Computational Statistics and Data Analysis (MVComp2)

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, $x_1, x_2, \ldots, x_k$?

(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.

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?

(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?

[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.