Online Appendix A Verifying Samples are Random from Discrete Uniform

Goodman (1954) suggests several ways to check that a given sample of serial numbers is a random sample from a discrete uniform distribution. First, denoting g the largest number in the sample, divide each of the remaining $k-1$ observations in the sample by g. These $k-1$ observations should then be statistically indistinguishable from a uniform on $[0,1]$, with a Kolmogorov-Smirnov test an appropriate way to check this. Second, one may break the data into equally spaced ordinal categories, and then conduct a χ^2-test on this coarsened data: if the null cannot be rejected, then the data is at least consistent with a discrete uniform. Both tests are straightforward to implement, and we use the former.

Online Appendix B Power of the Uniformity Tests

The techniques applied in this *Letter* work optimally when the sample is one drawn from a discrete uniform distribution of serial numbers. This does not mean that any bias induced by non-uniformity invalidates the general thrust of the results presented in the *Letter*, but for completeness, we analyze this first order concern here.

We verified uniformity with the Komologorov-Smirnov (KS) test noted above, but one may be concerned as to its power and want to know the circumstances under which that test is able to correctly reject the null of non-uniformity. While there is a theoretical literature on the general issue of the power of the KS test (e.g. Durbin, 1961; Lewis, 1965), we wanted to obtain specific results for our data. Thus we set up a series of simulation experiments in which we allowed discrete serial numbers to be generated from a set of Beta distributions (including the uniform as a special case) and considered the performance of the KS test

therein. The basic idea is to verify that in 'reasonable' sample sizes, the test can distinguish (i.e., return a $p < 0.05$) between the simulated Beta sample and an actually discrete random uniform sample.

In what follows, consider a Beta distribution $\mathcal{B}(\alpha, \beta)$; for clarity in the printing of our plots, we refer to α as shape1 and β as s2. Our simulation sample size varies from 5 (around the minimum in our empirical study) to 1300 (close to the maximum in our empirical study) at intervals of 40—thus it increases $\{5, 45, 85 \ldots 1220, 1260, 1300\}$. For each sample size, we fix α, and then iterate between values of β $(1, 2, 3, 4, 5)$, before iterating α (again $1, 2, 3, 4, 5$) thus covering all 25 possible combinations: $(\mathcal{B}(1,1), \mathcal{B}(1,2), \mathcal{B}(1,3), \ldots, \mathcal{B}(5,4), \mathcal{B}(5,5))$. For each value of α and β, we conduct the drawing of the simulated sample and the uniform with which it is paired, a total of 50 times. We then take the mean of these 50 p-values. Pseudo-code is as follows:

1. for given $s\{$
2. for $\alpha \in \{1, 2, 3, 4, 5\}$ {
3. for $\beta \in \{1, 2, 3, 4, 5\}$ {
4. for $i \in \{1, \ldots, 50\}\{$
   (a) draw a random sample of size s from $\mathcal{B}(\alpha, \beta)$
   (b) conduct a Kolmogorov-Smirnov test of this sample against a uniform sample of size s
   (c) record the p-value of this test
5. take the <u>mean</u> p-value of these 50 trials, store$\}$
6. iterate the value of β (i.e., $\beta + 1$), do the 50 trials keeping value of α fixed $\}$

7. iterate the value of α (i.e., $\alpha + 1$), iterate through values of β for that value of α }

8. iterate the size of the sample to the next entry in the sample size vector }

Recall, we would like to see that the p-values are below 0.05 for any 'reasonable' sample size: this would imply that the test is correctly distinguishing between cables from a simulated (non-uniform) distribution versus a 'truly' uniform one. For our various sample sizes and Beta distributions, our results are displayed in Figure 2. In each of the plots, the broken line represents $p < 0.05$. Thus, points (that is, particular Beta distribution samples) falling below the line are successfully differentiated from a uniform sample by the KS test. As a sanity check, we included a simulation set for $\mathcal{B}(1, 1)$ (i.e., a uniform) in the top panel: helpfully, it is not differentiable from the uniform it is paired with for the tests (i.e., the black squares are everywhere above the broken line).

Our immediate observation from the the bottom panels (shape1=α=4 and shape1=β=5) is that at essentially anything above a small number of cables in the sample (45), the KS test can differentiate between a uniform sample and a Beta. For the top three panels, the evidence is more complicated. Basically, for small sample sizes, say fewer than 100 cables, the KS test sometimes commits type II errors: e.g. for a sample size of 5, the ('average') KS test reports $p > 0.05$ for any Beta distribution with shape1=α=1 (top panel). The KS test does worst when the Beta is symmetric and its parameters take low values: i.e when shape1=α=β=s2=2 or 3. That is, when the distribution of serial numbers is quite close to uniform around its median. This can be seen by the [red] circles above the plot on the second panel, and the [green] triangle in the third panel for a sample size around 45. Of course, it is not obvious that such unimodel symmetric distributions of cables are likely in practice; more importantly, the bias induced by non-uniformity in e.g. the Goodman estimator is not necessarily troubling *per se*: in Online Appendix C we give much more discussion of this

(potential) issue. Finally, notice that when we get up to our mean sample size (around 100) the KS test generally gets it right and has the power we need, with the exception of the case where $\alpha = \beta = 2$ and we need a sample size of around 300 to be confident we have a uniform.

