BU-390-M.

Estimation of season total number of different households utilizing a park: Preliminary report to W. H. Gauger

D. S. Robson

Abstract

Daily records of the identification of the households utilizing a public park produce, at the end of the season, a count of the total number (K) of different households which are directly benefitting from this public facility. If household identification records are collected on only a random sample of n days during the N day season then an unbiased estimator of K can be constructed in the form

$$\hat{K} = N\bar{r}_{(1)} - \binom{N}{2}\bar{r}_2 + \ldots + (-1)^{n^*-1}\binom{N}{n^*}\bar{r}_{(n^*)}$$

where $\bar{r}_{(m)}$ is the average size of the intersection of m sample days,

$$\binom{n}{m}\bar{r}_{(m)} = \sum_j \binom{c_j}{m}$$

where c_j is the number of days that household j was in the sample. The sample size n must exceed the maximum number of visits (n*) of any household in order to guarantee that $\hat{K}$ is unbiased. This estimation procedure can be extended to the case of stratified sampling of days.

Estimation of season total number of different households utilizing a park: Preliminary report to W. H. Gauger

BU-390-M | D. S. Robson | September, 1971

Introduction

One measure of the extent of utilization of a park or other public facility is the number of different households served by the park in a year. This number could be determined from a complete daily record of the household identification of all park visitors. Compiled in the form illustrated below

Day of Visit	Household Identification Numbers 1	2	3	4	5	. . .	K-1	K	Total
1	x	x	x			. . .			R_1
2		x		x		. . .			R_2
3					x	. . .			R_3
.						.			.
.						.			.
.						.			.
N-1			x			. . .		x	R_{N-1}
N					x		x		R_N
Total	C_1	C_2	C_3	C_4	C_5	. . .	C_{K-1}	C_K	T

the daily records for an entire season of N days would reveal the frequency of use by each household and, in particular, would reveal the desired information on the total number (K) of households utilizing the park at least once during the season.

Maintaining such records for the entire season would generally be a costly and impractical procedure, and we consider here the statistical problem of estimating K from the records obtained on a sample of n days drawn at random. As in tag-recapture experiments, which bear strong resemblance to this sampling procedure, no uniformly unbiased estimator of K can be found; but if $n \geq \max(C_1, \ldots, C_K)$ -- that is, if the enumerator collecting these records visits the park more frequently than any household -- then the estimator described here is unbiased.

Estimation Procedure

If the collection of households visiting the park on day i is denoted by S_i then R_i is the number of elements (households) in S_i, say $R_i = \#(S_i)$, and

$$K = \#\left(S_1 \cup S_2 \cup \cdots \cup S_N\right) .$$

The cardinality of the union is also given by the finite series

$$K = \sum_{i}^{N} \#\left(S_i\right) - \sum_{i_1<i_2}^{N} \#\left(S_{i_1} \cap S_{i_2}\right) + \sum_{i_1<i_2<i_3} \#\left(S_{i_1} \cap S_{i_2} \cap S_{i_3}\right) - \cdots$$

or, letting

$$R_{i_1 \cdots i_m} = \#\left(S_{i_1} \cap \cdots \cap S_{i_m}\right)$$

and

$$\bar{R}_{(m)} = \frac{1}{\binom{N}{m}} \sum_{i_1<\cdots<i_m}^{N} R_{i_1 \cdots i_m}$$

then

$$K = \sum_{i_1}^{N} R_{i_1} - \sum_{i_1<i_2}^{N} R_{i_1 i_2} + \cdots + (-1)^{n^*-1} \sum_{i_1<\cdots<i_{n^*}} R_{i_1 \cdots i_{n^*}}$$

$$= N\bar{R}_{(1)} - \binom{N}{2} \bar{R}_{(2)} + \cdots + (-1)^{n^*-1} \binom{N}{n^*} \bar{R}_{(n^*)}$$

where

$$n^* = \max\left(C_1, \ldots, C_K\right) .$$

Another and computationally more convenient representation of K is obtained by noting that

$$\bar{R}_{(m)} = \frac{1}{\binom{N}{m}} \sum_{j}^{K} \binom{C_j}{m} .$$

The same representation applies to a sample of n days; if we use lower case letters to denote sample values then

