Sufficient Statistics and Estimation

Marianna Bolla, Prof, DSc.
Institute of Mathematics, BME

August 31, 2023

Descriptive statistics

$(\mathcal{S}, \mathcal{A}, \mathcal{P})$ is a *statistical space* if $(\mathcal{S}, \mathcal{A}, \mathbb{P})$ is probability space for all $\mathbb{P} \in \mathcal{P}$, where $\mathcal{P}$ is a family of distributions.
Parametric case: $\mathcal{P} = \{\mathbb{P}_\theta \,|\, \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$ is the *parameter space*.
Statistical sample: $X_1, X_2, \dots, X_n$ i.i.d.
Sample space ($\mathcal{X}$): set of all possible *realizations* $\mathbf{x} = (x_1, \dots, x_n)$ of $\mathbf{X} = (X_1, \dots, X_n)$.
Statistic: $T = T(\mathbf{X}) = T(X_1, \dots, X_n)$ measurable function of the sample elements.
Basic **descriptive statistics**:

- *Sample mean*: $\bar{X} = \dfrac{1}{n} \sum_{i=1}^n X_i$. (Sometimes $\bar{X}_n$, $\bar{x}$, $\bar{x}_n$.)
- *Steiner's Theorem*: $\dfrac{1}{n} \sum_{i=1}^n (x_i - c)^2 = \dfrac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + (\bar{x} - c)^2$.
- *Empirical variance*: $S^2 = \dfrac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \overline{X^2} - \bar{X}^2$.
- *Corrected empirical variance*: $S^{*2} = \dfrac{n}{n-1} S^2 = \dfrac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.
- *Standard Error of Mean*: $\bar{X}\sqrt{n}/S^*$.
- *k-th empirical moment*: $M_k = \dfrac{1}{n} \sum_{i=1}^n X_i^k$. *Centered* version: $M_k^c = \dfrac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$. ($S^2 = M_2^c = M_2 - M_1^2$.)
- *Skewness*: $M_3^c/(M_2^c)^{3/2}$. *Kurtosis*: $M_4^c/(M_2^c)^2 - 3$.
- *Empirical covariance* based on $(X_1, Y_1)^T, \dots, (X_n, Y_n)^T$ i.i.d.:

$$C = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}.$$

- *Empirical correlation coefficient*: $R = \frac{C}{S_X S_Y} = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^n Y_i^2 - n\bar{Y}^2\right)}}$.

- **Order statistics**: $X_1^* \leq X_2^* \leq \cdots \leq X_n^*$ (neither independent, nor identically distributed).
  - *Sample range*: $X_n^* - X_1^*$.
  - *Empirical median*: X_{k+1}^* (if $n = 2k+1$), and $(X_k^* + X_{k+1}^*)/2$ (if $n = 2k$).
  - *Proposition* (Steiner in L_1-norm): $\min_{c\in\mathbb{R}} \dfrac{1}{n}\sum_{i=1}^n |x_i - c| = \dfrac{1}{n}\sum_{i=1}^n |x_i - m|$.
  - *Empirical c.d.f.*: $F_n^*(x) := \frac{\sum_{i=1}^n I(X_i<x)}{n}$ (stochastic process, x is the time).
  - **Glivenko–Cantelli Theorem** (fundamental theorem of statistics): $\sup_{x\in\mathbb{R}} |F_n^*(x) - F(x)| \to 0$, almost surely ($n\to\infty$).

Sufficient statistics

We take an i.i.d. sample $X_1, \dots, X_n$ from a population with distribution $\mathbb{P}_\theta$, where θ is unknown parameter, and it is in the *parameter space* Θ, so $\theta \in \Theta$. For example, if $\mathbf{X} := (X_1, \dots, X_n)$ follow Poisson distribution, then the parameter, now denoted by λ is in the parameter space $\Theta = (0, \infty)$. The *sample space* is the set of all possible n-tuples $(x_1, \dots, x_n)$ that are possible *realizations* of the sample. For fixed *simple size* n, let $\mathcal{X} \subset \mathbb{R}^n$ denote the *sample space*, that is the set of all possible realizations. In the Poisson case, it is $\mathcal{X} = \{0, 1, 2, \dots\}^n$.

Point estimation means that we want to conclude for θ based on a sample. For this, we need a statistic that contains all important information from the sample.