Online Appendix C Simulation Study

We assume the sample of cable serial numbers observed in each embassy year to be draws from a discrete uniform distribution, in keeping with the practice in Goodman (1952) and Goodman (1954) and other studies. In the context of our applied research question, the discrete uniformity assumption means that each serial number, within a given embassy year, has the same ex ante probability of being included in the final Manning sample.

In this appendix, we briefly discuss conditions under which the discrete uniformity assumption is appropriate to estimate the cable population size for all cables originating from a particular embassy in a given year, and how the Goodman estimator tends to perform in settings where discrete uniformity is violated. To perform these analyses, we simulate cables being written at the daily level (and being released over the course of a year) and observe how non-constant probabilities of cables arriving in the final Manning sample may bias estimates of cable population sizes. In brief, we find that temporal shifts in the probability cables are excluded from the Manning sample are more likely to bias estimates of the total population size than vicissitudes in the daily rate of cables being written.

We close this document with a replication of our serial number analysis using a regression-based approach that incorporates information on the timing of each cable observed in the sample to help inform our estimates of the cable population size in each embassy-year. In

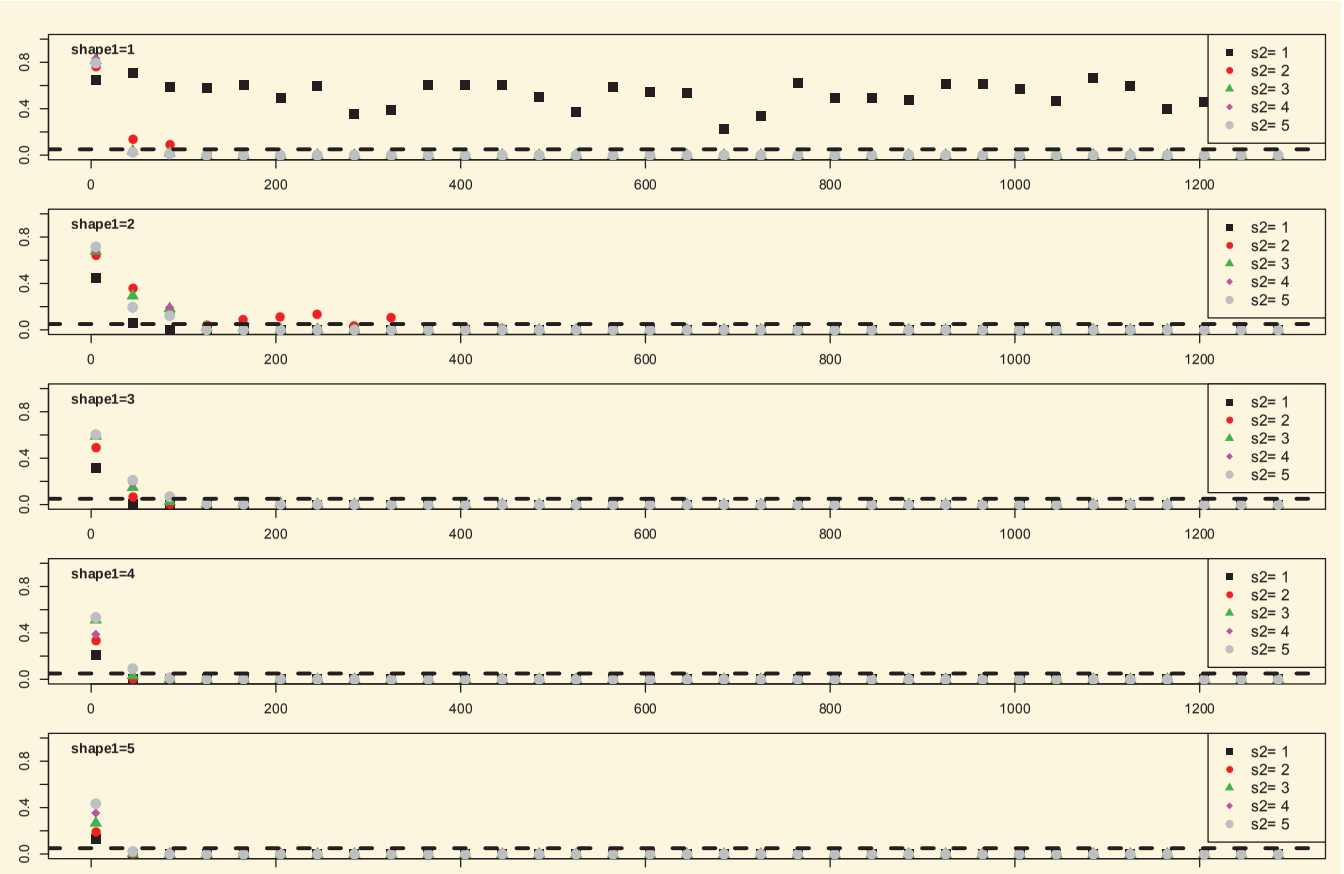

Figure 2: Power of the Kolmogorov-Smirnov test (literally, y-axis is a mean p-value) for various values of a Beta distribution versus a uniform, at various sample sizes (x-axis). Top plot fixes first Beta parameter at 1 ('shape1=1'), varies the second (s2; see legend). Second plot fixes first Beta parameter at 2, varies the second. Third plot fixes first Beta parameter at 3 and so on. Broken line represents $p = 0.05$, and thus all points below this line are successfully distinguished from the uniform by the KS test.

general, when the discrete uniformity assumption is satisfied, both Goodman-type estimates and regression-based approaches provide unbiased estimates of the population size; under some conditions, however, regression-based techniques may be preferable to the Goodman estimator if there are sharp changes in the probability is excluded from the Manning sample near the end of a calendar year, or if reasonable assumptions can be made about a fixed expected rate of cable generation across periods.

C.1 Overview

Our objective is to observe how various population size estimators perform on simulated data when (a) there may be seasonality in the rate at which cables are written, (b) there may exist seasonality in the sensitivity of cables being written. We also seek to inspect how such biases may manifest in large versus small sample settings. Evaluating such concerns through simulations, however, will require our making somewhat stylized assumptions about the data generation process of our sample. The main conclusions of our simulation studies are as follows: large shifts in the probability cables are excluded from the Manning sample are more likely to bias Goodman-type estimates of population size than shifts in the number of cables created per day. For the Goodman estimator, the bias introduced is greater as the probability of inclusion in the Manning sample is decreasing over time.

