendors are summoned to court each year. We believe the latter is more realistic.

There may be several reasons why our estimates are significantly larger than the ACS. The fact that the Street Vendor Project distributed relief aid may have encouraged respondents who are unlikely to respond to the ACS. Moreover, the Street Vendor Project, through membership lists, referrals, and canvassing may have covered individuals missed by the Census Bureau's sampling frame.

Another explanation is that the Street Vendor Project survey and the ACS may be measuring different populations or concepts. For example, the ACS estimate reflects the average number of vendors between 2017 and 2021, while our estimate reflects 2020 and 2021. For another example, the Census Bureau determines the occupation of an ACS respondent by autocoding responses to write-in questions. It is possible that the Street Vendor Project's definition of a vendor is more inclusive. Street vending often provides supplemental income, and thus the discrepancy may reflect the broader challenge of studying the gig economy. In this case, either estimate may be relevant, depending on the intended use case.

6. References

Burrows, Edwin G, and Mike Wallace. 1998. *Gotham: A History of New York City to 1898.* Oxford University Press. https://isbnsearch.org/isbn/9780195116342.

Chen, Louis HY. 1975. "Poisson Approximation for Dependent Trials." *The Annals of Probability* 3 (3): 534–45. https://doi.org/10.1214/aop/1176996359.

Cochran, William G. 1978. "Laplace's Ratio Estimator." In *Contributions to Survey Sampling and Applied Statistics*, 3–10. Elsevier. https://doi.org/10.1016/B978-0-12-204750-3.50008-3.

Daley, Daryl J, and David Vere-Jones. 2007. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure.* Springer Science & Business Media. https://doi.org/10.1007/978-0-387-49835-5.

Fordham Road BID, and Street Vendor Project. 2024. *Fordham, the Bronx: Commercial District Needs Assessments.* New York City Department of Small Business Services. https://www.nyc.gov/assets/sbs/downloads/pdf/neighborhoods/avenyc-cdna-fordham.pdf.

Freedman, David. 1974. "The Poisson Approximation for Dependent Events." *The Annals of Probability*, 256–69. https://doi.org/10.1214/aop/1176996707.

Hald, Anders. 1998. "A History of Mathematical Statistics from 1750 to 1930." https://isbnsearch.org/isbn/9780471179122.

Kingman, John Frank Charles. 1992. *Poisson Processes.* Vol. 3. Clarendon Press. https://isbnsearch.org/isbn/9780198536932.

Lohr, Sharon L. 2021. *Sampling: Design and Analysis.* Chapman; Hall/CRC. https://isbnsearch.org/isbn/9780495105275.

Mosher, Eric, and Alaina Turnquist. 2024. *Fiscal Impact of Eliminating Street Vendor Permit Caps in New York City.* New York City Independent Budget Office. https://ibo.nyc.ny.us/iboreports/Fiscal_Impact_of_Eliminating_Street_Vendor_Permit_Caps_Jan2024.pdf.

NYC Office of Administrative Trials and Hearings (OATH). accessed 2024-04-03. *OATH Hearings Division Case Status.* https://data.cityofnewyork.us/City-Government/OATH-Hearings-Division-Case-Status/jz4z-kudi.

Ross, Sheldon M. 2014. *Introduction to Probability Models.* Academic press. https://isbnsearch.org/isbn/9780124079489.

Serfling, Robert J. 1975. "A General Poisson Approximation Theorem." *The Annals of Probability*, 726–31. https://doi.org/10.1214/aop/1176996313.

U.S. Census Bureau. accessed 2022-12-22. *2017-2021 American Community Survey 5-Year Public Use Microdata Samples.* https://www2.census.gov/programs-surveys/acs/data/pums/.

Yule, George Udny. 1925. "II.—a Mathematical Theory of Evolution, Based on the Conclusions of Dr. JC Willis, FR s." *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character* 213 (402-410): 21–87. https://doi.org/10.1098/rstb.1925.0002.