$$k = n\bar{r}_{(1)} - \binom{n}{2}\bar{r}_{(2)} + \ldots + (-1)^{n^*-1}\binom{n}{n^*}\bar{r}_{(n^*)}$$

where

$$\bar{r}_{(m)} = \frac{1}{\binom{n}{m}} \sum_{j=1}^{k} \binom{c_j}{m} .$$

For all $m \leq n$ the sample means $\bar{r}_{(m)}$ are unbiased estimators of the corresponding population means $\bar{R}_{(m)}$; hence, if $n \geq n^*$ then

$$\hat{K} = N\bar{r}_{(1)} - \binom{N}{2}\bar{r}_{(2)} + \ldots + (-1)^{n-1}\binom{N}{n}\bar{r}_{(n)}$$

is an unbiased estimator of K.

The value of n^* is determined by the households making most frequent use of the park. In practice these regular visitors can be identified by park personnel and treated as a separate segment, resulting in a smaller value of n^* for the remaining population of visitors. The number of sample days (n) is thus required only to exceed this reduced maximum frequency in order to assure unbiasedness.

Stratified sampling and estimation: Preliminary considerations

Since daily attendance at a park differs substantially between weekdays and weekends the sampling of days is conventionally stratified by this criterion. Thus, if the season includes N_1 weekend days and holidays and N_2 non-holiday weekdays, $N_1 + N_2 = N$, then the sample of n days is conventionally obtained by a random selection of n_1 weekend-holiday days and an independent random selection of n_2 non-holiday weekdays, $n_1 + n_2 = n$. We consider the modifications in $\hat{K}$ required for unbiased estimation in this circumstance.

A somewhat more complicated notation is needed to describe a stratified population, and to this end we define

$$A = \{S_1, \ldots, S_{N_1}\}$$

$$A_{(2)} = \{S_1 \cap S_2,\ S_1 \cap S_3, \ldots, S_{N_1-1} \cap S_{N_1}\}$$

and

$$B = \{S_{N_1+1}, \ldots, S_{N_1+N_2}\}$$

$$B_{(2)} = \{S_{N_1+1} \cap S_{N_1+2}, \ldots, S_{N_1+N_2-1} \cap S_{N_1+N_2}\}$$

and, in general, $A_{(m)}$ is the collection of all $\binom{N_1}{m}$ m-fold intersections of sets in A, and $B_{(m)}$ is similarly defined. The mean $\bar{R}_{(m)}$ appearing in

$$K = N\bar{R}_{(1)} - \binom{N}{2}\bar{R}_{(2)} + \ldots + (-1)^{m-1}\binom{N}{m}\bar{R}_{(m)} + 111$$

is now given by

$$\binom{N}{m}\bar{R}_{(m)} = \sum_{\nu=0}^{m}\binom{N_1}{\nu}\binom{N_2}{m-\nu}\bar{R}_{(\nu,m-\nu)}$$

where

$$\binom{N_1}{\nu}\binom{N_2}{m-\nu}\bar{R}_{(\nu,m-\nu)} = \sum_{\substack{1\le i_1<\ldots<i_\nu\le N_1 \\ N_1+1\le i_{\nu+1}<\ldots<i_m\le N_1+N_2}} R_{i_1\ldots i_m} \quad .$$

Note that

$$K = K_1 + K_2 - K_{1,2}$$

where

$$K_1 = N_1\bar{R}_{(1,0)} - \binom{N_1}{2}\bar{R}_{(2,0)} + \ldots + (-1)^{m-1}\binom{N_1}{m}\bar{R}_{(m,0)} + \ldots$$

$$K_2 = N_2\bar{R}_{(0,1)} - \binom{N_2}{2}\bar{R}_{(0,2)} + \ldots + (-1)^{m-1}\binom{N_2}{m}\bar{R}_{(0,m)} + \ldots$$

$$K_{1,2} = N_1N_2\bar{R}_{(1,1)} - \binom{N_1}{1}\binom{N_2}{2}\bar{R}_{(1,2)} - \binom{N_1}{2}\binom{N_2}{1}\bar{R}_{(2,1)} + \ldots$$

$$\ldots + (-1)^m\sum_{\nu=1}^{m-1}\binom{N_1}{\nu}\binom{N_2}{m-\nu}\bar{R}_{(\nu,m-\nu)} + \ldots$$