To reach these conclusions, in first set of simulations, we will assume the number of cables written on a particular day is a draw from a Poisson distribution with a fixed rate parameter. In a second set, we model the number of cables written per day as realizations of a Hawkes process (e.g., Hawkes, 1971; Ogata, 1988), which allows the instantaneous rate of cable generation to vary as a part of a "self-exciting" point process, where the occurrence of any event (i.e., a cable being written) increases the short term probability of another cable being written. In our applied context, the simulated Hawkes process will lead to clustered

periods of time with higher than baseline (i.e., random) patterns of cable generation. For one set of Poisson simulations, we will set the rate parameter to equal $\lambda = 5$. For one set of Hawkes simulations, we will set the initial conditions to equal $\mu = 10/3$, $\alpha_1 = 1$, and $\beta_1 = 3$, and simulate events in continuous time for $T = 365$. These parameters were selected because they generate, in expectation, equal totals of cables over the course of an entire year, but vary in terms of their temporal clustering and variance.[9] The virtue in maintaining approximately equal yearly sample sizes in the Poisson and Hawkes study conditions is that it allows for easy inspection of how clustered periods of higher cable generation rates—rather than sample size on its own, or variation in the probability serial numbers are out of sample—influence population size estimates.[10] As the next section will show, however, the 'burstiness' of cables being generated over time may be more likely to bias regression-based estimates of population size when such periods of time are correlated with large shifts in the probability cables are excluded from the Manning sample. Goodman-type estimators (which rely more heavily on the observed value of the sample maximum serial number) may be less sensitive to burst-induced biases if the probability that cables appear in the Manning sample is sufficiently high near the end of a calendar year.

C.2 Sensitivity of Assumptions for Goodman and Regression-based Estimators

Absent large shifts over time in the probability that serial numbers are excluded from the Manning sample, daily cable counts being produced from Poisson and Hawkes processes are

[9] In addition to the "Large N" case where the expected number of cables written per year is 1825, we will also replicate our analysis on a "Small N" case when the expected yearly total is 365.

[10] If the number of cables created on any given day is $n_t \sim \text{Pois}(\lambda)$, then the expected number of cables being created over the course of a year is simply $\sum_{t=1}^{365} E[n_t] = 365 \cdot 5 = 1825$. In a Hawkes process, the instantaneous rate parameter in time t is $\lambda(t) = \mu + \sum_{t_i<t} \alpha e^{\beta(t-t_i)}$. Under the condition that the exponential rate of decay is greater than the self-excitation growth rate ($\beta > \alpha$), and as the number of periods $T \to \infty$, the expected value of the rate parameter is $E[\lambda] = \frac{\mu}{1-\int_0^\infty \alpha e^{-\beta t} dt} = \frac{\mu}{1-(\alpha/\beta)}$.