7. Appendix

7.1 Tables

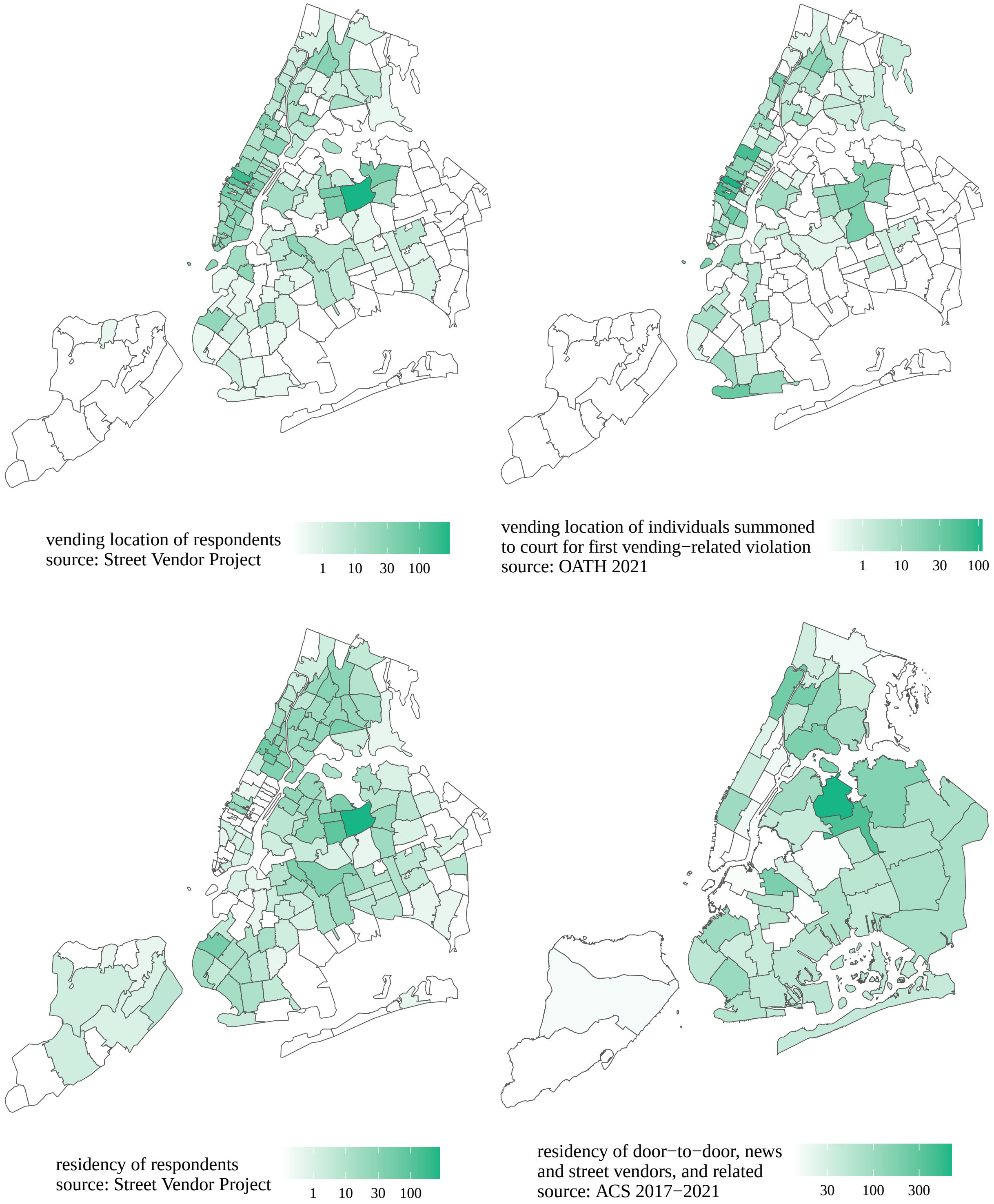
Table 1: Number of Street Vendors[1]

