MS-A0503 First course in probability and statistics
Department of mathematics and systems analysis
Aalto SCI

J Kohonen
Spring 2021
Exercise 4B

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

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

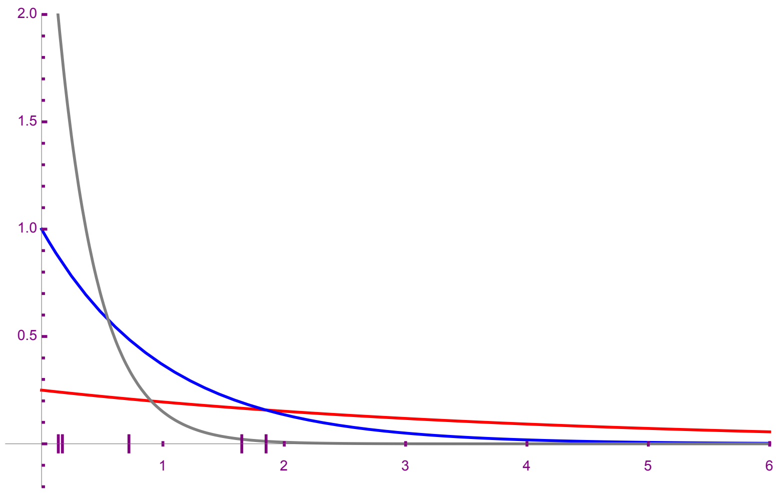
where $\lambda > 0$ is an unknown parameter. We have measured the intervals $0.16, 1.85, 0.15, 0.72, 1.65$.

(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.



Solution.

(a) (One possible way of arguing.) Three of the five data points are in interval $[0, 1]$, so the density function should be such that it gives reasonable (roughly $3/5$) area for this interval. The remaining two points are in $[1.5, 2]$ so we should also have reasonable area (roughly $2/5$) in this interval. Comparing the three proposals, the red curve seems to have far too small areas here; and the gray curve seems a bad fit in the $[1.5, 2]$ interval. The blue curve seems the most reasonable.

(b) Let the observations be $x_1, \ldots, x_5$. The likelihood function (given this data) is

$$L(\lambda) = \prod_{i=1}^{5} \lambda e^{-\lambda x_i} = \lambda^5 e^{-\lambda(x_1+\cdots+x_5)}$$

The maximum of $L(\lambda)$ is at the same λ as the maximum of the logarithmic likelihood $\ell(\lambda) = \log L(\lambda)$. The logarithmic likelihood is

$$\ell(\lambda) = \log\left(\lambda^5 e^{-\lambda(x_1+\cdots+x_5)}\right) = 5\log(\lambda) - \lambda(x_1 + \cdots + x_5),$$

and its first two derivatives are

$$\ell'(\lambda) = 5\lambda^{-1} - (x_1 + \cdots + x_5).$$

and

$$\ell''(\lambda) = -5\lambda^{-2}.$$

The first derivative is zero only at one point,

$$\lambda = \frac{5}{x_1 + \cdots + x_5}.$$

Because $\ell''(\lambda) \le 0$ for all $\lambda > 0$, the ℓ is maximized at the zero point of its first derivative. This is also where the likelihood function $L(\lambda)$ is maximized. So the maximum likelihood estimate for λ is

$$\hat{\lambda}(\vec{x}) = \frac{5}{x_1 + \cdots + x_5} = \frac{5}{0.16 + 1.85 + 0.15 + 0.72 + 1.65} \approx 1.104.$$

Compare the result to what we argued in (a).

In general, if the density function has the form given in this exercise (you may recognize it is the exponential distribution), then for any n-element data set $\vec{x} = (x_1, \ldots, x_n)$ we will have $\hat{\lambda}(\vec{x}) = 1/m(\vec{x})$, where $m(\vec{x}) = \frac{1}{n}(x_1 + \cdots + x_n)$. You may want to explain to yourself why and how this makes sense, given that λ is the "rate" parameter of the process.

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) From the previous observations, 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 esimate?

Solution.

(a) Since we saw a tank with number 145, surely b cannot be smaller than that. In particular it cannot be 140. But b could be 200.

(b) If $b < 145$, we cannot observe a tank with number 145, so our data has probability zero.

(c) Each time we observe a tank, its serial number is one of the numbers $1, \ldots, b$ with equal probabilities, $1/b$ for each possible value. In particular, $P(x_1 = 13) = 1/b$, and also $P(x_2 = 77) = 1/b$, and $P(x_3 = 111) = 1/b$, and $P(x_4 = 145) = 1/b$. (Recall that here we assumed $b \geq 145$ so all these four serial numbers were actually possible.) So the probability of our particular data is $1/b^4$.

(d)

$$L(b) = \begin{cases} 0 & \text{if } b < 145 \\ 1/b^4 & \text{if } b \geq 145 \end{cases}$$

(e) From the functional expression, we directly see that $L(b) > 0$ when $b \geq 145$, so the maximum point is somewhere here. Also, if $b \geq 145$, then $1/b^4$ decreases when b increases, so clearly the maximum occurs at $b = 145$. (You could plot some of the values to see how L behaves.) So the ML estimate is $\hat{b} = 145$.

(f) Let $M = \max\{x_1, \ldots, x_n\}$ be the biggest serial number we have seen. If $b < M$, then $L(b) = 0$ because we cannot see serial number M if the enemy only has b tanks.

If $b \geq M$, then $L(b) = 1/b^n$, which is positive but decreases as b increases. So the ML estimate is $\hat{b} = M$.

