POBABILITY A4, Lessons 10-11: Statistics, ML Estimation, and Confidence Intervals

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

November 19, 2024

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

Estimation

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 convenient statistic.

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* $T := T(X_1, \dots, X_n) = T(\mathbf{X})$.

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 $\mathbf{X} = (X_1, \dots, X_n)$. The point estimator is sometimes denoted by $\hat{\theta}$ or $\hat{\psi}$. Some 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

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

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 $\sigma^2 = \text{Var}\,(X_1)$; (this is a **BONUS** exercise).

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.

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 $\hat{\sigma^2} = S_n^2$, 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.

Definition: Let $\xi_1, \dots, \xi_n \sim \mathcal{N}(0,1)$ be i.i.d. rv's. Then the distribution of the rv $\xi = \sum_{i=1}^n \xi_i^2$ is called χ^2 (chi2) distribution with degrees of freedom (d.f.) n.

Definition: Let $\eta \sim \mathcal{N}(0,1)$ and $\xi \sim \chi^2(n)$ be independent rv's. Then the distribution of

$$t = \frac{\eta}{\sqrt{\xi/n}}$$

is called Student t-distribution with degrees of freedom (d.f.) n and denoted by $t(n)$ (Student=V. Gosset).

Lukács' Theorem. Let $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma)$ be i.i.d. rv's. Then

1. $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$;
2. $nS_n^2/\sigma^2 \sim \chi^2(n-1)$, or equivalently, $(n-1){S_n^*}^2/\sigma^2 \sim \chi^2(n-1)$;
3. $\bar{X}$ and S_n^2 are independent rv's, or equivalently, $\bar{X}$ and ${S_n^*}^2$ are independent rv's.

Consequences:

- Recall that in case of $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma_0)$ i.i.d. sample, where σ_0 is known, for any $0 < \alpha < 1$, the $1-\alpha$ level confidence interval for μ is

$$I_{1-\alpha} = \bar{X} \pm \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}, \tag{1}$$

  where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile value of the standard normal distribution.

- In case of $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma)$ i.i.d. sample, where σ is unknown, by Lukacs' Theorem,

$$t = \frac{\frac{\bar{X}-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{(n-1){S_n^*}^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S_n^*}\sqrt{n} \sim t(n-1),$$

  therefore, for any $0 < \alpha < 1$, the $1-\alpha$ level confidence interval for μ is

$$I_{1-\alpha} = \bar{X} \pm \frac{t_{\alpha/2}(n-1)S_n^*}{\sqrt{n}}, \tag{2}$$

  where $t_{\alpha/2}(n-1)$ is the $1-\alpha/2$ quantile value of the $t(n-1)$ distribution.

- Going further, in view of the expectation and variance of the $\chi^2(n-1)$ distribution,

$$\mathbb{E}\left((n-1){S_n^*}^2/\sigma^2\right) = n-1,$$

  so

$$\mathbb{E}\left({S_n^*}^2\right) = \sigma^2.$$

This is another proof that the corrected empirical variance is an unbiased estimator of the true (population) variance of the normal distribution. Also,

$$\mathrm{Var}\left((n-1)S_n^{*2}/\sigma^2\right) = 2(n-1),$$

so

$$\mathrm{Var}\,(S_n^{*2}) = \frac{2(n-1)}{(n-1)^2}\sigma^4 = \frac{2\sigma^4}{(n-1)} \to 0$$

as $n \to \infty$. Consequently, S_n^{*2} is an unbiased estimator with "small" variance in the normal case.

- Therefore, for "large" n ($n \geq 30$), even in case of unknown variance the confidence interval of (1) can be updated to

$$I_{1-\alpha} = \bar{X} \pm \frac{z_{\alpha/2} S_n^*}{\sqrt{n}},$$

whereas (2) is mainly applicable for "small" ($n < 30$) sample sizes.

Steiner's theorem, covariance, correlation

- *Steiner's Theorem*: $\mathbb{E}(X-c)^2 = \mathbb{E}(X-\mathbb{E}X)^2 + (\mathbb{E}X-c)^2 \geq \mathrm{Var}\,X$, min. if $c = \mathbb{E}X$.
- p-**quantile** value or 100p-*percentile* of X is x_p if $F(x_p) = p$. **Median**: 0.5-quantile value.
- The **covariance** between X and Y (having finite second moments) is

$$\mathrm{Cov}\,(X,Y) = \mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y) = \mathbb{E}(XY) - \mathbb{E}(X)\cdot\mathbb{E}(Y),$$

while their **correlation** is

$$\mathrm{Corr}\,(X,Y) = \frac{\mathrm{Cov}\,(X,Y)}{\sqrt{\mathrm{Var}\,(X)\cdot\mathrm{Var}\,(Y)}}.$$

By the Cauchy–Schwarz inequality: $|\mathrm{Corr}\,(X,Y)| \leq 1$, and it is ± 1 if and only if $Y = aX + b$.

- $\mathrm{Var}\,(aX+bY) = a^2\mathrm{Var}\,(X) + b^2\mathrm{Var}\,(Y) + 2ab\mathrm{Cov}\,(X,Y)$.
- If X and Y are independent, then $\mathrm{Cov}\,(X,Y) = 0$. The reverse is not usually true, but it is true in case of the following bivariate distribution.
- (X,Y) has **2-variate normal distribution** with parameters $\boldsymbol{\mu}$ and $\boldsymbol{C}$ if its density is

$$f(x,y) = \frac{1}{2\pi|\boldsymbol{C}|^{1/2}} e^{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

where the expectation vector $\boldsymbol{\mu}$ contains the expectations of X and Y in their components, and the 2×2 positive definite **covariance matrix** is

$$\boldsymbol{C} = \begin{pmatrix} \mathrm{Var}\,(X) & \mathrm{Cov}\,(X,Y) \\ \mathrm{Cov}\,(X,Y) & \mathrm{Var}\,(Y) \end{pmatrix}.$$