Computational Statistics and Data Analysis (MVComp2)

Solutions to exercise 6

Lecturer Tristan Bereau

Semester Wi23/24
Due Nov. 30, 2023, 23:59

1 NYC taxicabs: frequentist inference (5 points)

While visiting New York city, you realize that each yellow taxicab displays a serial number. You assume that each cab i displays a unique number, x_i, and that they are sequentially numbered starting from 1. Can you infer the total number of taxicabs, N, given a set of k observations, $x_1, x_2, \dots, x_k$?

(a) Show that the conditional probability that the largest serial number observed is $M = m$, given that there are $N = n$ taxicabs and you make $K = k$ observations is given by

$$P(M = m | N = n, K = k) = \begin{cases} \binom{m-1}{k-1}\binom{n}{k}^{-1}, & \text{if } k \leq m \text{ and } m \leq n \\ 0, & \text{otherwise} \end{cases}$$

(b) The expression in (a) is, in fact, the likelihood. Use maximum likelihood estimation to derive an estimator for N as a function of M, denoted $\hat{N}_1(M)$. Is it a biased estimator?

(c) You propose to build a more robust estimator: Estimate the number of unobserved labels that are *above* the largest number observed, M. Assume that this number is equal to the average gap between observations. Show that your estimator for the total population size leads to

$$\hat{N}_2(M) = \frac{k+1}{k} M - 1$$

(d) Use the likelihood in (a) to show that $\hat{N}_2$ is an unbiased estimator.

(e) The variance of the estimator is given by the expression

$$\mathrm{Var}[\hat{N}_2] = \frac{1}{k} \frac{(n-k)(n+1)}{k+2}.$$

In the regime of few observations, show that $\mathrm{Var}[\hat{N}_2]$ behaves in agreement to your assumptions.

1.1 Solution

(a) We're drawing a sample of size k from a population N *without* replacement. This immediately leads to using the Binomial coefficient $\binom{n}{k}$ for the denominator. Furthermore, we order the datapoints according to their serial number:

$$x_1 < x_2 < \cdots < x_k.$$

The probability that the largest serial number is $M = m$ corresponds to freely choosing the other $k-1$ points between the values 1 and $m-1$, $\binom{m-1}{k-1}$.

(b) The likelihood is a monotonically decreasing function of n. Its maximum is at the lowest allowed value, $M = m$. Thus $\hat{N}_1 = m$. Clearly it is a *biased* estimator, because the true population can be more than m, but never below.

(c) Order once again the datapoints according to their serial numbers

$$x_1 < x_2 < \cdots < x_k,$$

where $M = x_k$. Build an estimator for the number of unobserved labels above the largest value

$$\hat{N}_2 - x_k = \frac{(x_1 - 1) + (x_2 - x_1 - 1) + \cdots + (x_k - x_{k-1} - 1)}{k} = \frac{x_k}{k} - 1.$$

This leads to an expression for the estimator

$$\hat{N}_2(M) = M + \frac{M}{k} - 1 = \frac{k+1}{k} M - 1.$$

(d) For the bias we compute the first moment of the distribution, where the highest observed number, x_k, can take on values $k \leq M \leq n$

$$\begin{aligned}
E[M] &= \sum_{j=k}^{n} j P(M = j | N = n, K = k) \\
&= \frac{1}{\binom{n}{k}} \sum_{j=k}^{n} j \binom{j-1}{k-1} \\
&= \frac{1}{\binom{n}{k}} \sum_{j=k}^{n} j \frac{(j-1)!}{(k-1)!(j-k)!} \\
&= \frac{k}{\binom{n}{k}} \sum_{j=k}^{n} \frac{j!}{k!(j-k)!} \\
&= \frac{k}{\binom{n}{k}} \sum_{j=k}^{n} \binom{j}{k}
\end{aligned}$$

but the normalization condition dictates

$$\begin{aligned}
\sum_{j=k}^{n} P(M = j | N = n, K = k) &\stackrel{!}{=} 1 \\
\sum_{j=k}^{n} \binom{j-1}{k-1} \binom{n}{k}^{-1} &= 1 \\
\sum_{j=k}^{n} \binom{j-1}{k-1} &= \binom{n}{k} \\
\sum_{j=k}^{n} \binom{j}{k} &= \binom{n+1}{k+1}.
\end{aligned}$$

Plugging this back into the expected value, we obtain

$$\begin{aligned} E[M] &= \frac{k}{\binom{n}{k}} \sum_{j=k}^{n} \binom{j}{k} \\ &= \frac{k}{\binom{n}{k}} \binom{n+1}{k+1} \\ &= \frac{k}{k+1}(n+1). \end{aligned}$$

This leads to the expression for the bias

$$\begin{aligned} b &= E[\hat{N}_2] - N \\ &= E\left[\frac{k+1}{k}M - 1\right] - N \\ &= \frac{k+1}{k}\frac{k}{k+1}(N+1) - 1 - N \\ &= 0. \end{aligned}$$

(e) In the regime of few observations, we have $k \ll n$, such that

$$\mathrm{Var}[\hat{N}_2] \approx \frac{(n-k)n}{k^2} \approx \frac{n^2}{k^2}.$$

The standard deviation is approximately n/k, which is the expected spacing of the gap between sorted observations.