(g) In this case $\hat{b} = x_1$. If we observe serial number 130, our ML estimate is that the enemy has 130 tanks and we just happened to see their biggest-numbered tank. This seems a bit odd.

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 this estimator is unbiased. (Hint: Linearity of expectation.) Does this estimator seem reasonable?

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

Grading. 1/3 points per item.

Solution.

(a) We recall that $L(b) = f_b(x_1)f_b(x_2)\dots f_b(x_5)$, the product of densities at the data points. Let $M = \max x_1, \dots, x_n$ be the biggest value we observed. If $M > b$, then at least one data point has density zero, so $L(b) = 0$. If $M \leq b$, then every data point is in the interval $[0, b]$ and has density $1/b$, so $L(b) = (1/b)^n$.

In our case, we have $n = 5$ and $M = 5.1$, so

$$L(b) = \begin{cases} 0 & \text{if } b < 5.1 \\ 1/b^5 & \text{if } b \geq 5.1 \end{cases}$$

Plot not shown here. The function is zero for $b < 5.1$ because such values are impossible. Then it jumps to $1/5.1^5$ at $b = 5.1$, then decreases as b increases.

(b) Based on what we said in (a), $L(b)$ is maximized at $b = M = 5.1$.

(c) By the same reasoning as before, $L(b) = 0$ for $b < M$, and $L(b) = 1/b^n$ for $b \geq M$. Thus $\hat{b}(\vec{x}) = M = \max\{x_1, \dots, x_n\}$. In words: The ML estimator of b is simply the value of the biggest data point.

(d) Since X_1 is uniform over $[0, b]$, we have $\mathrm{E}(X_1) = b/2$. With $n = 1$ our estimator is simply our data point, $\hat{b}(X_1) = X_1$, so $\mathrm{E}(\hat{b}(X_1)) = \mathrm{E}(X_1) = b/2$. Our estimator is biased seriously downwards because $b/2 < b$.

(e) With our new estimator,

$$\begin{aligned} \mathrm{E}(\tilde{b}) &= \mathrm{E}\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{2}{n} \cdot \sum_{i=1}^{n} \mathrm{E}(X_i) \\ &= \frac{2}{n} \cdot n \cdot \frac{b}{2} \\ &= b, \end{aligned}$$

so our new estimator is unbiased, which is nice.

This kind of estimator is called a *moment estimator*. First we estimate some of the “moments” of the unknown distribution, here just the expected value. Then we find what the unknown distribution must be, to have those moments.

(f) $\tilde{b}(2, 3, 16) = \frac{2}{3}(2 + 3 + 16) = 14$. The estimated value does not really make sense because one of the data points is bigger; we cannot have obtained 16 from a uniform distribution over $[0, 14]$.

It seems that it is not straightforward to define an estimator for the endpoint of a uniform distribution, so that the estimator would make sense in all cases. Our first (maximum-likelihood) estimator is downbiased, sometimes very badly. Our second (moment-based) estimator is unbiased, but can give impossible results. There are solutions for this, see e.g. Ross Example 7.7c, and Wikipedia: German tank problem.

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

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

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.

Grading.
+1 p for forming the correct likelihood function.
+1 p for finding the maximum point. (For the point, it is enough to find where the first derivative is zero — looking at the second derivative is not required.)

Solution. The likelihood function is

$$\begin{aligned} L(p) &= f_p(5) \cdot f_p(3) \cdot f_p(10) \\ &= p(1-p)^5 \cdot p(1-p)^3 \cdot p(1-p)^{10} \\ &= p^3(1-p)^{18}, \end{aligned}$$

and the logarithmic likelihood is

$$\ell(p) \;=\; \log L(p) \;=\; 3 \log p + 18 \log(1-p).$$

The derivative of ℓ is

$$\ell'(p) \;=\; \frac{3}{p} - \frac{18}{1-p}.$$

The derivative is zero when

$$\frac{3}{p} \;=\; \frac{18}{1-p} \qquad \text{that is,} \qquad p \;=\; 1/7.$$

This point is indeed a maximum point of the logarithmic likelihood function, because the second derivative $\ell''(p) = -\frac{3}{p^2} - \frac{18}{(1-p)^2}$ is negative for all $0 < p < 1$, so the ℓ function is curving down.

Because logarithm preserves order, this point is also where the (non-logarithmic) likelihood function takes its maximum. So the maximum likelihood estimate for p is $1/7$.

(Possible) examples of processes that might produce such data:

- Spin a roulette that has seven slots numbered $1 \ldots 7$, until we get our lucky number (whatever it is). First we spun the roulette 5 times without luck, giving $x_1 = 5$, and got the lucky number on the sixth spin. Then we went on, and again spun 3 times without luck, giving $x_2 = 3$, and got the lucky number on the fourth spin. Finally, we spun 10 times without luck, giving $x_3 = 10$, and got the lucky number on the eleventh spin.

- In large population, $1/7$ of people have a certain property. We wish to find three such people for our medical experiments. Because we have no other information of where to find these people, we pick random people repeatedly until we have the three persons we need. This time, counting both failures and successes, we gathered $5+1+3+1+10+1 = 21$ people and found the three we need.

Of course, the same data could have been obtained even if the true population parameter was, say, 0.12 or 0.15. The value $p = 1/7 \approx 0.143$ is just the value that has the *highest* probability of generating this data. It may be instructive to compare the likelihoods of nearby values of p, and see that they are not very different!