	Respondents	Population	Margin of Error
Bronx			
Bronx Park and Fordham	68	757	237
Southeast Bronx	28	385	159
High Bridge and Morrisania	20	233	126
Central Bronx	9	96	81
Other Bronx	31	264	133
Total	**156**	**1,735**	**373**
Brooklyn			
Northwest Brooklyn	46	613	202
Bushwick and Williamsburg	41	433	174
Sunset Park	26	344	152
Flatbush	19	242	126
Borough Park	10	134	93
Southwest Brooklyn	6	64	65
Other Brooklyn	28	374	158
Total	**176**	**2,205**	**415**
Manhattan			
Chelsea and Clinton	341	3,289	477
Lower Manhattan	105	1,250	278
Gramercy Park and Murray Hill	97	1,062	266
Lower East Side	65	808	226
Greenwich Village and Soho	61	726	216
Upper East Side	56	534	187
Central Harlem	46	483	181
Inwood and Washington Heights	33	459	175
Upper West Side	48	429	165
East Harlem	38	425	171
Total	**890**	**9,464**	**872**
Queens			
West Queens	396	4,650	690
North Queens	67	896	248
Northwest Queens	23	312	143
Jamaica	14	181	110
Other Queens	31	406	167
Total	**531**	**6,445**	**858**
Staten Island			
Total	**4**	**58**	**57**
New York City			
Total	**1,905**	**21,857**	**1,941**

[1] This table provides estimates for the number of mobile food vendors and non-veteran general merchandise vendors. The number of veteran general merchandise vendors is approximately 1,000. First amendment vendors and vendors in NYC Parks are not included.

7.2 Figures

Figure 1: Location of Street Vendors


vending location of respondents
source: Street Vendor Project
1
10
30
100
vending location of individuals summoned
to court for first vending−related violation
source: OATH 2021
1
10
30
100
residency of respondents
source: Street Vendor Project
1
10
30
100
residency of door−to−door, news
and street vendors, and related
source: ACS 2017−2021
30
100
300

These maps compare the spatial distribution of respondents (left) to the distribution of vendors suggested by administrative data (New York City Office of Administrative Trials and Hearings (OATH), top right) and federal data (American Community Survey (ACS), bottom right).

7.3 Estimation and Standard Errors

7.3.1 Preliminaries We calculate estimates and standard errors using increasing domain asymptotics. That is, we partition the region $A \subset \mathbb{R}^2$ into I disjoint sets, $\{A_i\}_{i=1}^I$, such that $A = \cup_{i=1}^I A_i$, and we consider the behavior of the model

$$\begin{aligned} n_0(A_i) &\sim \text{Poisson}\big(p\,\theta(A_i)\big) \\ n_1(A_i) \mid N_1(A_i) &\sim \text{Binomial}\big(p,\, N_1(A_i)\big) \end{aligned}$$

as $I \to \infty$.

Let $n_0(A) = \sum_{i=1}^I n_0(A_i)$, $n_1(A) = \sum_{i=1}^I n_1(A_i)$, and $N_1(A) = \sum_{i=1}^I N_1(A_i)$ so that the likelihood can be written

$$\begin{aligned} \mathcal{L}_I(\{\theta(A_i)\}_{i=1}^I,\, p) &= \prod_{i=1}^I \exp(-p\;\theta(A_i))\;(p\;\theta(A_i))^{n_0(A_i)}\;\frac{1}{n_0(A_i)!}\binom{N_1(A_i)}{n_1(A_i)}\;p^{n_1(A_i)}\;(1-p)^{N_1(A_i)-n_1(A_i)} \\ &= p^{n_1(A)+n_0(A)}\;(1-p)^{N_1(A)-n_1(A)}\;\prod_{i=1}^I \exp(-p\;\theta(A_i))\;\theta(A_i)^{n_0(A_i)}\;\frac{1}{n_0(A_i)!}\binom{N_1(A_i)}{n_1(A_i)}\quad. \end{aligned}$$

