How to Estimate Statistical Characteristics Based on a Sample: Nonparametric Maximum Likelihood Approach Leads to Sample Mean, Sample Variance, etc.

Vladik Kreinovich and Thongchai Dumrongpokaphan

Abstract In many practical situations, we need to estimate different statistical characteristics based on a sample. In some cases, we know that the corresponding probability distribution belongs to a known finite-parametric family of distributions. In such cases, a reasonable idea is to use the Maximum Likelihood method to estimate the corresponding parameters, and then to compute the value of the desired statistical characteristic for the distribution with these parameters.
In some practical situations, we do not know any family containing the unknown distribution. We show that in such nonparametric cases, the Maximum Likelihood approach leads to the use of sample mean, sample variance, etc.

1 Need to Estimate Statistical Characteristics Based on a Sample: Formulation of the Problem

Need to estimate statistical characteristics. In many practical situations, we need to estimate statistical characteristic of a certain random phenomenon based on a given sample.

For example, to check that for all the mass-produced gadgets from a given batch, the valued of the corresponding physical quantity are within the desired bounds, the ideal solution would be to measure the quantity for all the gadgets. This may be reasonable to do if these gadgets are intended for a spaceship, where a minor fault can lead to catastrophic results. However, in most applications, it is possible

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

Thongchai Dumrongpokaphan
Department of Mathematics, Faculty of Science, Chiang Mai University, Thailand, e-mail: tcd43@hotmail.com

to save time and money by testing only a small sample, and by making statistical conclusions based on the results of this testing.

How do we estimate the statistical characteristics – finite-parametric case: main idea. In many situations, we know that the actual distribution belongs to a known finite-parametric family of distributions. For example, it is often known that the distribution is Gaussian (normal), for some (unknown) values of the mean μ and standard deviation σ. In general, we know that the corresponding probability density function (pdf) has the form $f(x\,|\,\theta)$ for some parameters $\theta = (\theta_1, \ldots, \theta_n)$.

In such situations, we first estimate the values of the parameters θ_i based on the sample, and then compute the values of the corresponding statistical characteristic (mean, standard deviation, kurtosis, etc.) corresponding to the estimates values θ_i.

How do we estimate the statistical characteristics – finite-parametric case: details. How do we estimate the values of the parameters θ_i based on the sample? A natural idea is to select the *most probable* values θ. How do we go from this idea to an algorithm?

To answer this question, let us first note that while theoretically, each of the parameters θ_i can take infinitely many values, in reality, for a given sample size, it is impossible to detect the difference between the nearby values θ_i and θ_i'. Thus, from the practical viewpoint, we have finitely many distinguishable cases.

In this description, we have finitely many possible combinations of parameters $\theta^{(1)}, \ldots, \theta^{(N)}$. We consider the case when all we know is that the actual pdf belongs to the family $f(x\,|\,\theta)$. There is no a priori reason to consider some of the possible values $\theta^{(k)}$ as more probable. Thus, before we start our observations, it is reasonable to consider these N hypotheses as equally probable: $P_0(\theta^{(k)}) = \dfrac{1}{N}$. This reasonable idea is known as the *Laplace Indeterminacy Principle*; see, e.g., [1].

We can now use the Bayes theorem to compute the probabilities $P(\theta^{(k)}\,|\,x)$ of different hypotheses $\theta^{(k)}$ after we have performed the observations, and these observations resulted in a sample $x = (x_1, \ldots, x_n)$:

$$P(\theta^{(k)}\,|\,x) = \frac{P(x\,|\,\theta^{(k)}) \cdot P_0(\theta^{(k)})}{\sum\limits_{i=1}^{N} P(x\,|\,\theta^{(i)}) \cdot P_0(\theta^{(i)})}.$$

Here, the probability $P(x\,|\,\theta^{(k)})$ is proportional to $f(x\,|\,\theta^{(k)})$. Dividing both numerator and denominator by $P_0 = \dfrac{1}{N}$, we thus conclude that

$$P(\theta^{(k)}\,|\,x) = c \cdot f(x\,|\,\theta^{(k)})$$

for some constant c.