both acceptable for the Goodman and regression-based estimators of cable population size at the embassy year level. This point can be demonstrated with a simple example. First let the number of cables observed on a given day be n_t, and the probability any given cable is included in the Manning sample on day t be $p_t = p = 0.5$. If cables are given serial numbers in the order in which they are released, and the probability a cable is included in the sample is independent of the day of the year, one can imagine data being generated over a full year like those listed on the lefthand side of Table 3. As should be clear, despite the daily variation in cable counts across days, the probability any given serial number is included in the sample is orthogonal to the day of the year on which it was written. This implies that the sample of serial numbers $\{1, 4, 5, 7, 8, 10, 11, \ldots, 1829\}$ is precisely a random sample from a discrete uniform distribution of size $N = 1831$, since each serial number has an equal probability of being drawn into the Manning sample. It is important this stylized example imposes no structure on how n_t is drawn. Regardless of whether the daily counts of cables result from Poisson or Hawkes processes (much less any stochastic process), the serial numbers included in the Manning sample are precisely a random draw from a discrete uniform distribution, which is guaranteed so long that p_t is fixed over time. On such a sample of data, to estimate the number of cables written in a given embassy year using Goodman estimator, therefore, would be a natural choice. In our simulation results we show that data generated and analyzed in such a fashion provide unbiased estimates of the population size.

The righthand side of Table 3 provides an example when the assumption of discrete uniformity (of serial numbers in the Manning sample, over the course of a full year) is not satisfied. The example provided is meant to be an extreme case in which there is a pronounced reduction in the probability of cables being included in the study sample in periods 4,..., 365, moving from $p_t = 0.5$ to $p_t = 0$. If such censorship were to occur in the data—i.e., for a fixed $N = 1831$, relying on a study sample of $\{1, 4, 5, 7, 8, 10, 11, 13, 16\}$ instead

Table 3: Serial numbers included in a hypothetical Manning sample over the course of a year, where p_t denotes each cable's probability of being drawn into the Manning sample in time t. The righthand column notes the maximum serial number observed in the Manning sample in period t, denoted M_t. Underlined serial numbers indicate cables included in a hypothetical Manning sample.

Discrete Uniformity Satisfied				
t	n_t	Serials$_t$	p_t	M_t
1	5	1 2 3 4 5	0.5	5
2	4	6 7 8 9	0.5	8
3	8	10 11 12 13 14 15 16 17	0.5	16
4	6	18 19 20 21 22 23	0.5	22
5	4	24 25 26 27	0.5	26
⋮	⋮	⋮	⋮	⋮
365	3	1829 1830 1831	0.5	1829

Discrete Uniformity Unsatisfied				
t	n_t	Serials$_t$	p_t	M_t
1	5	1 2 3 4 5	0.5	5
2	4	6 7 8 9	0.5	8
3	8	10 11 12 13 14 15 16 17	0.5	16
4	6	18 19 20 21 22 23	0	NA
5	4	24 25 26 27	0	NA
⋮	⋮	⋮	⋮	⋮
365	3	1829 1830 1831	0	NA

of $\{1, 4, 5, 7, 8, 10, 11, \ldots, 1829\}$—the Goodman estimator would severely underestimate the true population size. However, if instead one were to use only the first three days of observed data, a linear extrapolation that relies only on the maximum serial observed on each day in the observed Manning sample may provide a more plausible estimate of the total population size. Extrapolation is particularly merited if one is willing to make assumptions about the expected number of cables being written per day being relatively constant over time.

The aim here is to provide a sketch of this intuition. Assume the number of cables written on day t is $n_t \sim g(\theta)$, with the expected number of documents written per day being $E[X]$. If a cable is written in time t, there is some p_t probability that cable is included in the Manning sample. Each cable written (i.e., both those in the Manning sample and outside the Manning sample) are given a serial number according to the order in which it is written, in line with the process shown in Table 3. On the first day, the expected number of cables included in the sample is $E[X] \cdot p_1$, on the second day the expected number of cables included in the Manning sample is $E[X] \cdot p_2$, and on the t-th day the expected number of cables included in the Manning sample is $E[X] \cdot p_t$. This is by definition true so long as the distribution from

which daily cable counts are drawn is fixed over time. The expected Manning sample size over T periods is simply $E[X] \cdot \sum_{t=1}^{T} p_t$.

