eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Seminar 2.5

1 Synopsis.

What we are set about to do. This activity will use a famous historical example on the use of Statistics to estimate the size of a population. You will face an estimation problem which nicely illustrates some properties of estimators.

What you need to know. In order to benefit from this activity you need to know:

1. What an estimator is, and the meaning of unbiasedness, consistency and regularity.
2. When an estimator extracts all available information in the sample about a given parameter (i.e., when it is *sufficient*).
3. How to use the method of maximum likelihood to construct estimators, and what the properties of such estimators are.

2 Context.

2.1 The problem

In the period between the First and Second World Wars, France, in particular, relied on an static conception of defense, whose backbone was a line of fortifications known as the Maginot line. Few had the expertise and foresight to anticipate the shortcomings of such a defense policy[1].

In 1939 and 1940, the Germans were able to overrun the defences of Poland an France leveraging on the strength of their armored divisions, which rendered static defense quite ineffectual. France was defeated in a matter of days, and the superiority of the German armored divisions remained absolute in Europe for some time. Understandably, in 1940-1942

a major concern of the Allied Command was to estimate the number of tanks that Germany could deploy.

As it turns out, the efforts of the Allied intelligence services were easily foiled by the German counter-intelligence. Then it was remarked that tanks captured or destroyed in combat had serial numbers, apparently assigned in sequence. It was thought that perhaps from the sample of numbers observed something might be inferred about the population of serial numbers, thus arriving at an estimate of the total number of tanks in existence. Your work will consist in recreating that inference process, in a suitably simplified framework.

2.2 A simplified setup

Let θ be at any given moment the total number of tanks of a given class in existence. Let $x_1, \ldots, x_n$ the serial numbers of $n < \theta$ units destroyed in combat or captured. One may think of these as a random sample without replacement from the set $\{0, 1, 2, \ldots, \theta\}$. *Nothing much is lost if we instead think of $x_1, \ldots, x_n$ as a random sample from a continuous uniform distribution, $U(0, \theta)$.* A further simplification you can make is to assume the observed serial numbers independent (which is not true if serial numbers are assigned once: you would never observe two identical serial numbers).

Your task is to estimate θ and answer some questions about your estimator along the way.

3 Questions.

1. We have seen in class that $\hat{\theta} = 2\overline{X}$ is the moment estimator of θ.
   (a) If we had a sample of only $n = 3$ observations, $x_1 = 1$, $x_2 = 5$ and $x_3 = 28$, what would be the moment estimate of θ?
   (b) Is there an *obvious* way of improving on such an estimate? (HINT: Without having any resort to the notion of sufficiency, that we will see along the way, what does the single observation $x_3 = 28$ tell you about the possible values of θ?)
2. What is the density function of any observation coming from a $U(0, \theta)$?
3. Assume the serial numbers you have observed are: $x_1 = 134$, $x_2 = 213$, $x_3 = 58$, $x_4 = 188$, $x_5 = 312$, $x_6 = 333$, $x_7 = 17$, $x_8 = 258$, $x_9 = 358$, $x_{10} = 281$.
   (a) What is the likelihood function?
   (b) Where does the likelihood function have its maximum? What is the maximum likelihood estimate of θ? What would be the moment estimator of θ?
   (c) What properties do the moment and MLE have among these: (i) Unbiasedness, (ii) Sufficiency, (iii) Consistency, (iv) Regularity.
4. The MLE has certain optimality properties, like minimum variance among regular unbiased estimators (i.e. unbiased regular MLE attain the Cramer-Rao lower bound). Explain why this does not help you much in establishing optimality of the MLE in this particular case.

5. To choose among the moment estimator or MLE in this case, you will need to work out their properties. You have two routes, and you can choose either one (but the teacher advises you to follow both!). The next two questions sketch route 1 and route 2 respectively.