Sample means $\bar{r}_{(\nu,m-\nu)}$ are defined in a completely analogous manner and the estimator

$$\hat{K} = \sum_{m=1}^{n^*} (-1)^{m-1} \sum_{\nu=0}^{m} \binom{N_1}{\nu}\binom{N_2}{m-\nu} \bar{r}_{(\nu,m-\nu)}$$

is then unbiased provided that $n_1 \geq n_1^*$, $n_2 \geq n_2^*$ and $n \geq n^*$. For computational purposes we note that

$$\binom{n_1}{\nu}\binom{n_2}{m-\nu} \bar{r}_{(\nu,m-\nu)} = \sum_{j=1}^{K} \binom{c_{1j}}{\nu}\binom{c_{2j}}{m-\nu}$$

where c_{1j} and c_{2j} are the observed frequency of visits of the j'th household in the samples of n_1 days and n_2 days respectively.

Numerical illustration with a hypothetical population

Hypothetical records for a N = 6 day season are given below, and the estimate $\hat{K}$ is computed for each of the $\binom{N}{n} = \binom{6}{3} = 20$ different possible samples of n = 3 days.

Season Record

Household identification number (j)

Day (i)	1	2	3	4	5	6	7	8	9	10	Total(R_i)
1	x	x	x	x							4
2		x			x	x					3
3				x		x	x				3
4								x			1
5				x			x		x		3
6		x								x	2
Total (C_j)	1	3	1	3	1	2	2	1	1	1	16 (T

The total number K = 10 of different households utilizing the park during the N = 6 day season can be expressed as

$$K = \sum_{j=1}^{K} \binom{C_j}{1} - \sum_{j=1}^{K} \binom{C_j}{2} + \sum_{j=1}^{K} \binom{C_j}{3}$$

$$= 16 - 8 + 2 \quad .$$

Only three terms were needed in this series since no C_j exceeds $n^* = 3$. A sample size n of at least $n^* = 3$ days is therefore required to achieve unbiasedness; for illustrative purposes we therefore enumerate the sampling distribution for the case n = 3.

The sample consisting of days 2,4, and 5, for example, would produce the sample table:

	Household							
Day	2	4	5	6	7	8	9	Total
2	x		x	x				3
4						x		1
5		x			x		x	3
Total	1	1	1	1	1	1	1	7

giving

$$\bar{r}_{(1)} = \frac{1}{n} \sum c_j = \frac{1}{3} (7)$$

$$\bar{r}_{(2)} = \frac{1}{\binom{n}{2}} \sum \binom{c_j}{2} = 0$$

$$\bar{r}_{(3)} = \frac{1}{\binom{n}{3}} \sum \binom{c_j}{3} = 0$$

and

$$\hat{K} = N\bar{r}_{(1)} - \binom{N}{2} \bar{r}_{(2)} + \binom{N}{3} \bar{r}_{(3)}$$

$$= 6 \left(\frac{7}{3}\right) - 15 \ (0) + 20 \ (0) = 14 \quad .$$

A complete enumeration of the $\binom{6}{3} = 20$ possible samples of size $n = 3$ gives the following frequency distribution of estimates:

$\hat{K}$	3	4	5	6	7	8	9	10	11	12	14	20	23	Total
frequency	1	1	1	2	1	2	2	1	4	2	1	1	1	20

The average of these 20 estimates is exactly $K = 10$, and $\sigma^2_{\hat{K}} = 22.9$. Regrettably, there are several samples among these 20 for which $\hat{K} < k$ -- i.e., where the estimated number of different households is less than the number actually observed. In practice such a $\hat{K}$ would certainly be increased to the value k; such adjustments do destroy the unbiasedness as a formal property of the estimator but improve the general properties of $\hat{K}$, as indicated below:

$\hat{K} \geq k$	5	6	7	8	9	10	11	12	14	20	23	Total
frequency	1	3	2	2	2	1	4	2	1	1	1	20

where the average value of the adjusted $\hat{K}$ is 10.3 instead of 10.