2 NYC taxicabs: Bayesian inference (5 points)

Let's solve the same problem as in question 1, but using Bayesian inference. We want to use the likelihood in question 1 (a), together with an improper uniform prior over N, while fixing $K = k$.[1]

(a) Show that the posterior distribution, $P(N = n|M = m, K = k)$, is given by

$$P(N = n|M = m, K = k) = \frac{k-1}{m}\binom{m}{k}\binom{n}{k}^{-1}.$$

Hint: you may find the following Binomial coefficient identity useful

$$\sum_{a=j}^{\infty} \binom{a}{b}^{-1} = \frac{b}{b-1}\frac{1}{\binom{j-1}{b-1}}.$$

(b) What is the maximum a-posteriori estimator?

[1] An *improper* uniform prior is not bounded, and as such does not strictly speaking integrate to 1 over its domain. A more rigorous approach would consist of taking a large, but finite, interval. You can then show that the final result does not depend on the value of the bound. We will not do that here.

(c) The posterior, $P(n|m,k)$, in fact corresponds to a shifted factorial distribution, such that $N - m \sim \text{Fact}(k,m)$. A random variable, Z, follows a factorial distribution with parameters n and m, i.e., $Z \sim \text{Fact}(n,m)$, such that

$$P(Z=z) = (n-1)\frac{(m-1)!}{(m-n)!}\frac{(m+z-n)!}{(m+z)!}.$$

One can show that the expected value of Z is given by $E[Z] = \frac{m-n+1}{n-2}$. Show that the posterior mean is given by

$$\bar{N} = E[P(n|m,k)] = \frac{k-1}{k-2}(m-1).$$

(d) Consider the following sequence of serial numbers: $\boldsymbol{x} = (41, 60, 17, 42)$. Compare the frequentist estimator, $\hat{N}_2$ in question 1 (c) with the present posterior mean, $\bar{N}$. Comment on the difference. What might be a more appropriate quantity for the posterior to better match the frequentist inference?

2.1 Solution

(a) Bayes' rule allows us to relate likelihood and prior to the posterior

$$\begin{aligned} P(N=n|M=m,K=k) &= \frac{P(M=m|N=n,K=k)P(N=n)}{P(M=m)} \\ &= \frac{P(m|n)P(n)}{\sum_{n'=m}^{\infty} P(m|n')P(n')} \\ &= \frac{P(m|n)}{\sum_{n'=m}^{\infty} P(m|n')}. \end{aligned}$$

We compute the denominator

$$\sum_{n'=m}^{\infty} P(m|n') = \binom{m-1}{k-1}\sum_{n'=m}^{\infty}\binom{n}{k}^{-1} = \binom{m-1}{k-1}\frac{k}{k-1}\frac{1}{\binom{m-1}{k-1}} = \frac{k}{k-1},$$

which allows us to simplify the expression for the posterior distribution

$$\begin{aligned} P(N=n|M=m,K=k) &= \frac{P(m|n)}{\sum_{n'=m}^{\infty} P(m|n')} \\ &= \frac{k-1}{k}\binom{m-1}{k-1}\binom{n}{k}^{-1} \\ &= \frac{k-1}{m}\binom{m}{k}\binom{n}{k}^{-1} \end{aligned}$$

(b) The MAP corresponds to the maximum likelihood estimator, which is $\hat{N} = M$ (same as in problem 1 (b)).

(c) The shift in the factorial distribution means that we need to compute

$$E[P(n|m,k)] = E[Z] + m = \frac{m-k+1}{k-2} + m = \frac{m-k+1+mk-2m}{k-2} = \frac{k-1}{k-2}(m-1).$$

(d) Given the sequence $\boldsymbol{x} = (41, 60, 17, 42)$ we have

$$\hat{N}_2 = \frac{k+1}{k} m - 1 = \frac{5}{4} 60 - 1 = 74.$$

On the other hand the posterior mean yields

$$\bar{N} = \frac{k-1}{k-2}(m-1) = \frac{3}{2}(60-1) = 88.5.$$

The significant difference is due to the skew of the posterior distribution. A more appropriate quantity to match the frequentist inference might be the median, which would be closer to $M = m$.