CONFIDENCE INTERVALS FOR THE NUMBER OF UNSEEN TYPES

By Mark Finkelstein, Howard G. Tucker, and Jerry Alan Veeh

University of California Irvine, University of California Irvine,
and Auburn University

July 31, 1997

Abstract. This paper finds the unique maximum likelihood estimator of, and conservative confidence intervals for, the unknown number of different coupons in the coupon collector's problem. This problem is also known as the problem of estimating the abundance of wildlife. The techniques developed here can be easily implemented, are valid without regard to sample size, and validate previous methods based on large sample theory when those methods apply.

Introduction

An infected organism is being studied. A sample of n pathogens (viruses, bacteria, etc.) of a single species from the organism is examined, and the strain of each is determined. The number of different strains observed among the n pathogens is c. Knowing only n and c, the number k of different strains of this pathogen present in the organism is to be estimated.

In a national park in India, an automatic camera photographed the number n of tigers passing by the camera over a 12 month period. According to *The New York Times*, from the photographs it was determined that the number of different tigers observed was c. From n and c, the total number k of tigers in the park is to be estimated.

In general, can a simple random sample of size n from a population of unknown size k be used to estimate k? The answer is in the affirmative if the sampling can be done *with* replacement, and the number c of distinct units in the sample can be determined.

These problems are related to the coupon collector's problem where one observes the number c of distinct coupons collected out of n coupons and wishes to estimate the unknown total number k of different coupons. The coupons are assumed to be selected either without replacement from an infinite population containing equal numbers of the k different coupons, or with replacement from a finite population containing equal numbers of the k different coupons. These problems are also related to the classical occupancy problem

1991 *Mathematics Subject Classification.* 62F10, 62F25, 62P10, 62P99.
Key words and phrases. Coupon collector's problem, occupancy problem, wildlife abundance.

in which the number k of boxes is to be estimated when the number n of balls which have been dropped at random into the k boxes and the number c of occupied boxes is known.

This paper is concerned with the problem of finding an estimator of, and confidence intervals for, the unknown parameter k in the settings described above. The conceptual model underlying the analysis is that an urn contains k coupons, each of a different color. Coupons are drawn from the urn one at a time with replacement and the color of each coupon drawn is noted. The value of k is unknown and is to be estimated using the number C_n of different colors that have appeared in n draws from the urn with replacement.

The first result established here is that when the maximum likelihood estimator of k exists, it is unique. Moreover, there is a simple, computationally effective way of determining this maximum likelihood estimator using only C_n and n. The problem of determining the maximum likelihood estimator of k has been studied extensively, especially in the context of estimating wildlife abundance. The maximum likelihood estimator has been previously derived, but the uniqueness of the estimator has apparently not been noticed. A simple computational method for obtaining the maximum likelihood estimator is also absent from the existing literature.

The second result obtained here is a procedure for finding conservative $100(1-\alpha)\%$ one and two sided confidence intervals for k. Confidence intervals obtained previously have depended on asymptotic theory in order to justify their validity. This means that for any finite sample size the coverage probability may be either more or less than the nominal level. The conservative confidence intervals here always have at least the specified nominal coverage probability. In limited testing, the intervals found here secure this advantage without sacrificing length.

The sufficiency of the statistic C_n for k is also established here, as is the strong asymptotic convergence of both C_n and the maximum likelihood estimator $\hat{k}_n$ to k. These results round out the theoretical developments connected with this problem.

The case of not equally likely colors presents complications which apparently cannot be treated by the methods presented here.

Basic Formula

The theorems and procedures contained here depend on a single basic formula. This formula was also obtained by Driml and Ullrich (1967).

Before establishing the basic formula, note that the actual colors used play no important role in the problem. The colors serve only as convenient labels to refer to the observed outcomes. For this reason it is useful to refer to a pattern determined by the observations. The pattern determined by the observations consists of the positions at which new colors are observed together with information about the positions which are occupied by the same color. Thus the two observations (red, green, red) and (blue, yellow, blue) correspond to the same pattern which will be denoted (1,2,1).

