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

Vladik Kreinovich[1] and Thongchai Dumrongpokaphan[2]

[1]University of Texas at El Paso, El Paso, Texas 79968, USA
vladik@utep.edu
[2]Department of Mathematics, Chiang Mai University, Thailand
tcd43@hotmail.com

1. Need to Estimate Statistical Characteristics

- In many practical situations, we need to estimate statistical characteristic based on a given sample.
- For example, we need to check that:
  - for all the mass-produced gadgets from a given batch,
  - the values 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 for a spaceship, where a minor fault can lead to catastrophic results.
- Usually, we can save time and money:
  - by testing only a small sample, and
  - making statistical conclusions from the results.

2. 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:

$$f(x\,|\,\theta) \text{ for some } \theta = (\theta_1, \ldots, \theta_n).$$

- For example, the distribution is Gaussian (normal), for some (unknown) mean μ and st. dev. σ.
- In such situations:
  - we first estimate the values of the parameters θ_i based on the sample, and then
  - we compute statistical characteristic (mean, standard deviation, etc.) corr. to the estimates θ_i.

3. 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.

4. Finite-Parametric Case (cont-d)

- 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 $f(x \mid \theta)$.
- 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:

$$P_0(\theta^{(k)}) = \frac{1}{N}.$$

- This reasonable idea is known as the *Laplace Indeterminacy Principle.*

5. Finite-Parametric Case (cont-d)

- 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 $x = (x_1, \ldots, x_n)$:

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

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

$$P(\theta^{(k)}\,|\,x) = c \cdot f(x\,|\,\theta^{(k)}) \text{ for some constant } c.$$

6. Finite-Parametric Case (cont-d)

- 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 $f(x\,|\,\theta)$ is the largest possible.
- The expression $f(x\,|\,\theta)$ is known as *likelihood*.
- The whole idea is thus known as the *Maximum Likelihood Method*.
- In particular, for Gaussian distribution, the Maximum Likelihood method leads:
  - to the sample mean $\widehat{\mu} \stackrel{\text{def}}{=} \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i$, and
  - to the sample variance $(\widehat{\sigma})^2 \stackrel{\text{def}}{=} \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} (x_i - \widehat{\mu})^2$.

7. What If We Do Not Know the Family?

- Often, we do not know a finite-parametric family of distributions containing 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?
- 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.

8. Continuous Case

- Let us first consider the case when the random variable is continuous.
- Theoretically, we can thus have infinitely many possible values of the random variable x.
- Ii 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

$$x^{(1)} < x^{(2)} < \ldots < x^{(M)}.$$

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

9. Let Us Apply the Maximum Likelihood Method: Resulting Formulation

- According to the Maximum Likelihood Method,
  - out of all possible probability distributions $\vec{p} = (p_1, \ldots, p_n)$,
  - 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 x_i is $p(x_i)$.
- It is usually assumed that different elements in the sample are independent.
- So, the probability $p(x \,|\, \vec{p})$ of observing the whole sample $x = (x_1, \ldots, x_n)$ is equal to the product:

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

10. Continuous Case (cont-d)

- 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

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

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

11. 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.

12. Analysis of the Problem (cont-d)

- Let I denote the set of all indices corresponding to observed values p_i.
- Then, in the optimal arrangement, $p_i = 0$ for $i \notin I$.
- So, $\sum\limits_{i=1}^{M} p_i = 1$ takes the form $\sum\limits_{i \in I} p_i = 1$.
- The likelihood optimization problem takes the following form: $\prod\limits_{i \in I} p_i \to \max$ under the constraint $\sum\limits_{i \in I} p_i = 1$.
- This is a known optimization problem.
- The corresponding maximum is attained when all the probabilities p_i are equal to each other: $p_i = \dfrac{1}{n}$.
- Thus, we arrive at the following conclusion.

13. 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 select a distribution in which all sample values $x_1, \ldots, x_n$ appear with equal probability $p_i = \dfrac{1}{n}$.
- 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.

14. Conclusion (cont-d)

- 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, the maximum likelihood method implies that we should use sample mean, sample variance, etc.
- So, we justify using sample mean, sample variance, etc., in situations beyond Gaussian case.

15. Discrete 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.
- For each sample $x_1, \ldots, x_n$, the corresponding likelihood $\prod_{i=1}^{n} p(x_i)$ takes the form $p(x \mid \vec{p}) = \prod_{i=1}^{M} p_i^{n_i}$.
- Here, n_i is the number of times the value $x^{(i)}$ appears in the sample.
- We must find p_i for which the likelihood is the largest under the constraint $\sum_{i=1}^{n} p_i = 1$.

16. Optimizing the Likelihood

- To solve the above constraint optimization problem, we can use the Lagrange multiplier method.
- This method reduces our problem 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 $A \stackrel{\text{def}}{=} \prod_{i=1}^{M} p_i^{n_i}$, we get
$$\frac{\partial A}{\partial p_i} = \prod_{j \ne i} p_j^{n_j} \cdot n_i \cdot p_i^{n_i - 1} = A \cdot \frac{n_i}{p_i}.$$
- Equating the derivative to 0, we conclude that
$$A \cdot \frac{n_i}{p_i} + \lambda = 0.$$

17. Optimizing the Discrete-Case Likelihood (cont-d)

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

18. Discrete Case: Conclusion

- In the discrete case, for each of the possible values $x^{(i)}$, we assign, as the probability p_i, the frequency $\dfrac{n_i}{n}$.
- 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.

19. Discrete Case: Conclusion (cont-d)

- In the continuous case, $p_i = \frac{1}{n}$.
- Thus, in the continuous case, the entropy is always equal to

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

- In the discrete case, we have a different value

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

20. Acknowledgments

- We acknowledge the support of Chiang Mai University, Thailand.
- This work was also supported in part:
  - by the National Science Foundation grants HRD-0734825, HRD-1242122, and DUE-0926721, and
  - by an award from Prudential Foundation.