Definition 1 *The likelihood function for* $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}$ *and* $\theta \in \Theta$ *is* $L_\theta(\mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i) = \prod_{i=1}^n p_\theta(x_i)$ *in the discrete, and* $L_\theta(\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i)$ *in the absolutely continuous case, where* $p_\theta(x)$ *is the probability mass function (p.m.f.) in the discrete, and* $f_\theta(x)$ *is the probability density function (p.d.f.) in the continuous case.*

Now we organize the sample entries into a *statistic* $T := T(X_1, \dots, X_n) = T(\mathbf{X})$ such that, by this compression, we would not loose any information for the parameter.

Definition 2 *The statistic* $T(\mathbf{X})$ *is sufficient for* θ *if the distribution of* $\mathbf{X}$ *conditioned on* $T(\mathbf{X})$ *does not depend on* θ.

It means that T contains all the information that can be retrieved from the sample for the parameter.

Theorem 1 (Neyman–Fisher factorization) *The statistic* $T(\mathbf{X})$ *is sufficient for* θ *if and only if the likelihood function can be factorized as*

$$L_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x}), \quad \forall \theta \in \Theta, \quad \mathbf{x} \in \mathcal{X}$$

with some measurable, nonnegative real functions g *and* h.

Sufficient statistics are many, even based on the same sample and for the same parameter (e.g., the ordered sample is such). A sufficient statistic is *minimal* if it is the function of any other sufficient statistic. Minimal sufficient statistic always exists, and it is unique up to equivalence.

Theory of point estimation

We want to estimate θ, or its measurable function $\psi(\theta)$ by means of the statistic $T(\mathbf{X})$ on the basis of the i.i.d. sample $\mathbf{X} = (X_1, \ldots, X_n)$. The point estimator is sometimes denoted by $\hat{\theta}$ or $\hat{\psi}$. Criteria for the 'goodness' of a point estimator:

- $T(\mathbf{X})$ is an **unbiased** estimator of $\psi(\theta)$, if $\mathbb{E}_\theta(T(\mathbf{X})) = \psi(\theta), \quad \forall\theta \in \Theta$.
- $T(\mathbf{X}_n)$ is an **asymptotically unbiased** estimator of $\psi(\theta)$, if

$$\lim_{n\to\infty} \mathbb{E}_\theta(T(\mathbf{X}_n)) = \psi(\theta), \quad \forall\theta \in \Theta.$$

- Let T_1 and T_2 be both unbiased estimators of $\psi(\theta)$. T_1 is **at least as efficient** than T_2, if $\mathrm{Var}\,_\theta^2(T_1) \leq \mathrm{Var}\,_\theta^2(T_2)$, $\forall\theta \in \Theta$. An unbiased estimator is **efficient**, if it is at least as efficient than any other unbiased estimator. An efficient estimator is sometimes called *minimum variance unbiased estimator.*

  Efficient estimator does not always exist, but if yes, then it is unique (with probability 1).
- $T(\mathbf{X}_n)$ is a weakly/strongly/mean square **consistent** estimator of $\psi(\theta)$, if $\forall\theta \in \Theta$:
  $T(\mathbf{X}_n) \to \psi(\theta)$ in probability/almost surely/mean square as $n \to \infty$.

Examples of 'good' estimators:

- the sample mean $\bar{X}$ is always an unbiased estimator of the population mean $\mathbb{E}(X_1)$;
- the empirical variance is asymptotically unbiased, whereas, the corrected empirical variance is unbiased estimator of the population variance $\mathrm{Var}\,(X_1)$;
- the above are also consistent in all of the three meanings (provided the first/second/fourth population moments exist).

Methods of point estimation:

- **Maximum Likelihood Estimation (MLE)**: given the sample, the MLE of θ is $\hat{\theta}$ if it maximizes the likelihood function. By common sense, in case of a discrete distribution, the MLE is a possible parameter value, for which having the actual sample is the most likely. However, $\hat{\theta} = T(\mathbf{X})$ is a statistic, and it is asymptotically unbiased and strongly consistent estimator of θ.
- **Method of moments**: if $\dim(\theta) = k$, then we find the first k moments of the $\mathbb{P}_{(\theta_1,\ldots,\theta_k)}$ distribution. If, vice versa, θ_j can be expressed by the first k moments, then the same function of the empirical moments gives $\hat{\theta}_j$, for $j = 1, \ldots, k$.

Examples

1. Let $X_1, \dots, X_n$ be i.i.d. sample from Poisson distribution with parameter λ.

$$L_\lambda(\mathbf{x}) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \left(\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}\right) \cdot \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) = g_\lambda(\sum_{i=1}^{n} x_i) \cdot h(\mathbf{x}),$$

so $\sum_{i=1}^n X_i$ is sufficient statistic for λ, akin to its one-to-one function $\bar{X}$.