6. Route 1: using R.

   (a) Assume $\theta = 350$ (your best estimate with the sample above will not be far from it). Generate a large number (for instance, $N = 1000$) of samples of size $n = 10$ from a $U(0, \theta = 350)$[2].

   (b) With each of the samples generated, compute the moment estimate $\hat{\theta}_M$ and the MLE, $\hat{\theta}_{MLE}$.

   (c) For each of the estimates compute $(\hat{\theta}_M - 350)^2$ and $(\hat{\theta}_{MLE} - 350)^2$. These are the squared errors of estimation of each estimator[3]. Save them.

   (d) Compare the $N = 1000$ squared errors after you are finished: you could for instance draw the histogram of the squared errors for the two estimators[4], compute their average[5], variance[6], draw side by side two boxplots[7], or do something else of your choice. You would then choose the estimator which in the $N = 1000$ simulated experiments appears to give the smallest squared error[8].

7. Route 2: with paper and pencil.

   (a) Find the mean and variance of the moment estimator (trivial). Find $E(\hat{\theta}_M - 350)^2)$.

   (b) Find the mean and variance of the MLE estimator[9], and compute also $E(\hat{\theta}_{MLE} - 350)^2)$. (HINT: remember that $E(X - c)^2 = E(X - E(X))^2 + (E(X) - c)^2$.)

   (c) Compare both.

8. Having made your choice among $\hat{\theta}_M$ and $\hat{\theta}_{MLE}$, on the light of your answer to questions 6 and/or 7, compute from the sample in question 3 your best estimate of the number of tanks in existence.

Recommended reading Using only the hints provided and the theory developed in class you should be equal to the tasks proposed. However, you might want to skim §1.5 of [6] or Chapter 8 in [2] (both relevant to question 7); or, for higher level presentations, books such as [1], [3], or [4].

If you are curious about the original problem (which we have conveniently simplified here) you may look in the Wikipedia a full description and solution, which drops both simplifications we have introduced (independent sampling and continuity of the uniform distribution). It is however of a level higher than required here, and of not much use to answer the questions above.

You may also look alternative discussions, such as in "The locomotive problem", [5], p. 10 (Problem 41).

Notes

[1] A notable exception was general de Gaulle, later to become President of France.

[2] Generating one of those random samples can be done in R invoking `runif(n=10,min=0,max=350)`. To do it $N = 1000$ times, you have to place that instruction inside a `for` loop.

[3] There is nothing sacred with the squared error; you could just as well compute $|\hat{\theta}_M - 350|$ and $|\hat{\theta}_{MLE} - 350|$; but computing the squared errors will make easy to compare with results of the next question.

[4] Using R function `hist`.

[5] With R function `mean`.

[6] With R function `var`.

[7] With R function `boxplot.`

[8] Or any other loss function you choose to compare your estimators.

[9] Make use of the fact that for the largest, $X_{(n)}$, among n sampled values, $F_{X_{(n)}}(s) = P(X_{(n)} \leq s) = P(\cap_{i=1}^{n} [X_i \leq s])$. With the distribution function you can easily obtain the density, the mean, the variance, and any other moment of $X_{(n)}$.

References

[1] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1979 edition, 1974.

[2] A. Garín and F. Tusell. *Problemas de Probabilidad e Inferencia Estadística*. Ed. Tébar-Flores, Madrid, 1991.

[3] P. H. Garthwaite, I. T. Jolliffe, and B. Jones. *Statistical Inference*. Prentice Hall, London, 1995.

[4] J. C. Kiefer. *Introduction to Statistical Inference*. Springer-Verlag, New York, 1987 edition, 1983.

[5] Frederick Mosteller. *Fifty Challenging Problems In Probability With Solutions*. Dover Publications, 1987.

[6] L. Ruíz-Maya and F.J. Martín-Pliego. *Fundamentos de Inferencia Estadística*. Thomson - Paraninfo, 2005. Signatura: AL-519.23 RUI.