Estimation

SARA A. VAN DE GEER

Volume 2, pp. 549–553

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Estimation

In the simplest case, a data set consists of observations on a single variable, say real-valued observations. Suppose there are n such observations, denoted by $X_1, \ldots, X_n$. For example, X_i could be the reaction time of individual i to a given stimulus, or the number of car accidents on day i, and so on. Suppose now that each observation follows the same probability law P. This means that the observations are relevant if one wants to predict the value of a new observation X (say, the reaction time of a hypothetical new subject, or the number of car accidents on a future day, etc.). Thus, a common underlying distribution P allows one to generalize the outcomes.

The emphasis in this paper is on the data and estimators derived from the data, and less on the estimation of population parameters describing a model for P. This is because the data exist, whereas population parameters are a theoretical construct (*see* **Model Evaluation**). An estimator is any given function $T_n(X_1, \ldots, X_n)$ of the data. Let us start with reviewing some common estimators.

The Empirical Distribution. The unknown P can be estimated from the data in the following way. Suppose first that we are interested in the probability that an observation falls in A, where A is a certain set chosen by the researcher. We denote this probability by $P(A)$. Now, from the frequentist point of view, a probability of an event is nothing else than the limit of relative frequencies of occurrences of that event as the number of occasions of possible occurrences n grows without limit. So, it is natural to estimate $P(A)$ with the frequency of A, that is, with

$$\begin{aligned}\hat{P}_n(A) &= \frac{\text{number of times an observation } X_i \text{ falls in } A}{\text{total number of observations}} \\ &= \frac{\text{number of } X_i \in A}{n}. \end{aligned} \tag{1}$$

We now define the empirical distribution $\hat{P}_n$ as the probability law that assigns to a set A the probability $\hat{P}_n(A)$. We regard $\hat{P}_n$ as an estimator of the unknown P.

The Empirical Distribution Function. The distribution function of X is defined as

$$F(x) = P(X \leq x), \tag{2}$$

and the empirical distribution function is

$$\hat{F}_n(x) = \frac{\text{number of } X_i \leq x}{n}. \tag{3}$$

Figure 1 plots the distribution function $F(x) = 1 - 1/x^2$, $x \geq 1$ (smooth curve) and the empirical distribution function $\hat{F}_n$ (stair function) of a sample from F with sample size $n = 200$.

Sample Moments and Sample Variance. The theoretical mean

$$\mu = E(X), \tag{4}$$

(E stands for Expectation), can be estimated by the sample average

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}. \tag{5}$$

More generally, for $j = 1, 2, \ldots$ the jth sample moment

$$\hat{\mu}_{j,n} = \frac{X_1^j + \cdots + X_n^j}{n}, \tag{6}$$

is an estimator of the jth moment $E(X^j)$ of P (*see* **Moments**).

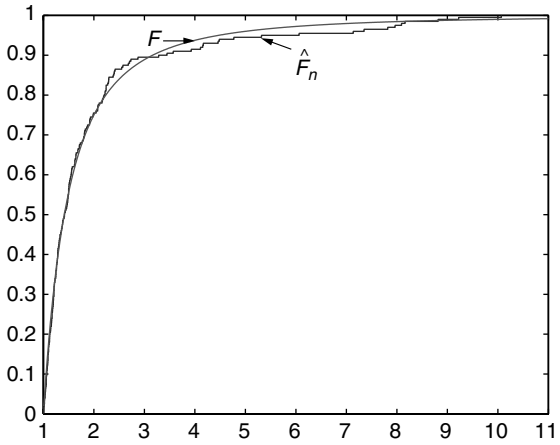

Figure 1 The empirical and theoretical distribution function

The sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \tag{7}$$

is an estimator of the variance $\sigma^2 = E(X-\mu)^2$.