The maximum likelihood estimates, obtained from solving the score function

$$0 \overset{\text{set}}{=} \nabla \log \mathcal{L}_I(\{\theta(A_i)\}_{i=1}^I,\, p) = \begin{cases} \dfrac{\partial}{\partial\theta(A_i)} \log \mathcal{L}_I(\{\theta(A_i)\}_{i=1}^I,\, p) = -p + \dfrac{n_0(A_i)}{\theta(A_i)} \\ \\ \dfrac{\partial}{\partial p} \log \mathcal{L}_I(\{\theta(A_i)\}_{i=1}^I,\, p) = \dfrac{n_0(A)+n_1(A)}{p} - \dfrac{N_1(A)-n_1(A)}{1-p} - \sum_{i=1}^I \theta(A_i) \end{cases}$$

are

$$\hat\theta(A_i) = \frac{N_1(A)\;n_0(A_i)}{n_1(A)} \qquad \text{and} \qquad \hat p = \frac{n_1(A)}{N_1(A)} \quad.$$

7.3.2 Ratio Estimator Under the conditions of Section 7.3.1, the maximum likelihood estimate of $\theta(A) = \sum_{i=1}^I \theta(A_i)$ is the ratio estimator

$$\hat\theta(A) = \sum_{i=1}^I \hat\theta(A_i) = \frac{N_1(A)\;n_0(A)}{n_1(A)} \quad.$$

The asymptotic standard error of $\hat\theta(A)$ can be obtained by noting that under the Lyapunov Central Limit Theorem,

$$\begin{bmatrix} p\,\theta(A) & 0 \\ 0 & p\,(1-p)\,N_1(A) \end{bmatrix}^{-1/2} \begin{bmatrix} n_0(A) - p\,\theta(A) \\ n_1(A) - p\,N_1(A) \end{bmatrix}$$

converges to a standard bivariate normal distribution as $I \to \infty$.

It follows from the Delta Method that when I is large,

$$\hat\theta(A) \mathrel{\dot\sim} \text{Normal}\left(\theta(A),\; \begin{bmatrix} p^{-1} & -\dfrac{\theta(A)}{p\,N_1(A)} \end{bmatrix} \begin{bmatrix} p\,\theta(A) & 0 \\ 0 & p\,(1-p)\,N_1(A) \end{bmatrix} \begin{bmatrix} p^{-1} \\ -\dfrac{\theta(A)}{p\,N_1(A)} \end{bmatrix}\right) \quad.$$

The asymptotic standard error simplifies to

$$\mathrm{SE}[\hat{\theta}(A)] = \theta(A)\sqrt{\frac{1}{p\,\theta(A)} + \frac{1-p}{p\,N_1(A)}} \quad .$$

Substituting $\hat{\theta}(A)$ and $\hat{p}$ for $\theta(A)$ and p yields a plug-in estimate of the standard error

$$\hat{\mathrm{SE}}[\hat{\theta}(A)] = \frac{N_1(A)\,n_0(A)}{n_1(A)}\sqrt{\frac{1}{n_0(A)} + \frac{1}{n_1(A)} - \frac{1}{N_1(A)}} \quad .$$

7.3.3 Subregion Estimator Now consider a subregion $B \subset A$ that can be written as the union of a subsequence of partition elements $\{A_{k_j}\}_{j=1}^J$, such that $B = \cup_{j=1}^J A_{k_j}$ and k_j is strictly increasing in j. Under the conditions of Section 7.3.1, the maximum likelihood estimate of $\theta(B) = \sum_{j=1}^J \theta(A_{k_j})$ is the subregion estimator

$$\tilde{\theta}(B) = \sum_{j=1}^J \hat{\theta}(A_{k_j}) = \frac{N_1(A)\;n_0(B)}{n_1(A)} \quad .$$

The asymptotic standard error of $\tilde{\theta}(B)$ can be obtained as in Section 7.3.2 by noting that under the Lyapunov Central Limit Theorem,

$$\begin{bmatrix} p\,\theta(B) & 0 \\ 0 & p\,(1-p)\,N_1(A) \end{bmatrix}^{-1/2} \begin{bmatrix} n_0(B) - p\,\theta(B) \\ n_1(A) - p\,N_1(A) \end{bmatrix}$$

converges to a standard bivariate normal distribution as $J \to \infty$.