To establish the basic formula, note that each configuration of n coupons has probability of k^{-n} of being selected. The event $C_n = c$ occurs by first selecting a sequence of c colors from the k available colors, and then using this sequence of colors, in order, to color a pattern of length n containing exactly the integers 1, 2, ... , c. Denote by $f(c,n)$ the

number of such patterns. It is clear that $f(c,n)$ does not depend on k. (In fact, it can be shown that $f(c,n) = \sum\left\{\frac{n!}{r_1!\cdots r_c!} : r_1 \geq 1, \ldots, r_c \geq 1, \sum_{i=1}^{c} r_i = n\right\}$.) For $1 \leq c \leq k$,

$$P_k[C_n = c] = c!\binom{k}{c} f(c,n)\, k^{-n},$$

which is our basic formula. Here and throughout, the notation $P_k[E]$ denotes the probability of the event E computed assuming there are k colors.

Sufficiency of C_n

As a consequence of the basic formula, the statistic C_n will be shown to be a sufficient statistic for k. Practically speaking, this means that C_n is the correct statistic to use when estimating k or finding confidence intervals for k. In particular, the pattern contains no additional information about k, other than the number of observed colors.

As remarked above, the actual colors observed are only considered as arbitrary labels used to make it easier to refer to the outcomes. The actual observation consists only of the pattern. Sufficiency of C_n will be established by showing that $P_k[\text{observed pattern}|C_n = c]$ does not depend on k.

Clearly, if the observed pattern does not contain exactly c distinct colors the conditional probability $P_k[\text{observed pattern}|C_n = c] = 0$. If the pattern does contain exactly c distinct colors then by the discussion above, $P_k[\text{observed pattern}, C_n = c] = c!\binom{k}{c}k^{-n}$. It follows that $P_k[\text{observed pattern}|C_n = c] = P_k[\text{observed pattern}, C_n = c]/P_k[C_n = c] = 1/f(c,n)$, which does not depend on k. The sufficiency of C_n is therefore established.

The Maximum Likelihood Estimator of k

The maximum likelihood estimator of k can be found by using the basic formula above. A derivation similar to that given here (except for the uniqueness assertion) can be found in Driml and Ullrich (1967). Denote by c the observed value of C_n. The objective is to show that when $c < n$ there is a unique finite value of k which maximizes $P_k[C_n = c]$. In the case in which $c = n$ this probability is a strictly increasing function of k, as is easily seen. Therefore when $c = n$ the maximum likelihood estimator of k does not exist.

Suppose $1 \leq c < n$ is fixed. Simple algebra and the basic formula show that

$$\frac{P_{k+1}[C_n = c]}{P_k[C_n = c]} = \frac{k+1}{k+1-c}\left(\frac{k}{k+1}\right)^n.$$

Notice that these formulas only hold for $k \geq c$, which will henceforth be assumed. The maximum likelihood estimator is determined by studying when this ratio is less than and greater than 1. In the special case $c = 1$, this last ratio is clearly < 1 for all k and therefore the unique maximum likelihood estimator of k is c. In the remainder of this discussion it is assumed that $1 < c < n$. Define $g(x) = \frac{1}{1-cx}(1-x)^n$ for $0 \leq x < 1/c$ and observe that $g(1/(k+1)) = \frac{k+1}{k+1-c}\left(\frac{k}{k+1}\right)^n$. Elementary calculus shows that g has a unique critical

point in the interval $(0, 1/c)$ and this critical point corresponds to a minimum of g. Also $g(0) = 1$ and $g(1/c-) = +\infty$. Hence there is a point $0 < x_0 < 1/c$ so that $g(x) \leq 1$ for $0 \leq x \leq x_0$ and $g(x) > 1$ for $x > x_0$. This implies the existence of an integer k_0 with the property that $\frac{P_{k+1}[C_n=c]}{P_k[C_n=c]} \geq 1$ for $c \leq k \leq k_0$ and $\frac{P_{k+1}[C_n=c]}{P_k[C_n=c]} \leq 1$ for $k \geq k_0$. Hence there is at least one value of k at which $P_k[C_n = c]$ attains its maximum value. This establishes the existence of at least one maximum likelihood estimator of k. Uniqueness will be established by showing that there is no integer k for which $\frac{P_{k+1}[C_n=c]}{P_k[C_n=c]} = 1$. From the above discussion, this last equality is possible if and only if $k^n = (k+1)^{n-1}(k+1-c)$. Since k and $k+1$ are relatively prime, this equation could hold only if k divides $k+1-c$. This would imply $c = 1$ and then $n = 1$ too. But the case $c = n$ has been excluded. Thus there is no such integer k. The uniqueness of the maximum likelihood estimator has been established.