Sample Median. The median of X is the value m that satisfies $F(m) = 1/2$ (assuming there is a unique solution). Its empirical version is any value $\hat{m}_n$ such that $\hat{F}_n(\hat{m}_n)$ is equal or as close as possible to 1/2. In the above example $F(x) = 1 - 1/x^2$, so that the theoretical median is $m = \sqrt{2} = 1.4142$. In the ordered sample, the 100th observation is equal to 1.4166, and the 101th observation is equal to 1.4191. A common choice for the sample median is taking the average of these two values. This gives $\hat{m}_n = 1.4179$.

Parametric Models. The distribution P may be partly known beforehand. The unknown parts of P are called *parameters* of the model. For example, if the X_i are yes/no answers to a certain question (the binary case), we know that P allows only two possibilities, say 1 and 0 (yes = 1, no = 0). There is only one parameter, say the probability of a yes answer $\theta = P(X = 1)$. More generally, in a parametric model, it is assumed that P is known up to a finite number of parameters $\theta = (\theta_1, \ldots, \theta_d)$. We then often write $P = P_\theta$. When there are infinitely many parameters (which is, for example, the case when P is completely unknown), the model is called nonparametric.

If $P = P_\theta$ is a parametric model, one can often apply the maximum likelihood procedure to estimate θ (*see* **Maximum Likelihood Estimation**).

Example 1 The time one stands in line for a certain service is often modeled as exponentially distributed. The random variable X representing the waiting time then has a density of the form

$$f_\theta(x) = \theta \mathrm{e}^{-\theta x}, \ x > 0, \tag{8}$$

where the parameter θ is the so-called *intensity* (a large value of θ means that - on average - the waiting time is short), and the maximum likelihood estimator of θ is

$$\hat{\theta}_n = \frac{1}{\bar{X}_n}. \tag{9}$$

Example 2 In many cases, one assumes that X is normally distributed. In that case there are two parameters: the mean $\theta_1 = \mu$ and the variance $\theta_2 = \sigma^2$. The maximum likelihood estimators of (μ, σ^2) are $(\hat{\mu}_n, \hat{\sigma}_n^2)$, where $\hat{\mu}_n = \bar{X}_n$ is the sample mean and $\hat{\sigma}_n^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2/n$.

The Method of Moments. Suppose that the parameter θ can be written as a given function of the moments $\mu_1, \mu_2, \ldots$. The *methods of moments estimator* replaces these moments by their empirical counterparts $\hat{\mu}_{n,1}, \hat{\mu}_{n,2}, \ldots$.

Example 3 Vilfredo Pareto [2] noticed that the number of people whose income exceeds level x is often approximately proportional to x^θ, where θ is a parameter that differs from country to country. Therefore, as a model for the distribution of incomes, one may propose the Pareto density

$$f_\theta(x) = \frac{\theta}{x^{\theta+1}}, \ x > 1. \tag{10}$$

When $\theta > 1$, one has $\theta = \mu/(\mu - 1)$. Hence, the method of moments estimator of θ is in this case $t_1(\hat{P}_n) = \bar{X}_n/(\bar{X}_n - 1)$. After some calculations, one finds that the maximum likelihood estimator of θ is $t_2(\hat{P}_n) = (n/\sum_{i=1}^{n}\log X_i)$. Let us compare these on the simulated data in Figure 1. We generated in this simulation a sample from the Pareto distribution with $\theta = 2$. The sample average turns out to be $\bar{X}_n = 1.9669$, so that the methods of moments estimate is 2.0342. The maximum likelihood estimate is 1.9790. Thus, the maximum likelihood estimate is a little closer to the true θ than the methods of moments estimate.

Properties of Estimators. Let $T_n = T_n(X_1, \ldots, X_n)$ be an estimator of the real-valued parameter θ. Then it is desirable that T_n is in some sense close to θ. A minimum requirement is that the estimator approaches θ as the sample size increases. This is called *consistency*. To be more precise, suppose the sample $X_1, \ldots, X_n$ are the first n of an infinite sequence $X_1, X_2, \ldots$ of independent copies of X. Then T_n is called consistent if (with probability one)

$$T_n \to \theta \text{ as } n \to \infty. \tag{11}$$

Note that consistency of frequencies as estimators of probabilities, or means as estimators of expectations, follows from the (strong) **law of large numbers**.