It follows from the Delta Method that when J is large,

$$\tilde{\theta}(B) \mathrel{\dot\sim} \text{Normal}\left(\theta(B),\;\; \begin{bmatrix} p^{-1} & -\dfrac{\theta(B)}{p\,N_1(A)} \end{bmatrix} \begin{bmatrix} p\,\theta(B) & 0 \\ 0 & p\,(1-p)\,N_1(A) \end{bmatrix} \begin{bmatrix} p^{-1} \\ -\dfrac{\theta(B)}{p\,N_1(A)} \end{bmatrix}\right) \quad .$$

The asymptotic standard error simplifies to

$$\mathrm{SE}[\tilde{\theta}(B)] = \theta(B)\sqrt{\frac{1}{p\,\theta(B)} + \frac{1-p}{p\,N_1(A)}} \quad .$$

Substituting $\tilde{\theta}(B)$ and $\hat{p}$ for $\theta(B)$ and p yields a plug-in estimate of the standard error

$$\tilde{\mathrm{SE}}[\tilde{\theta}(B)] = \frac{N_1(A)\,n_0(B)}{n_1(A)}\sqrt{\frac{1}{n_0(B)} + \frac{1}{n_1(A)} - \frac{1}{N_1(A)}} \quad .$$

The expected total in subregion B is $\theta(B) + N_1(B)$. When $N_1(B)$ is not observed, it can be estimated by

$$\tilde{N}_1(B) = \frac{N_1(A)\,n_1(B)}{n_1(A)} \quad .$$

The estimated total in subregion B is then

$$\tilde{\tau}(B) = \tilde{\theta}(B) + \tilde{N}_1(B) = \frac{N_1(A)\;(n_0(B) + n_1(B))}{n_1(A)} = \frac{N_1(A)\;(n_0(B) + n_1(B))}{n_1(B) + n_1(A \backslash B)}$$

where $A\backslash B$ denotes the compliment of subregion B.

As before, the asymptotic standard error of $\tilde{\tau}(B)$ can be obtained by noting that

$$\begin{bmatrix} p\,\theta(B) & 0 & 0 \\ 0 & p\,(1-p)\,N_1(B) & 0 \\ 0 & 0 & p\,(1-p)\,N_1(A\backslash B) \end{bmatrix}^{-1/2} \begin{bmatrix} n_0(B) - p\,\theta(B) \\ n_1(B) - p\,N_1(B) \\ n_1(A\backslash B) - p\,N_1(A\backslash B) \end{bmatrix}$$

converges to a standard trivariate normal distribution as $J \to \infty$.

It follows from the Delta Method that when J is large, $\tilde{\tau}(B)$ is approximately normal with mean $\theta(B) + N_1(B)$ and variance

$$\begin{bmatrix} p^{-1} & \dfrac{N_1(A\backslash B) - \theta(B)}{p\,N_1(A)} & -\dfrac{\theta(B) + N_1(B)}{p\,N_1(A)} \end{bmatrix} \begin{bmatrix} p\,\theta(B) & 0 & 0 \\ 0 & p\,(1-p)\,N_1(B) & 0 \\ 0 & 0 & p\,(1-p)\,N_1(A\backslash B) \end{bmatrix} \begin{bmatrix} p^{-1} \\ \dfrac{N_1(A\backslash B) - \theta(B)}{p\,N_1(A)} \\ -\dfrac{\theta(B) + N_1(B)}{p\,N_1(A)} \end{bmatrix} .$$

The asymptotic standard error simplifies to

$$\mathrm{SE}[\tilde{\tau}(B)] = \theta(B)\sqrt{\frac{1}{p\,\theta(B)} + \frac{1-p}{p\,N_1(A)}\left(1 + \frac{N_1(B)\,N_1(A\backslash B)}{\theta(B)^2}\right)} .$$

Substituting $\tilde{\theta}(B)$, $\hat{p}$, $\tilde{N}_1(B)$, and $\tilde{N}_1(A\backslash B)$ for $\theta(B)$, p, $N_1(B)$, and $N_1(A\backslash B)$ yields a plug-in estimate of the standard error

$$\tilde{\mathrm{SE}}[\tilde{\theta}(B)] = \frac{N_1(A)\,n_0(B)}{n_1(A)}\sqrt{\frac{1}{n_0(B)} + \left(\frac{1}{n_1(A)} - \frac{1}{N_1(A)}\right)\left(1 + \frac{n_1(B)\,n_1(A\backslash B)}{n_0(B)^2}\right)} .$$