Thus, selecting the most probable hypotheses $P(\theta^{(k)}\,|\,x) \to \max\limits_{k}$ is equivalent to finding the values θ for which, for the given sample x, the expression $f(x\,|\,\theta)$ attains its largest possible value. The expression $f(x\,|\,\theta)$ is known as *likelihood*, and the whole idea is known as the *Maximum Likelihood Method*; see, e.g., [2].

In particular, for Gaussian distribution, the Maximum Likelihood method leads to the sample mean

$$\widehat{\mu} \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

as the estimate for the mean, and to the sample variance

$$(\widehat{\sigma})^2 \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

as the estimate for the variance.

What if we do not know the family? In some practical situations, we do not know a finite-parametric family of distributions that contains the actual one. In such situations, all we know is a sample. Based on this sample, how can we estimate the statistical characteristics of the corresponding distribution?

What we do in this paper. In this paper, we apply the Maximum Likelihood method to the above problem. It turns out that the resulting estimates are sample mean, sample variance, etc.

Thus, we get a justification for using these estimates beyond the case of the Gaussian distribution.

2 Nonparametric Maximum Likelihood Approach to Estimating Statistical Characteristics Based on a Sample: Continuous Case

Description of the case. Let us first consider the case when the random variable is continuous, i.e., when, in principle, it can take either all possible real values – or at least all possible real values from a certain interval.

Possibility to discretize. Similar to the above case, while theoretically, we can thus have infinitely many possible values of the random variable x, in reality, due to measurement uncertainty, very close values $x \approx x'$ are indistinguishable. Thus, in practice, we can safely assume that there are only finitely many distinguishable values $x^{(1)} < x^{(2)} < \ldots < x^{(M)}$.

Possible probability distributions. In the discretized representations, to describe the corresponding random variable, we need to describe the probabilities $p_i = p(x^{(i)})$ of each of M values $x^{(i)}$, $1 \le i \le M$. The only restriction on these probabilities is that they should be non-negative and add up to 1: $\sum\limits_{i=1}^{M} p_i = 1$.

Let us apply the Maximum Likelihood Method: resulting formulation. According to the Maximum Likelihood Method, out of all possible probability distributions $\mathbf{p} = (p_1, \ldots, p_n)$, we should select a one for which the probability of observing a given sequence $x_1, \ldots, x_n$ is the largest.

The probability of observing each value x_i is equal to $p(x_i)$. It is usually assumed that different elements in the sample are independent, so the probability $p(x\,|\,\mathbf{p})$ of observing the whole sample $x = (x_1, \ldots, x_n)$ is equal to the product of these probabilities:

$$p(x\,|\,\mathbf{p}) = \prod_{i=1}^{n} p(x_i).$$

In the continuous case, the probability of observing the exact same number twice is zero, so we can safely assume that all the values x_i are different. In this case, the above product takes the form

$$p(x\,|\,\mathbf{p}) = \prod \{x_i : x_i \text{ has been observed}\}.$$

We need to find the values $p_1, \ldots, p_M$ that maximize this probability under the constraints that $p_i \geq 0$ and $\sum\limits_{i=1}^{M} p_i = 1$.

Analysis of the problem. Let us explicitly describe the probability distribution that maximizes the corresponding likelihood.

First, let us notice that when the maximum is attained, the values p_i corresponding to un-observed values should be 0. Indeed, if $p_i > 0$ for one of the indices i corresponding to an un-observed value x_i, then we can, without changing the constraint $\sum\limits_{i=1}^{M} p_i = 1$, decrease this value to 0 and instead increase one of the probabilities p_i corresponding to an observed value x_i.

Let I denote the set of all indices corresponding to observed values p_i. Then, in the optimal arrangement, we have $p_i = 0$ for $i \notin I$. So, the constraint $\sum\limits_{i=1}^{M} p_i = 1$ takes the form $\sum\limits_{i \in I} p_i = 1$, and the likelihood optimization problem takes the following form: $\prod\limits_{i \in I} p_i \to \max$ under the constraint that $\sum\limits_{i \in I} p_i = 1$.