A stratified sampling procedure can also be illustrated with this hypothetical population. If days 1,2,3,4 represent one stratum and days 5,6 another stratum:

Stratified Season Record

Household identification number

Day	1	2	3	4	5	6	7	8	9	10	Total
1	x	x	x	x							4
2		x			x	x					3
3				x		x	x				3
4								x			1
Subtotal	1	2	1	2	1	2	1	1			11
5				x			x		x		3
6		x								x	2
Subtotal	0	1	0	1	0	0	1	0	1	1	5
Total	1	3	1	3	1	2	2	1	1	1	16

then

$$K_1 = 11 - 3 + 0 = 8$$

$$K_2 = 5 - 0 + 0 = 5$$

$$K_{1,2} = \binom{N_1}{1}\binom{N_2}{1}\frac{\sum\binom{C_{1j}}{1}\binom{C_{2j}}{1}}{\binom{N_1}{1}\binom{N_2}{1}} - \binom{N_1}{2}\binom{N_2}{1}\frac{\sum\binom{C_{1j}}{2}\binom{C_{2j}}{1}}{\binom{N_1}{2}\binom{N_2}{1}}$$

$$= \binom{N_1}{1}\binom{N_2}{1}\bar{R}_{(1,1)} - \binom{N_1}{2}\binom{N_2}{1}\bar{R}_{(2,1)}$$

$$= 5 - 2 = 3$$

giving, again

$$K = K_1 + K_2 - K_{1,2} = 8 + 5 - 3 = 10 .$$

Since $n_1^* = \max\left(C_{1j}\right) = 2$, $n_2^* = \max\left(C_{2j}\right) = 1$, $n^* = \max\left(C_j\right) = 3$ then a stratified sample of total size $n = 3$ with $n_1 = 2$ and $n_2 = 1$ will produce an unbiased estimator. There are $\binom{N_1}{n_1}\binom{N_2}{n_2} = \binom{4}{2}\binom{2}{1} = 12$ different possible and equally likely samples in this case; for example, choosing days 1 and 3 from the first stratum and day 5 from the second stratum gives

	Household							
Day	1	2	3	4	6	7	9	Total
1	x	x	x	x				4
3				x	x	x		3
Subtotal	1	1	1	2	1	1		7
5				x		x	x	3
Subtotal	0	0	0	1	0	1	1	3
Total	1	1	1	3	1	2	1	10

$$\hat{K}_1 = 4\left(\frac{7}{2}\right) - 6(1) = 8$$

$$\hat{K}_2 = 2(3) = 6$$

$$\hat{K}_{1,2} = 4(2)\left(\frac{3}{2}\right) - 6(1)(2) = 0$$

$$\hat{K} = \hat{K}_1 + \hat{K}_2 - \hat{K}_{1,2} = 14$$

(Note that the observed overlap of the two strata is $k_{1,2} = 2$, while $\hat{K}_{1,2} = 0$). The sampling distribution of $\hat{K}$ over the 12 possible samples is:

$\hat{K}$	4	6	8	10	14	16	Total
frequency	1	2	2	2	2	1	12

giving $E(\hat{K}) = K = 10$ and $E(\hat{K} - K)^2 = \sigma^2_{\hat{K}} = 9\ 1/3$; this stratification thus reduces the variance of the estimator from 22.9 to 9 1/3. Anomolies such as the one noted above $\left(k_{1,2} = 2,\ \hat{K}_{1,2} = 0\right)$ can be eliminated to produce a slightly improved but slightly biased estimator.

Estimation of the frequency distribution of visits per household

As an extension of the preceding method for estimating K by estimating the individual terms in a finite series expansion of K we note that similar expansions may be employed to calculate the frequency distribution of visits per household. If we simplify the notation by defining

$$T_{(m)} = \binom{N}{m} \bar{R}_{(m)}$$

and introduce the frequencies

$$f_m = \#\left(C_j = m\right) \qquad m = 1,2,\ldots,n^*$$

$$f_m = T_{(m)} - \binom{m+1}{m} T_{(m+1)} + \binom{m+2}{m} T_{(m+2)} + \ldots + (-1)^{n^*-m} \binom{n^*}{m} T_{(n^*)} \ .$$

Letting

$$\hat{T}_{(m)} = \binom{N}{m} \bar{r}_{(m)}$$

then

$$\hat{f}_{(m)} = \sum_{\nu=0}^{n^*-m} (-1)^{\nu} \binom{m+\nu}{m} \hat{T}_{(m+\nu)}$$

is an unbiased estimator of f_m provided that $n \geq n^*$.