In general, however, the expected number of cables appearing in the Manning sample per day is not the same as *the expected value of the maximum serial number observed on day* t, which we will denote $E[M_t]$. In a trivial case, for example, $E[M_t]$ may be undefined if $p_t = p = 0$, even if $E[X] > 0$. The expected value of the sample maximum observed on day t will depend on several quantities: the number of periods that have passed prior to period t (accounting for the serial numbering pattern), the expected number of cables written per day, and the expected number of cables entering into the Manning sample on day t. More formally, we denote the expected value of the maximum serial number observed on day 1 as $E[M_1] = f(p_1, \theta)$, the maximum serial number observed on day 2 as $E[M_2] = E(X) + f(p_2, \theta)$, and $E[M_t] = (t-1)E[X] + f(p_t, \theta)$. The first summand of $E[M_t]$ accounts for the *expected starting point* of serial numbers written in period t, regardless of whether they appear in the sample; the second summand adjusts directly for the expected maximum serial number observed in the Manning sample, which is a function of the underlying daily count function and p_t. If we differentiate $E[M_t]$ with respect to t, we observe $\frac{\partial E[M_t]}{\partial t} = E(X) + \frac{\partial f}{\partial p_t}\frac{\partial p_t}{\partial_t}$. By necessity the sign on $\frac{\partial f}{\partial p_t}$ will always be positive, but $\frac{\partial p_t}{\partial_t}$ may positive, negative, or equal to zero. Clearly, if $\frac{\partial p_t}{\partial_t} = 0$, then $\frac{\partial E[M_t]}{\partial t} = E(X)$.

C.2.1 A Regression-based Estimator of Cable Population Size

The observation that $\frac{\partial E[M_t]}{\partial t} = E(X) + \frac{\partial f}{\partial p_t}\frac{\partial p_t}{\partial_t}$ is valuable because it motivates a linear regression-based approach to estimate the total number of cables written in a given year. It also provides intuition on the bias that may be introduced through such an estimation approach if $\frac{\partial p_t}{\partial_t} \neq 0$. Let us first consider the case when $\frac{\partial p_t}{\partial_t} = 0$. If changes over time in the probability cables are included in the Manning sample are not linearly associated with time,

the subsequent regression-based approach will be appropriate to estimate the total number of cables produced at an embassy in a given year. Namely, for each embassy year, aggregate the observed sample of data at the daily level, and estimate the following bivariate regression equation:

$$M_t = \beta_0 + \beta_1 \cdot t + \varepsilon, \tag{1}$$

where M_t is the maximum serial observed (e.g., $\{5, 8, 16\}$) in on day t, t is the numeric calendar day (e.g., $\{1, 2, 3\}$), and ε is the error. At the embassy level, we estimate the total number of cables written over a 365 day period as

$$\hat{N} = \widehat{M_{365}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 365. \tag{2}$$

Straightforwardly, the quantity $\hat{\beta}_1$ is an estimate of $\frac{\partial E[M_t]}{\partial t}$. In leap years, the fitted value for day 366 would be used.

When there exists an association between the expected change in p_t and t, however, this estimator may be biased. If $\frac{\partial p_t}{\partial_t} > 0$, the estimator will tend to produce estimates that are somewhat larger than the true population size, and when $\frac{\partial p_t}{\partial_t} < 0$ the estimates will tend to undershoot the true population size. So too, if there are associations between changes in rate of cable generation and changes in the probability probability with which cables are included in the sample, the estimator may be biased in expectation. The magnitude of this bias will depend precisely on magnitude of the unobserved shifts in cable generation and p_t.

There may be cases in which sharp shifts in p_t do not threaten the validity of linear extrapolation, however. Consider the case when $p_t = 0.1$ for the first half of a calendar year, and $p_t \approx 0$ in the second half. (This scenario is approximated in the "Second Half Censored" study condition mentioned in the next section.) In this extreme case, linear extrapolation

given the observed data may be reasonable: even though the range of the observed data is weighted exclusively to the first half of a calendar year, if the true rate of cable generation in the first half of the year is close to the rate of cable generation in the second half of the year, the estimates $\widehat{\frac{\partial E[M_t]}{\partial t}}$ obtained from the first half of the calendar year should appropriately map to the second half of the year, even if no data are observed in sample from that period.

C.2.2 Study Conditions and Outcomes of Interest

To assess how various estimators perform across various hypothetical data generation processes, we vary both the distribution from which daily cable counts are drawn, in addition to the probability that any cable written on day t is to be included in the Manning sample. As before, we denote the probability that a cable is included in the Manning sample, given that it is written on day t, as p_t.

We report simulation results for eight different manipulations of p_t. The names of these study conditions are presented along with their formal definitions in Table 4.

Table 4: Study Manipulations: Variation in the Probability of a Cable's Inclusion in the Manning Sample, given that it was written on day t

Condition	Definition
Fixed Probabilities	$p_t = p = 0.1$
First Half Censored	$p_t = 0.001, \text{if } t < 183;\ p_t = 0.1, \text{if } t \geq 183$
Second Half Censored	$p_t = 0.1, \text{if } t < 183;\ p_t = 0.001, \text{if } t \geq 183$
Random Uniform	$p_t \sim \text{Unif}(0, 0.1)$
Inverted U-Shape	$p_t = \sin(t \cdot \pi/365)/10$
U-Shape	$p_t = (1 - \sin(t \cdot \pi/365))/10$
Linear Increase	$p_t = (t/365)/10$
Linear Decrease	$p_t = (1 - t/365)/10$

In addition to varying the probability with with written cables appear in the Manning sample, we vary whether daily cables counts arise as a result of a Poisson process or a Hawkes process. For both the Poisson and the Hawkes study conditions, we have "Large N" and a

"Small N" variants. In the "Large N" conditions, the Poisson parameter is $\lambda = 5$, while the respective Hawkes parameters are defined as $\mu = 10/3$ (the baseline rate), $\alpha = 1$ (the excitation parameter), and $\beta = 3$ (the exponential decay). In the "Small N" study conditions, the Poisson parameter $\lambda = 1$, and the Hawkes parameters are $\mu = 0.2$, $\alpha = 0.8$, and $\beta = 1$.