The *bias* of an estimator T_n of θ is defined as its mean deviation from θ:

$$\text{bias}(T_n) = E(T_n) - \theta. \tag{12}$$

We remark here that the distribution of $T_n = T_n(X_1, \ldots, X_n)$ depends on P, and, hence, on θ. Therefore, the expectation $E(T_n)$ depends on θ as well. We indicate this by writing $E(T_n) = E_\theta(T_n)$. The estimator T_n is called *unbiased* if

$$E_\theta(T_n) = \theta \text{ for all possible values of } \theta. \tag{13}$$

Example 4 Consider the estimators $S_n^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ and $\hat{\sigma}_n^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$. Note that S_n^2 is larger than $\hat{\sigma}_n^2$, but that the difference is small when n is large. It can be shown that S_n^2 is an unbiased estimator of the variance $\sigma^2 = \text{var}(X)$. The estimator $\hat{\sigma}_n^2$ is biased: it underestimates the variance.

In many models, unbiased estimators do not exist. Moreover, it often heavily depends on the model under consideration, whether or not an estimator is unbiased. A weaker concept is *asymptotic unbiasedness* (see [1]).

The mean square error of T_n as estimator of θ is

$$MSE(T_n) = E(T_n - \theta)^2. \tag{14}$$

One may decompose the MSE as

$$MSE(T_n) = \text{bias}^2(T_n) + \text{var}(T_n), \tag{15}$$

where $\text{var}(T_n)$ is the variance of T_n.

Bias, variance, and mean square error are often quite hard to compute, because they depend on the distribution of all n observations $X_1, \ldots, X_n$. However, one may use certain approximations for large sample sizes n. Under regularity conditions, the maximum likelihood estimator $\hat{\theta}_n$ of θ is asymptotically unbiased, with asymptotic variance $1/(nI(\theta))$, where $I(\theta)$ is the Fisher information in a single observation (*see* **Information Matrix**). Thus, maximum likelihood estimators reach the minimum variance bound asymptotically.

Histograms. Our next aim is estimating the density $f(x)$ at a given point x. The density is defined as the derivative of the distribution function F at x:

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \to 0} \frac{P(x, x+h]}{h}. \tag{16}$$

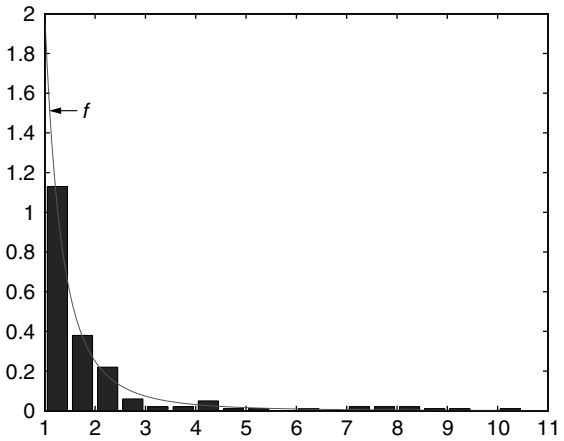

Figure 2 Histogram with bandwidth $h = 0.5$ and true density

Here, $(x, x+h]$ is the interval with left endpoint x (not included) and right endpoint $x+h$ (included). Unfortunately, replacing P by $\hat{P}_n$ here does not work, as for h small enough, $\hat{P}_n(x, x+h]$ will be equal to zero. Therefore, instead of taking the limit as $h \to 0$, we *fix* h at a (small) positive value, called the bandwidth. The estimator of $f(x)$, thus, becomes

$$\hat{f}_n(x) = \frac{\hat{P}_n(x, x+h]}{h} = \frac{\text{number of } X_i \in (x, x+h]}{nh}. \tag{17}$$

A plot of this estimator at points $x \in \{x_0, x_0 + h, x_0 + 2h, \ldots\}$ is called a **histogram**.