The discussion is summarized in the following theorem.

Theorem. *If $C_n < n$ the maximum likelihood estimator $\hat{k}_n(C_n)$ of k is unique, and is the smallest integer $j \geq C_n$ which satisfies $\frac{j+1}{j+1-C_n}\left(\frac{j}{j+1}\right)^n < 1$.*

Here is a short table illustrating the behavior of the maximum likelihood estimator $\hat{k}$. The endpoints of a two sided 95% confidence interval for k, computed using the method below, are also included.

Maximum Likelihood Estimator of k
and Two Sided 95% Confidence Interval Endpoints

n	C_n	$\hat{k}_n(C_n)$	$\mathcal{L}_n(C_n)$	$\mathcal{U}_n(C_n)$
20	5	5	5	7
20	10	12	10	22
20	15	31	17	92
20	19	183	38	7512

Large Sample Behavior

The behavior of the maximum likelihood estimator as the sample size $n \to \infty$ can be obtained from the following result about the asymptotic behavior of C_n.

Theorem. *The sequence $C_n \to k$ almost surely as $n \to \infty$.*

proof. Since $1 \leq C_n \leq k$ for all n and C_n is non-decreasing in n, $P_k[C_n \not\to k] = P_k[\bigcap_{i=1}^{\infty}[C_i < k]] \leq \lim_{i\to\infty} P_k[C_i < k] \leq \lim_{i\to\infty} k(1-1/k)^i = 0$, which proves the theorem. □

Theorem. *The maximum likelihood estimator $\hat{k}_n(C_n) \to k$ almost surely as $n \to \infty$.*

proof. Because of the preceding theorem it is enough to show that $\lim_{n\to\infty} \hat{k}_n(C_n) - C_n = 0$ almost surely. From the earlier discussion of $\hat{k}$ it follows first that $\hat{k}_n(C_n) \geq C_n$. An elementary computation using the criterion for $\hat{k}$ shows that if $n \geq -\ln(C_n+1)/\ln(C_n/(C_n+1))$ then $\hat{k}_n(C_n) = C_n$. □

Confidence Intervals for k

Finding the distribution of the maximum likelihood estimator, or some function of the estimator, is clearly a difficult task. In order to find confidence intervals for k the so-called statistical method will be used. This method is described in Bickel and Doksum (pp. 180-182), as well as in Kendall and Stuart (pp. 114-116, Example 20.2).

As motivation for the method, suppose the null hypothesis $H_0 : k = k_0$ is to be tested against the alternative $H_1 : k > k_0$ at level of significance α. The hypothesis is clearly rejected if C_n is too large. Suppose c is the (fixed) observed value of C_n. One concludes that C_n was too large if $P_{k_0}[C_n \geq c] \leq \alpha$.

This hypothesis testing technique suggests the following method of finding a one sided confidence interval for k. Let k_0 be the largest value of i so that $P_i[C_n \geq c] \leq \alpha$. Then the interval $[k_0, \infty)$ is an intuitively reasonable choice of a confidence interval for k. Note that k_0 depends on the observed value c of C_n and therefore is a random variable. The notes following justify this procedure.

To apply the method it will be helpful to know that each of the probabilities $P_k[C_n \leq j]$ and $P_k[C_n \geq j]$ are monotone functions of k for each fixed j.

Lemma. *In the notation above, $P_k[C_n \geq j]$ is a non-decreasing function of k for each fixed integer $j \geq 1$. Moreover, $\lim_{k\to\infty} P_k[C_n \geq j] = 1$, for $1 \leq j \leq n$.*

proof. While this result is intuitively reasonable, a proof will be given by using a coupling argument. Suppose the number of colors is fixed. Denote by G_i the number of draws required to see a new color after $i-1$ colors have already been observed. The random variable G_i will be defined to be $+\infty$ if i exceeds the number of available colors. Then $G_1 = 1$ and the G's form a sequence of independent random variables. It is also clear that $[C_n \geq j] = [\sum_{i=1}^{j} G_i \leq n]$.