Using Poisson and Hawkes data generation process, across both the "Large N" and "Small N" study conditions, we perform 2,500 random simulations of each of the study conditions listed in Table 4. In each of these simulations we record the "true" number of cables generated by either the Poisson or Hawkes processes, in addition to the estimates of each of the MLE, Goodman, and regression-based estimators. In each iteration of the simulation, we divide each estimator's estimate of the total population size by the true number, yielding $\hat{N}/N$, and we store this value. If across multiple simulations a particular estimator systematically yields values of $\hat{N}/N > 1$, this provides evidence that an estimator tends to overestimate the true number of cables. Similarly, if a particular estimator on average yields values of $\hat{N}/N < 1$, this provides evidence that an estimator, given the study conditions, tends to underestimate the true number of cables in a given embassy year.

C.3 Results

Figures 3 through 6 present the results of this simulation study. In each subplot, the mean value of $\hat{N}/N$ across simulations is presented beneath each estimator's name. The upper and lower boundaries of each boxplot denote the interquartile range of simulation results for each estimator. The median result is presented as a solid, horizontal line. The upper and lower whiskers denote values 1.5 above or below the interquartile range of the plot.

Overall, the regression-based estimator performs consistently well. When the discrete uniformity assumption is satisfied, however, the Goodman estimator is unbiased and exhibits

the lowest variance. The bias and variance of each estimator appears to be larger in the "Small N" study conditions. In our applied example, the Goodman estimator is most biased cases in which p_t is decreasing over time. Relative to the "Inverted U-Shape", the "U-Shape" study conditions have distributions of $\hat{N}/N$ closer to 1.

Figure 3: *Simulation results for "Poisson, Large N" study.* The results of 2500 random simulations reflected in each subplot. In this condition, $\lambda = 5$, such that the expected number of cables per year is 1825.