Example 3 continued Figure 2 shows the histogram, with bandwidth $h = 0.5$, for the sample of size $n = 200$ from the Pareto distribution with parameter $\theta = 2$. The solid line is the density of this distribution.

Minimum Chi-square. Of course, for real (not simulated) data, the underlying distribution/density is not known. Let us explain in an example a procedure for checking whether certain model assumptions are reasonable. Suppose that one wants to test whether data come from the exponential distribution with parameter θ equal to 1. We draw a histogram of the sample (sample size $n = 200$), with bandwidth $h = 1$ and 10 cells (see Figure 3). The cell counts are (151, 28, 4, 6, 1, 1, 4, 3, 1, 1). Thus, for example, the number of observations that falls in

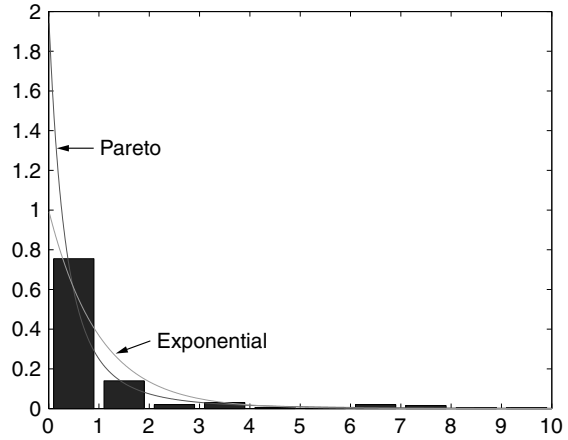

Figure 3 Histogram with bandwidth $h = 1$, exponential and Pareto density

the first cell, that is, that have values between 0 and 1, is equal to 151. The cell probabilities are, therefore, (0.755, 0.140, 0.020, 0.030, 0.005, 0.005, 0.020, 0.015, 0.005, 0.005). Now, according to the exponential distribution, the probability that an observation falls in cell k is equal to $\mathrm{e}^{-(k-1)} - \mathrm{e}^{-k}$, for $k = 1, \ldots, 10$. These cell probabilities are (.6621, .2325, .0855, .0315, .0116, .0043, .0016, .0006, .0002, .0001). Because the probabilities of the last four cells are very small, we merge them together. This gives cell counts $(N_1, \ldots, N_7) = (151, 28, 4, 6, 1, 1, 4, 9)$ and cell probabilities $(\pi_1, \ldots, \pi_7) = (.6621, .2325, .0855, .0315, .0116, .0043, .0025)$. To check whether the cell frequencies differ significantly from the hypothesized cell probabilities, we calculate Pearson's χ^2. It is defined as

$$\chi^2 = \frac{(N_1 - n\pi_1)^2}{n\pi_1} + \cdots + \frac{(N_7 - n\pi_7)^2}{n\pi_7}. \quad (18)$$

We write this as $\chi^2 = \chi^2(\text{exponential})$ to stress that the cell probabilities were calculated assuming the exponential distribution. Now, if the data are exponentially distributed, the χ^2 statistic is generally not too large. But what is large? Consulting a table of Pearson's χ^2 at the 5% significance level gives the critical value $c = 12.59$. Here we use 6 degrees of freedom. This is because there are $m = 7$ cell probabilities, and there is the restriction $\pi_1 + \cdots + \pi_m = 1$, so we estimated $m - 1 = 6$ parameters. After some calculations, one obtains $\chi^2(\text{exponential}) = 168.86$. This exceeds the critical value c, that is, $\chi^2(\text{exponential})$ is too large to support the assumption of the exponential distribution. In fact, the data considered here are the simulated sample from the Pareto distribution with parameter $\theta = 2$. We shifted this sample one unit to the left. The value of χ^2 for this (shifted) Pareto distribution is

$$\chi^2(\text{Pareto}) = 10.81. \quad (19)$$

This is below the critical value c, so that the test, indeed, does not reject the Pareto distribution. However, this comparison is not completely fair, as our decision to merge the last four cells was based on the exponential distribution, which has much lighter tails than the Pareto distribution.