7.3.4 Overdispersion Finally, consider the behavior of the estimators

$$\hat{\theta}(A) = \frac{N_1(A)\ n_0(A)}{n_1(A)} \qquad \text{and} \qquad \hat{p} = \frac{n_1(A)}{N_1(A)} .$$

under the model

$$n_0(A) \sim \text{Negative Binomial}\left(p\,\theta(A),\ p\,\theta(A)\,\frac{\theta(A) - M(A)}{M(A)}\right)$$
$$n_1(A) \mid N_1(A) \sim \text{Negative Hypergeometric}\big(N(A) - 1,\ M(A) - 1,\ p\,M(A)\big) .$$

Note that we define the negative binomial and negative hypergeometric distributions as the total number of trials required to get a predetermined number of successes when sampling with and without replacement, respectively. We parameterize the negative binomial distribution by its mean and variance. The mean and variance of $n_1(A)$ given $N_1(A)$ are

$$\mathbb{E}\left[n_1(A) \mid N_1(A)\right] = p\,N_1(A) \qquad \text{and} \qquad \mathbb{V}\big(n_1(A) \mid N_1(A)\big) = w_1\,p\,(1-p)\,N_1(A) .$$

We proceed using increasing domain asymptotics as in Sections 7.3.1 and 7.3.2. We consider A to be the union of I disjoin sets, $\{A_i\}_{i=1}^{I}$, and we are interested in the behavior of $\hat{\theta}(A)$ and $\hat{p}$ as $I \to \infty$. Note again that $n_0(A) = \sum_{i=1}^{I} n_0(A_i)$, $n_1(A) = \sum_{i=1}^{I} n_1(A_i)$, $N_1(A) = \sum_{i=1}^{I} N_1(A_i)$ so that under the Lyapunov

Central Limit Theorem,

$$\begin{bmatrix} w_0\, p\, \theta(A) & 0 \\ 0 & w_1\, p\,(1-p)\, N_1(A) \end{bmatrix}^{-1/2} \begin{bmatrix} n_0(A) - p\, \theta(A) \\ n_1(A) - p\, N_1(A) \end{bmatrix}$$

converges to a standard bivariate normal distribution as $I \to \infty$ for

$$w_0 = \frac{\theta(A) - M(A)}{M(A)} \qquad \text{and} \qquad w_1 = \frac{N_1(A) - M(A)}{M(A) + 1} \qquad .$$

The weights w_0 and w_1 depend on A, although this dependence is not reflected in the notation.

It follows from the Delta Method that when I is large,

$$\hat{\theta}(A) \dot{\sim} \text{ Normal}\left(\theta(A),\ \begin{bmatrix} p^{-1} & -\dfrac{\theta(A)}{p\, N_1(A)} \end{bmatrix} \begin{bmatrix} w_0\, p\, \theta(A) & 0 \\ 0 & w_1\, p\,(1-p)\, N_1(A) \end{bmatrix} \begin{bmatrix} p^{-1} \\ -\dfrac{\theta(A)}{p\, N_1(A)} \end{bmatrix} \right) \qquad .$$

The asymptotic standard error simplifies to

$$\text{SE}[\hat{\theta}(A)] = \theta(A) \sqrt{w_0 \left(\frac{1}{p\, \theta(A)} \right) + w_1 \left(\frac{1-p}{p\, N_1(A)} \right)} \qquad .$$

Substituting $\hat{\theta}(A)$ and $\hat{p}$ for $\theta(A)$ and p yields a plug-in estimate of the standard error

$$\hat{\text{SE}}[\hat{\theta}(A)] = \frac{N_1(A)\, n_0(A)}{n_1(A)} \sqrt{\hat{w}_0 \left(\frac{1}{n_0(A)} \right) + w_1 \left(\frac{1}{n_1(A)} - \frac{1}{N_1(A)} \right)}$$

where

$$\hat{w}_0 = \frac{\dfrac{N_1(A)\ n_0(A)}{n_1(A)} - M(A)}{M(A)} \qquad \text{and} \qquad w_1 = \frac{N_1(A) - M(A)}{M(A) + 1} \qquad .$$