A specific realization of the random variables G_i for two different number of colors, k and k', will now be constructed on a single underlying probability space. Suppose that $k < k'$ are given. Let $\{U_{i,m} : i \geq 1, m \geq 1\}$ be a collection of independent random variables each of which is uniformly distributed on the interval $(0, 1)$. Define $G_i = \min\{m : U_{i,m} \leq (k - i + 1)/k\}$ for $1 \leq i \leq k$ and set $G_i = +\infty$ for $i > k$. Define $G'_i = \min\{m : U_{i,m} \leq (k' - i + 1)/k'\}$ for $1 \leq i \leq k'$ and set $G'_i = +\infty$ for $i > k'$. It is easily verified that since $k < k'$, then $(k - i + 1)/k < (k' - i + 1)/k'$ and so $G_i \geq G'_i$ for all i. The equality of the two events in the previous paragraph leads to the computation

$$\begin{aligned} P_k[C_n \geq j] &= P[\sum_{i=1}^{j} G_i \leq n] \\ &\leq P[\sum_{i=1}^{j} G'_i \leq n] \\ &= P_{k'}[C_n \geq j] \end{aligned}$$

which proves the first assertion of the lemma. The second assertion follows by noting that each of the random variables G_i converges to 1 in probability as $k \to \infty$. $\square$

A similar computation proves the following result.

Lemma. *In the notation above, $P_k[C_n \leq j]$ is a non-increasing function of k for each fixed integer $j \geq 1$. Moreover, $\lim_{k\to\infty} P_k[C_n \leq j] = 0$ for $1 \leq j < n$.*

A one sided confidence interval for k will now be found. To be precise, a statistic $\mathcal{L}_n(C_n)$ will be defined with the property that $P_k[\mathcal{L}_n(C_n) \leq k] \geq 1 - \alpha$ for all $k \geq 1$.

Define a function $\mathcal{L}_n$ with domain $\{1, \ldots, n\}$ by first setting $\mathcal{L}_n(1) = 1$. For $2 \leq j \leq n$ let $\mathcal{L}_n(j)$ be the largest integer i satisfying $P_i[C_n \geq j] \leq \alpha$. This inequality has a unique maximal solution since $P_i[C_n \geq j]$ is a non-decreasing function of i which has the value 0 at $i = 1$ and limit 1 as $i \to \infty$. Notice that $1 = \mathcal{L}_n(1) \leq \mathcal{L}_n(2) \leq \cdots \leq \mathcal{L}_n(n)$.

Theorem. *In the notation above, $P_k[\mathcal{L}_n(C_n) \leq k] \geq 1 - \alpha$ for each $k \geq 1$.*

proof. The theorem will be proved by showing that $P_k[k < \mathcal{L}_n(C_n)] \leq \alpha$ for each $k \geq 1$. The proof proceeds by considering various cases. First suppose $1 = \mathcal{L}_n(1) \leq k < \mathcal{L}_n(2)$. It is clear that for such k

$$[k < \mathcal{L}_n(C_n)] = [C_n \geq 2].$$

Thus

$$P_k[k < \mathcal{L}_n(C_n)] = P_k[C_n \geq 2] \leq P_{\mathcal{L}_n(2)}[C_n \geq 2] \leq \alpha$$

where the next to last inequality follows from the fact that $P_k[C_n \geq 2]$ is an increasing function of k with $P_{\mathcal{L}_n(2)}[C_n \geq 2] \leq \alpha$. Next consider the case $\mathcal{L}_n(2) \leq k < \mathcal{L}_n(3)$. For k in this interval $[k < \mathcal{L}_n(C_n)] = [C_n \geq 3]$, so $P_k[k < \mathcal{L}_n(C_n)] = P_k[C_n \geq 3] \leq \alpha$ since $P_k[C_n \geq 3]$ is an increasing function of k with $P_{\mathcal{L}_n(3)}[C_n \geq 3] \leq \alpha$. A similar argument holds for $\mathcal{L}_n(j-1) \leq k < \mathcal{L}_n(j)$ for $4 \leq j \leq n$. Finally consider the case in which $\mathcal{L}_n(n) \leq k$. For such k, $[k < \mathcal{L}_n(C_n)] = \emptyset$ and $P_k[k < \mathcal{L}_n(C_n)] = 0 \leq \alpha$. This concludes the proof of the theorem. □