In Figure 3, the histogram is shown, together with the densities of the exponential and Pareto distribution. Indeed, the Pareto distribution fits the data better in the sense that it puts more mass at small values.

Continuing with the test for the exponential distribution, we note that, in many situations, the intensity θ is not required to be fixed beforehand. One may use an estimator for θ and proceed as before, calculating χ^2 with the estimated value for θ. However, the critical values of the test then become smaller. This is because, clearly, estimating parameters using the sample means that the hypothesized distribution is pulled towards the sample. Moreover, when using, for example, maximum likelihood estimators of the parameters, critical values will in fact be hard to compute. The minimum χ^2 estimator overcomes this problem. Let $\pi_k(\vartheta)$ denote the cell probabilities when the parameter value is ϑ, that is, in the exponential case $\pi_k(\vartheta) = \mathrm{e}^{-\vartheta(k-1)} - \mathrm{e}^{-\vartheta k}$, $k = 1, \ldots, m-1$, and $\pi_m(\vartheta) = 1 - \sum_{k=1}^{m-1} \pi_k(\vartheta)$. The minimum χ^2 estimator $\hat{\theta}_n$ is now the minimizer over ϑ of

$$\left\{ \frac{(N_1 - n\pi_1(\vartheta))^2}{n\pi_1(\vartheta)} + \cdots + \frac{(N_m - n\pi_m(\vartheta))^2}{n\pi_m(\vartheta)} \right\}. \quad (20)$$

The χ^2 test with this estimator for θ now has $m - 2$ degrees of freedom. More generally, the number of degrees of freedom is $m - 1 - d$, where d is the number of estimated free parameters when calculating cell probabilities. The critical values of the test can be found in a χ^2 table.

Sufficiency. A goal of statistical analysis is generally to summarize the (large) data set into a small

number of characteristics. The sample mean and sample variance are such summarizing statistics, but so is, for example, the sample median, and so on. The question arises, to what extent one can summarize data without throwing away information. For example, suppose you are given the empirical distribution function $\hat{F}_n$, and you are asked to reconstruct the original data $X_1, \ldots, X_n$. This is not possible since the ordering of the data is lost. However, the index i of X_i is just a label: it contains no information about the distribution P of X_i (assuming that each observation X_i comes from the same distribution, and the observations are independent). We say that the empirical distribution $\hat{F}_n$ is *sufficient*. More generally, a statistic $T_n = T_n(X_1, \ldots, X_n)$ is called *sufficient* for P if the distribution of the data given the value of T_n does not depend on P. For example, it can be shown that when P is the exponential distribution with unknown intensity, then the sample mean is sufficient. When P is the normal distribution with unknown mean and variance, then the sample mean and sample variance are sufficient. Cell counts are not sufficient when, for example, P is a continuous distribution. This is because, if one only considers the cell counts, one throws away information on the distribution within a cell. Indeed, when one compares Figures 2 and 3 (recall that in Figure 3 we shifted the sample one unit to the left), one sees that, by using just 10 cells instead of 20, the strong decrease in the second half of the first cell is no longer visible.

Sufficiency depends very heavily on the model for P. Clearly, when one decides to ignore information because of a sufficiency argument, one may be ignoring evidence that the model's assumptions may be wrong. Sufficiency arguments should be treated with caution.

References

[1] Bickel, P.J. & Doksum, K.A. (2001). *Mathematical Statistics*, 2nd Edition, Holden-Day, San Francisco.

[2] Pareto, V. (1897). *Course d'Économie Politique*, Rouge, Lausanne et Paris.

SARA A. VAN DE GEER