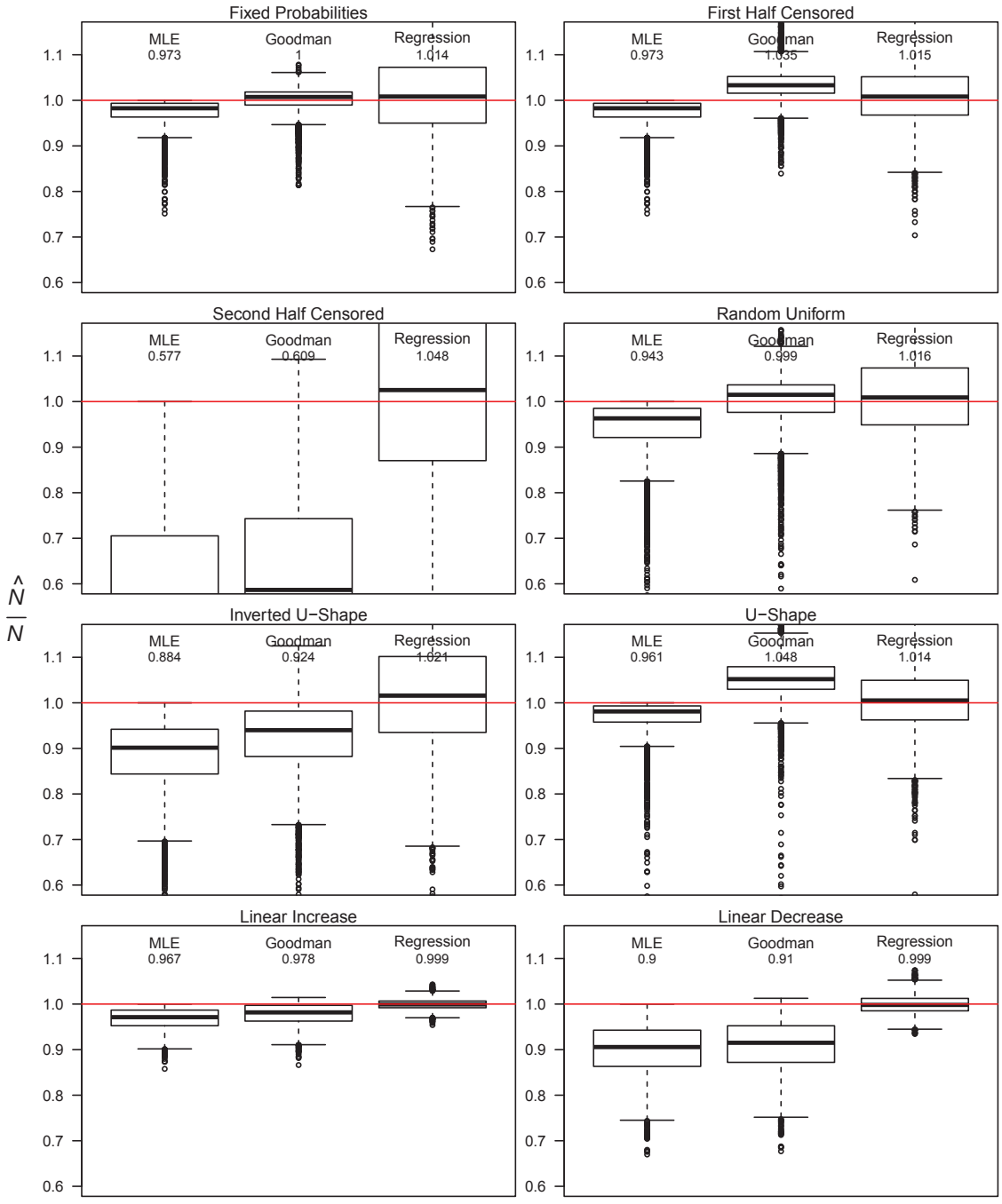

Figure 4: *Simulation results for "Poisson, Small N" study.* The results of 2500 random simulations reflected in each subplot. In this condition, $\lambda = 1$, such that the expected number of cables per year is 365.