This is a known and easy-to-solve optimization problem. The corresponding maximum is attained when all the probabilities p_i are equal to each other, i.e., when $p_i = \dfrac{1}{n}$. Thus, we arrive at the following conclusion.

Conclusion: we should use sample mean, sample variance, etc. In the non-parametric case, the maximum likelihood method implies that out of all possible probability distributions, we should select a distribution in which all sample values $x_1, \ldots, x_n$ appear with equal probability $p_i = \dfrac{1}{n}$, and no other values can appear.

So, as estimates of the desired statistical characteristics, we should select characteristics corresponding to this sample-based distribution. The mean of this distribution is equal to $\widehat{\mu} = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i$, i.e., to the sample mean. The variance of this distribution is equal to $\dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} (x_i - \widehat{\mu})^2$, i.e., to the sample variance.

Thus, for the nonparametric case, the maximum likelihood method implies that we should use sample mean, sample variance, etc.

Discussion. Thus, we get a justification for using sample mean, sample variance, etc., in situations beyond their usual Gaussian-based justification.

3 Nonparametric Maximum Likelihood Approach to Estimating Statistical Characteristics Based on a Sample: Discrete Case

Description of the case. In the discrete case, we have a finite list of possible values $x^{(1)}, \ldots, x^{(M)}$. To describe a probability distribution, we need to describe the probabilities $p_i = p(x^{(i)})$ of these values.

Maximum Likelihood Approach: formulation of the optimization problem. For each sample $x_1, \ldots, x_n$, the corresponding likelihood $\prod\limits_{i=1}^{n} p(x_i)$ takes the form

$$p(x \,|\, \mathbf{p}) = \prod_{i=1}^{M} p_i^{n_i},$$

where n_i is the number of times the value $x^{(i)}$ appears in the sample.

We must find the probabilities p_i for which the likelihood attains its largest possible value under the constraint $\sum\limits_{i=1}^{n} p_i = 1$.

Optimizing the likelihood. To solve the above constraint optimization problem, we can use the Lagrange multiplier method that reduces it to the unconstrained optimization problem

$$\prod_{i=1}^{M} p_i^{n_i} + \lambda \cdot \left(\sum_{i=1}^{M} p_i - 1 \right) \to \max_{p}.$$

Differentiating this objective function with respect to p_i, taking into account that for $A \stackrel{\text{def}}{=} \prod\limits_{i=1}^{M} p_i^{n_i}$, we get

$$\frac{\partial A}{\partial p_i} = \prod_{j \neq i} p_j^{n_j} \cdot n_i \cdot p_i^{n_i - 1} = A \cdot \frac{n_i}{p_i},$$

and equating the derivative to 0, we conclude that

$$A \cdot \frac{n_i}{p_i} + \lambda = 0.$$

Thus, $p_i = \text{const} \cdot n_i$. The constraint that $\sum_{i=1}^{M} p_i = 1$ implies that the constant is equal to 1 over the sum $\sum_{i=1}^{n} n_i = n$. Thus, we get $p_i = \frac{n_i}{n}$. So, we arrive at the following conclusion.

Conclusion. In the discrete case, for each of the possible values $x^{(i)}$, we assign, as the probability p_i, the frequency $\frac{n_i}{n}$ with which this value appears in the observed sample.

This is the probability distribution that we should use to estimate different statistical characteristics. For this distribution, the mean is still equal to the sample mean, and the variance is still equal to the sample variance – same as for the continuous case.

However, e.g., for entropy, we get a value which is different from the continuous case: there, the entropy is always equal to

$$-\sum_{i\in I} p_i \cdot \ln(p_i) = -n \cdot \frac{1}{n} \cdot \ln\left(\frac{1}{n}\right) = \ln(n),$$

while in the discrete case, we have a different value

$$-\sum_{i\in I} p_i \cdot \ln(p_i) = -\sum_{i=1}^{M} \frac{n_i}{n} \cdot \ln\left(\frac{n_i}{n}\right).$$

Acknowlegments

This work was supported by Chiang Mai University, Thailand. This work was also supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

References

1. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
2. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.