The problem of finding a two sided confidence interval will now be examined. Let $\mathcal{L}_n$ be defined as above. Define a function $\mathcal{U}_n$ with domain $\{1, \ldots, n\}$ by first setting $\mathcal{U}_n(n) = \infty$. For $1 \leq j \leq n-1$ let $\mathcal{U}_n(j)$ be the smallest integer i satisfying $P_i[C_n \leq j] \leq \alpha$. This inequality has a smallest solution since $P_i[C_n \leq j]$ is a non-increasing function of i which has the value 1 at $i = \min\{j, n\}$ and limit 0 as $i \to \infty$. Notice that $\mathcal{U}_n(1) \leq \cdots \leq \mathcal{U}_n(n-1) < \mathcal{U}_n(n) = \infty$. An argument parallel to that above can be used to prove the following theorem.

Theorem. *In the notation above, $P_k[k \leq \mathcal{U}_n(C_n)] \geq 1 - \alpha$ for each $k \geq 1$.*

The only difference in the proof of this theorem is that the cases are determined by the inequalities $\mathcal{U}_n(j-1) < k \leq \mathcal{U}_n(j)$.

In order to obtain two sided confidence intervals it is important to establish an order relation between $\mathcal{L}_n(j)$ and $\mathcal{U}_n(j)$. It is clear that $\mathcal{L}_n(1) = 1 \leq \mathcal{U}_n(1)$ and $\mathcal{L}_n(n) < \infty = \mathcal{U}_n(n)$. For $2 \leq j \leq n-1$ note that $P_i[C_n \leq j-1] = 1 - P_i[C_n \geq j]$ for all i. Replacing i by $\mathcal{L}_n(j)$ gives $P_{\mathcal{L}_n(j)}[C_n \leq j-1] \geq 1 - \alpha > \alpha$ for $\alpha < 1/2$. This shows that $\mathcal{L}_n(j) < \mathcal{U}_n(j-1)$ and hence that $\mathcal{L}_n(j) < \mathcal{U}_n(j)$.

Theorem. *In the notation above,* $P_k[\mathcal{L}_n(C_n) \le k \le \mathcal{U}_n(C_n)] \ge 1 - 2\alpha$ *for each* $k \ge 1$.

proof. The theorem is proved by noting that $1 - P_k[\mathcal{L}_n(C_n) \le k \le \mathcal{U}_n(C_n)] = P_k[k < \mathcal{L}_n(C_n)] + P_k[\mathcal{U}_n(C_n) < k] \le \alpha + \alpha = 2\alpha$ by the preceding two theorems. □

It is worth noting that these 3 theorems are optimal in a certain sense. The following theorem and its proof have immediate analogs in the case of upper and two sided confidence intervals.

Theorem. *There is no statistic* $S(C_n)$ *with the property that* $P_k[S(C_n) \le k] = 1 - \alpha$ *for each* $k \ge 1$.

proof. The proof of the theorem is obtained by considering 2 cases. In the first case suppose $S(1) \le 1$. Then when $k = 1$, $P_k[S(C_n) \le k] = P_1[S(C_n) \le 1] = 1 > 1 - \alpha$. In the second case suppose $S(1) > 1$. Then when $k = 1$, $P_k[S(C_n) \le k] = 0 < 1 - \alpha$. This proves the theorem. □

In connection with the confidence interval problem, it would have been nice to show that the confidence intervals obtained always contained the maximum likelihood estimator. This would be the case if $P_k[\mathcal{L}_n(C_n) \le \hat{k}_n(C_n) \le \mathcal{U}_n(C_n)] = 1$ for all $0 < \alpha < 1/2$ and all k. Unfortunately, this property fails to hold. To see this, first observe from the definitions that this property is equivalent to the two assertions $P_{\hat{k}_n(c)}[C_n \le c] \ge 1/2$ and $P_{\hat{k}_n(c)}[C_n \ge c] \ge 1/2$ for all c and n. Numerical computation shows that this second inequality fails for $n = 16$ and $c = 7$.