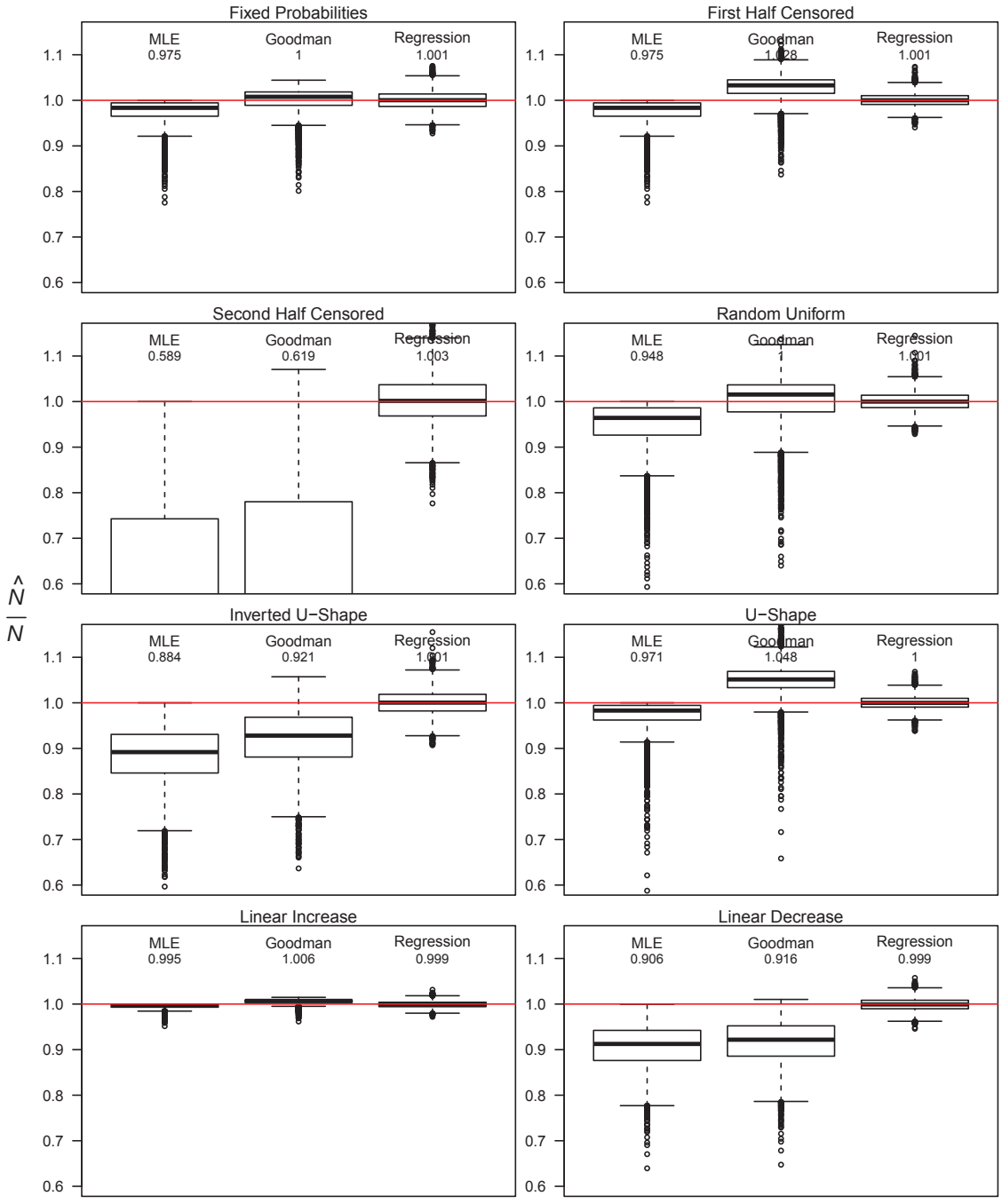

Figure 5: *Simulation results for "Hawkes, Large N" study.* The results of 2500 random simulations reflected in each subplot. In this condition, $\mu = 10/3, \alpha = 1$, and $\beta = 3$, such that the expected number of cables per year is 1825.

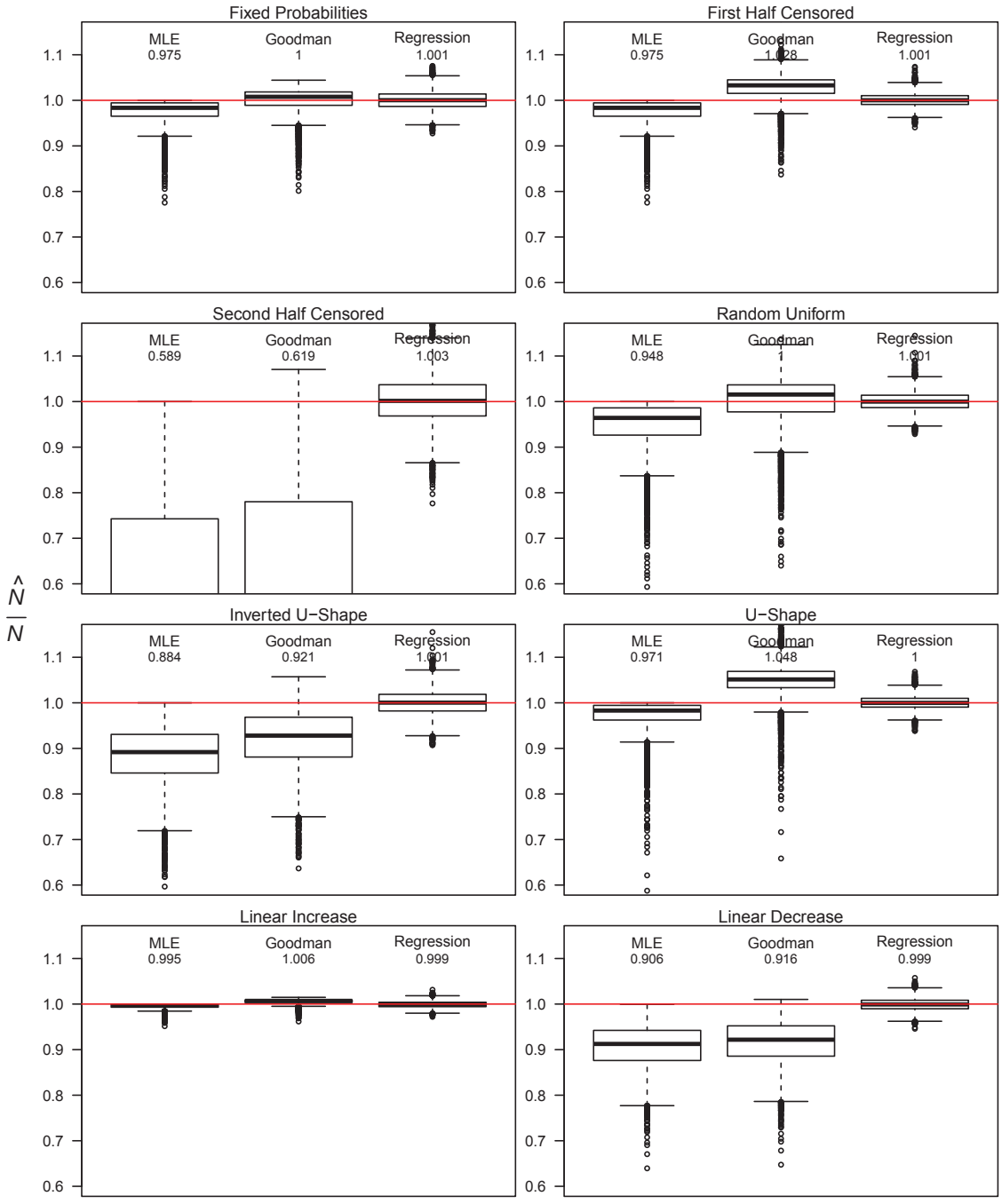

Figure 6: *Simulation results for "Hawkes, Small N" study.* The results of 2500 random simulations reflected in each subplot. In this condition, $\mu = 0.2, \alpha = 0.8$, and $\beta = 1$, such that the expected number of cables per year is 365.