4B Parameter estimation

About notation. Here we will use notation like $f_\lambda(x)$ to mean "the density function of the given form, when its **parameter** has value λ". Note that the subscript now refers to the parameter and *not* the random variable, like in $f_X(x)$. You must understand from context how the subscript is meant. — There are also other ways of showing the parameter; Ross writes $f(x \mid \lambda)$, and some people write $f(x\,;\,\lambda)$. Varying notation is a fact of life.

Class problems

4B1 (Service requests) A computing server receives service requests at random intervals. The intervals between each two consecutive requests are independent, and follow the density function

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

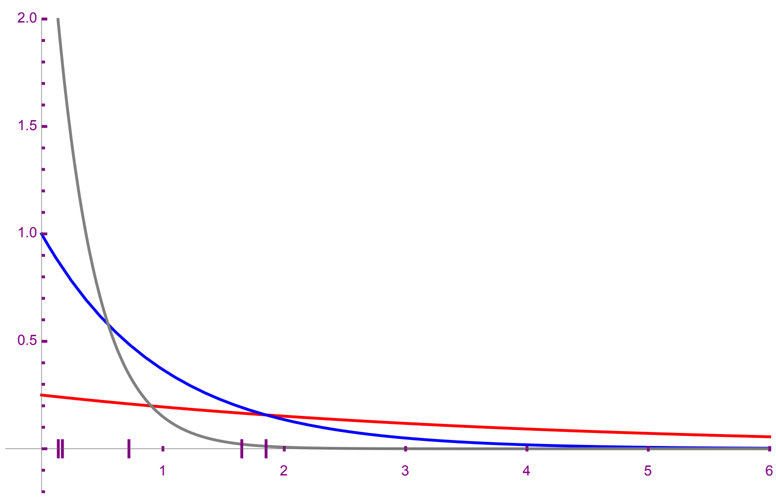
where $\lambda > 0$ is an unknown parameter. We have measured the intervals $0.16, 1.85, 0.15, 0.72, 1.65$.

(a) In the graph below, the measured values are marked on the x axis as small bars. There are also three proposed density functions for the data, corresponding to the parameter values $\lambda = 0.25$ (red), $\lambda = 1.00$ (blue) and $\lambda = 3.00$ (gray).

By *looking* at the data, give your opinion on which of the three proposed density functions might the best match for (the empirical distribution of) the data.

(b) Find the maximum likelihood estimate for the parameter λ.

The likelihood function $L(\lambda)$ and its logarithm $\ell(\lambda) = \log(L(\lambda))$ are maximized at the same value of λ, so you can use either function. The latter may have more convenient derivatives.



4B2 (Serial numbers) Battle tanks of a foreign army are numbered serially $1, 2, \ldots, b$. Our observers have seen tanks and their serial numbers four times: $x_1 = 13$, $x_2 = 77$, $x_3 = 111$ and $x_4 = 145$. We assume each time the tank was uniformly randomly chosen from the b tanks that the enemy has, but b is unknown to us. (See lecture 4A.)

(a) Based on the data, is it possible that $b = 140$? Is it possible that $b = 200$?

(b) If the enemy has b tanks and $b < 145$, what is the probability that you observe these particular four serial numbers?

(c) If the enemy has b tanks and $b \geq 145$, what is the probability that you observe these particular four serial numbers? (The answer is an expression that contains b in some form.)

(d) Based on what you now know, write down the likelihood function $L(b)$, when b is an arbitrary positive integer. Hint: You want to break it into two cases.

(e) By looking at the expression of $L(b)$, find where it has its maximum value. In other words, find the maximum likelihood estimate $\hat{b}$.

(f) Generalize: What is the maximum likelihood estimate $\hat{b}(\vec{x})$ if we have seen n tanks $(x_1, \ldots, x_n)$?

(g) If we have seen only one tank with serial number x_1, what is the maximum likelihood estimate for b? Do you think this is a good estimate?

Home problems

4B3 (Continuous uniform distribution) Our data are coming from a uniform distribution over the interval $[0, b]$, which has density

$$f_b(x) = \begin{cases} \frac{1}{b}, & 0 \leq x \leq b, \\ 0, & \text{elsewhere.} \end{cases}$$

The parameter b is an unknown (but positive) real number.

(a) We have seen data five data points $(1.3, 1.9, 3.6, 1.1, 5.1)$. Write down the likelihood function $L(b)$ (hint: two cases). Plot it by hand or computer with b varying e.g. from 1 to 10. Explain in words: How is the function shaped and why? Hint: This is similar to the battle tank problem but now b need not be integer. Be careful with the density function: when is it zero and when is it nonzero?

(b) Find the maximum likelihood estimate $\hat{b}$ for our data.

(c) Generalize to any data of any size: If we have seen data $\vec{x} = (x_1, x_2, \ldots, x_n)$, what is the maximum likelihood estimate of b?

(d) Let b be fixed (but unknown). Suppose we only observe only one data point. Treat the data point as a random number X_1 coming from the uniform distribution over $[0, b]$. What is its expected value? What is the expected value of our ML estimator $\hat{b}(X_1)$? Is our estimator biased or unbiased?

(e) Another, quite different estimator could be defined as follows. We know that the expected value of the (unknown) generating distribution is $\mu = b/2$. If we use the sample mean $m(\vec{x})$ as an estimate of μ, then it makes sense that $2m(\vec{x})$ would be an estimate of $b = 2\mu$.

So let's define our new estimator

$$\tilde{b}(\vec{x}) = 2m(\vec{x}) = \frac{2}{n} \sum_{i=1}^{n} x_i.$$

Find out whether the estimator $\tilde{b}(\vec{X})$ is unbiased, when each observation in $\vec{X} = (X_1, \ldots, X_n)$ comes from the uniform distribution. (Hint: Linearity of expectation.) Does this estimator seem reasonable?

(f) Using the estimator $\tilde{b}$ from (e), estimate b from data $\vec{x} = (2, 3, 16)$. Is the estimate reasonable?

4B4 (Fitting a geometric distribution) A random variable X has *geometric distribution* with parameter p, that is, it has density

$$f_p(x) = \begin{cases} p(1-p)^x, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise.} \end{cases}$$

Application: X is obtained when an experiment has a constant probability p of succeeding, each time; the experiment is repeated until it succeeds, and we count the number of failures. For example, tossing a coin until heads is obtained, or asking random people until you find a supporter of party P.

From this distribution, we have three independent observations $x_1 = 5$, $x_2 = 3$ and $x_3 = 10$. Find the maximum likelihood estimate for the parameter p. Looking at the value of p, explain what kind of experiment might have produced the data.

Using the logarithmic likelihood is probably convenient here.