If we further assume that for subregion $B \subseteq A$,

$$n_0(B) \sim \text{ Negative Binomial}\left(p\, \theta(B),\ p\, \theta(B)\, \frac{\theta(B) - M(B)}{M(B)} \right) \qquad ,$$

then by an argument analogous to Section 7.3.3, the asymptotic standard error of the subregion estimator

$$\tilde{\theta}(B) = \frac{N_1(A)\ n_0(B)}{n_1(A)} \qquad .$$

is

$$\text{SE}[\tilde{\theta}(B)] = \theta(B) \sqrt{v_0 \left(\frac{1}{p\, \theta(B)} \right) + w_1 \left(\frac{1-p}{p\, N_1(A)} \right)} \qquad .$$

Substituting $\tilde{\theta}(B)$ and $\hat{p}$ for $\theta(B)$ and p yields a plug-in estimate of the standard error

$$\tilde{\text{SE}}[\tilde{\theta}(B)] = \frac{N_1(A)\, n_0(B)}{n_1(A)} \sqrt{\hat{v}_0 \left(\frac{1}{n_0(B)} \right) + w_1 \left(\frac{1}{n_1(A)} - \frac{1}{N_1(A)} \right)}$$

where

$$v_0 = \frac{\theta(B) - M(B)}{M(B)} \qquad \text{and} \qquad \hat{v}_0 = \frac{\dfrac{N_1(A)\; n_0(B)}{n_1(A)} - M(B)}{M(B)} \qquad .$$

The asymptotic standard error for the estimated subregion total $\tilde{\tau}(B) = \tilde{\theta}(B) + \tilde{N}_1(B)$ is

$$\mathrm{SE}[\tilde{\tau}(B)] = \theta(B) \sqrt{v_0 \left(\frac{1}{p\, \theta(B)} \right) + \left(\frac{v_1\, N_1(B)\, (N_1(A \backslash B) - \theta(B))^2 + v_2\, N_1(A \backslash B)\, (N_1(B) + \theta(B))^2}{N_1(A)\, \theta(B)^2} \right) \left(\frac{1-p}{p\, N_1(A)} \right)}$$

where

$$v_1 = \frac{N_1(B) - M(B)}{M(B) + 1} \qquad \text{and} \qquad v_2 = \frac{N_1(A \backslash B) - M(A \backslash B)}{M(A \backslash B) + 1}$$

Substituting $\tilde{\theta}(B)$, $\hat{p}$, $\tilde{N}_1(B)$, and $\tilde{N}_1(A \backslash B)$ for $\theta(B)$, p, $N_1(B)$, and $N_1(A \backslash B)$ yields a plug-in estimate of the standard error

$$\widetilde{\mathrm{SE}}[\tilde{\theta}(B)] = \frac{N_1(A)\, n_0(B)}{n_1(A)} \sqrt{\hat{v}_0 \left(\frac{1}{n_0(B)} \right) + \hat{v} \left(\frac{1}{n_1(A)} - \frac{1}{N_1(A)} \right)}$$

where

$$\hat{v} = \frac{\hat{v}_1\, n_1(B)\, (n_1(A \backslash B) - n_0(B))^2 + \hat{v}_2\, n_1(A \backslash B)\, (n_1(B) + n_0(B))^2}{n_1(A)\, n_0(B)^2} \qquad ,$$

$$\hat{v}_1 = \frac{\dfrac{N_1(A)\, n_1(B)}{n_1(A)} - M(B)}{M(B) + 1} \qquad \text{and} \qquad \hat{v}_2 = \frac{\dfrac{N_1(A)\, n_1(A \backslash B)}{n_1(A)} - M(A \backslash B)}{M(A \backslash B) + 1} \qquad .$$