To find the MLE,

$$\ln L_\lambda(\mathbf{x}) = \ln \left[\prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right] = \ln \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln x_i! - \lambda n.$$

Differentiating with respect to λ, the likelihood equation is

$$\frac{\partial \ln L_\lambda(\mathbf{x})}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0.$$

The solution is $\hat{\lambda} = \bar{x}$, which indeed gives a local and global maximum. So $T(\mathbf{X}) = \bar{X}$ is the MLE of λ, provided it is not 0, i.e., not all the sample entries are zero at the same time (it can happen with positive, albeit 'small' probability).

2. Let $X_1, \dots, X_n$ be i.i.d. sample from exponential distribution with parameter λ). Then

$$L_\lambda(\mathbf{x}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i},$$

that is $g_\lambda(T(\mathbf{x}))$, and $h(\mathbf{x}) = 1 \cdot I_{(0,\infty)}$. Therefore, $\sum_{i=1}^n X_i$ is sufficient akin to $\bar{X}$ or $\frac{1}{\bar{X}}$.

As for the MLE of λ,

$$\ln L_\lambda(\mathbf{x}) = \ln \left[\prod_{i=1}^{n} \lambda e^{-\lambda x_i}\right] = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i,$$

from which, after differentiating, we get that $\hat{\lambda} = 1/\bar{x}$, that gives a local and global maximum. Consequently, $T(\mathbf{X}) = 1/\bar{X}$ is the MLE of λ with probability 1 ($\bar{X}$ can be 0 only with probability 0).

3. Let $X_1, \dots, X_n$ be i.i.d. sample from normal (Gaussian) distribution with unknown parameter $\theta = (\mu, \sigma^2)$. Then

$$L_\theta(\mathbf{x}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right) =$$

$$= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right]\right).$$

It is $g_\theta(T(\mathbf{x}))$, where $T(\mathbf{X}) = (\bar{X}, S^2)$ sufficient for θ, and $h(\mathbf{x}) = 1$. Obviously, $(\bar{X}, S^{*2})$ or $(\sum_{i=1}^n X_i,\ \sum_{i=1}^n X_i^2)$ are also sufficient.

To find MLE,

$$\ln L_\theta(\mathbf{x}) = \ln \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \sum_{i=1}^n \left[-\ln(\sqrt{2\pi\sigma^2}) - \frac{(x_i-\mu)^2}{2\sigma^2} \right] =$$

$$= -\frac{n}{2}(\ln(2\pi) + \ln \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Taking partial derivatives,

$$\frac{\partial \ln L_\theta(\mathbf{x})}{\partial \mu} = -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(x_i - \mu)(-1) = 0 \Longrightarrow \hat{\mu} = \bar{x}.$$

and

$$\frac{\partial \ln L_\theta(\mathbf{x})}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (x_i - \mu)^2 = 0.$$

Since the solution $\hat{\mu} = \bar{x}$ does not depend on the actual value of σ^2 substituting it to the second equation, we get that $\widehat{\sigma^2} = S_n^2$, that is only asymptotically unbiased for σ^2. The Hessian at $(\bar{x}, s_n^2)$ is:

$$H = \begin{pmatrix} -\frac{n}{s_n^2} & 0 \\ 0 & -\frac{n}{2(s_n^2)^2} \end{pmatrix},$$

which is negative definite, so we indeed have a local and global maximum here.

4. Let $X_1, \dots, X_n$ be i.i. sample from continuous uniform distribution on $[a, b]$. Here $\theta = (a, b)$.

$$L_\theta(\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i) = \frac{1}{(b-a)^n}, \quad \text{if} \quad x_1, \dots, x_n \in [a, b],$$

and 0, otherwise. $L_\theta(\mathbf{x}) = (b-a)^{-n} I(x_1^* \geq a,\ x_n^* \leq b) = g_\theta(x_1^*, x_n^*)$ and $h(\mathbf{x}) = 1$. So the pair (X_1^*, X_n^*) is sufficient for (a, b). It also gives the MLE, as we maximize the likelihood on the constraint that $[a, b]$ should contain all the sample entries.

Here the moment estimate of the parameters is not the same as the MLE, in contrast to the first three examples.

Interval estimation: The random interval $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ is a *confidence interval* of level at least $1 - \varepsilon$ for $\psi(\theta)$, if $\mathbb{P}_\theta(T_1 < \psi(\theta) < T_2) \geq 1 - \varepsilon$ $(\forall \theta \in \Theta)$.

Note that in case of a continuous distribution, exactly $1 - \varepsilon$ level confidence interval can be attained. ε is usually 'small', e.g., 0.05 or 0.01, in which cases we speak about 95% or 99% confidence intervals.