Comparisons

Numerical comparison of the confidence intervals found by the method described here and previous methods based on asymptotic theory show that the present method has the advantage of maintaining the nominal coverage probability without suffering the drawback of producing overly long intervals. Additionally, the present method applies in all circumstances, while the asymptotic methods depend on the size and relative magnitude of n, C_n, and k for their (approximate) validity.

As an example, the method of Ivchenko and Timonina (1983) will be examined. The validity of their method depends on the asymptotic normality of a statistic which is closely related to $\hat{k}_n$. This asymptotic normality should approximately hold if n and k are large and the ratio k/n is not close to 0 or ∞. Applying their method to the cases considered earlier yields the following intervals with endpoints denoted $\mathcal{L}_n^*$ and $\mathcal{U}_n^*$.

Comparison with Ivchenko and Timonina
Two Sided 95% Confidence Interval Endpoints

n	C_n	$\hat{k}_n(C_n)$	$\mathcal{L}_n(C_n)$	$\mathcal{U}_n(C_n)$	$\mathcal{L}_n^*$	$\mathcal{U}_n^*$
20	5	5	5	7	5	5
20	10	12	10	22	10	19
20	15	31	17	92	20	116
20	19	183	38	7512	67	fails

This demonstrates the favorable comparison of the two methods under circumstances in which both should work well, and the failure of the asymptotic method in certain situations.

Using the Craig (1953) butterfly data and the Darroch (1958) asymptotic variance estimates, Seber (1982) obtains, for $c = 341$ and $n = 435$, an (approximate) 95% confidence interval of [729,1026]. Our algorithm yields a conservative 95% confidence interval of [722, 1029].

Implementation Details

Some simple functions were written in C++ to perform the computations necessary to find the maximum likelihood estimator as well as the upper and lower endpoints of the confidence intervals discussed above. The source code is available from Howard Tucker's home page at `http://www.math.uci.edu/~htucker`.

In order to find the confidence intervals, an effective way of computing the distribution function of C_n is needed. Feller (1957) presents a formula which in the notation above reads

$$P_k[C_n \leq c] = \binom{k}{c} \sum_{i=0}^{c} (-1)^i \binom{c}{i} \left(\frac{c-i}{k}\right)^n \left(\frac{k-c}{k-c+i}\right).$$

This formula is difficult to work with, due to the relative sizes of the terms and the alternating signs. By conditioning on the outcome of the

final draw the recursive formula

$$P_k[C_n = c] = \left(\frac{k-c+1}{k}\right) P_k[C_{n-1} = c-1] + \left(\frac{c}{k}\right) P_k[C_{n-1} = c]$$

is obtained. This formula is numerically stable and was used to compute the distribution function of C_n.

References

P. J. Bickel and K. A. Doksum, *Mathematical Statistics*, Holden-Day, San Francisco, 1977.

C. C. Craig, *Use of Marked Specimens in Estimating Populations*, Biometrika **40** (1953), 170–176.

J. N. Darroch, *The Multiple-Recapture Census. I: Estimation of a Closed Population*, Biometrika **45** (1958), 343–359.

M. Driml and M. Ullrich, *Maximum Likelihood Estimate of the Number of Types*, Acta Technica ČSAV (1967), 300–303.

W. Feller, *Introduction to Probability and Its Applications*, volume 1 second edition, John Wiley, New York, 1957.

Tim Hilchey, *Snapshots to Improve the Tallies of Tigers*, The New York Times, July 30, 1996, B7.

G. I. Ivchenko and E. E. Timonina, *Estimating the Size of a Finite Population*, Theory of Probability and Its Applications **27** (1983), 403–406.

M. G. Kendall and A. Stuart, *Advanced Theory of Statistics*, volume 2 fourth edition, Charles Griffin & Co., Ltd., London, 1979.

G. A. F. Seber, *The Estimation of Animal Abundance*, second edition, Charles Griffin & Co., Ltd., London, 1982.

Department of Mathematics, University of California, Irvine, 92697-3875 and Department of Discrete and Statistical Sciences, Auburn University, 36849-5307