Probability and Mathematical Statistics

Miskolc, 2025.

Dr. Tamás Glavosits

Lecture 10

Statistics I.,
Point Estimation

1. Required Mathematical Background

Properties of Logarithms

Let $x, y > 0$, $c \in \mathbb{R}$.

1. $\ln(xy) = \ln(x) + \ln(y)$;
2. $\ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y)$;
3. $\ln(x^c) = c\ln(x)$;
4. $\ln(e^x) = x$.

Property 1 also holds for a finite number of factors, that is, if $x_1, x_2, \ldots, x_n$ are positive real numbers, then

$$\ln\left(\prod_{i=1}^{n} = x_i\right) = \sum_{i=1}^{n} \ln(x_i).$$

Properties of Summation

Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be arbitrary real sequences, and $\lambda \in \mathbb{R}$. Then

1. $\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$;
2. $\sum_{i=1}^{n} \lambda x_i = \lambda \sum_{i=1}^{n} x_i$;
3. $\sum_{i=1}^{n} 1 = n$.

Properties of Differentiation

Let $f_1, f_2, \ldots, f_n$ be differentiable functions, and $\lambda \in \mathbb{R}$. Then

1. $\frac{\partial}{\partial v} \sum_{i=1}^{n} f_i(v) = \sum_{i=1}^{n} \frac{\partial}{\partial v} f_i(v)$;
2. $\frac{\partial}{\partial v} \lambda f(v) = \lambda \frac{\partial}{\partial v} f(v)$.

Property 1 expresses the additivity of differentiation, while property 2 expresses its homogeneity.

2. Main Branches of Statistics

Branches of Statistics

- **Sampling**
- **Estimation Theory**
  - point estimation
  - interval estimation
- **Hypothesis Testing**

Sample and Sample Realization

- **Sample:** $x_1, x_2, \ldots, x_n$ are identically distributed random variables. If they are independent, we say that the sample is independent.
- **Sample Realization:** $\xi_1, \xi_2, \ldots, \xi_n$ are real numbers or vectors.
- **Sample size:** n.
  if $n \leqslant 50$ then it is a small sample
  if $50 \leqslant n \leqslant 500$ then it is a medium sample
  if $n \geqslant 500$ then it is a large sample.

Sampling

- **Representative Sample:** every member of the population has an equal chance to be in the sample.
- **Stratified Sampling:** The investigated parameter are related with certain categories of the population. The sampling must reflects this composition of the population.

Estimation

The purpose of statistics (among others) is to determine the unknown parameter(s) of a distribution.

- **Point Estimation:** we provide the number (or vector) which esimates the unknown parameter.
- **Interval Estimation:** we provide an interval which contains the unknown parameter with a given probability.

Statistics, and unbiased estimators

Definition

A sequence of functions $g_n : \mathbb{R}^n \to \mathbb{R}$ or $(\mathbb{R}^n \to \mathbb{R}^k)$ is called a **statistic**.

Definition

A statistic (g_n) for an unknown parameter $\vartheta \in \Theta$ is

- **Unbiased estimator** if $\mathbb{E}(g_n(\xi_1, \xi_2, \ldots, \xi_n)) = \vartheta$.
- **Asymptotically unbiased estimator** if $\lim_{n\to\infty} \mathbb{E}(g_n(\xi_1, \xi_2, \ldots, \xi_n)) = \vartheta$.

Consistent estimator, and More efficient statistic

Definition

Statistic (g_n) is called a **consistent estimator** of the unknown parameter ϑ if

$$\lim_{n\to\infty} \mathbb{P}(|g_n(\xi_1, \xi_2, \ldots, \xi_n) - \vartheta| < \varepsilon) = 0$$

for all $\varepsilon > 0$ (in other words the random variable $g_n(\xi_1, \xi_2, \ldots, \xi_n)$ stochastically convergens to ϑ).

Definition

Let g_n and $g_n^{'}$ be two unbiased estimators of the parameter ϑ (i.e. $\mathbb{E}(g_n(\xi_1, \xi_2, \ldots, \xi_n)) = \vartheta$ and $\mathbb{E}(g_n^{'}(\xi_1, \xi_2, \ldots, \xi_n)) = \vartheta$). We say that the g_n **estimator is more efficient than the** $g_n^{'}$ **estimator** if it has smaller variance, that is

$$\mathbb{D}^2(g_n(\xi_1, \xi_2, \ldots, \xi_n) \leqslant \mathbb{D}^2(g_n^{'}(\xi_1, \xi_2, \ldots, \xi_n).$$

3. Basic Statistics

Basic Statistics

We will get acquainted with the following basic statistics:

- mean ($\overline{\xi}$);
- empirical variance (s_n^2);
- corrected empirical variance (s_n^{*2});
- empirical median (med);
- median absolute deviation (MAD);
- empirical distribution function ($\mathbb{F}_n^*$);
- histograms.

Mean

Definition

The **mean** of the sample $\xi_1, \xi_2, \ldots, \xi_n$, denoted by $\overline{\xi}$, is defined by **sample mean** of the sample

$$\overline{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i.$$

Theorem (Sample Mean)

The sample mean is an unbiased estimator of the theoretical expected value (m).

Proof.

The proof is evident by the additivity, and homogenity of the expectation. □

Sample variance

Definition

Sample (or empirical) variance:

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \overline{\xi})^2.$$

Properties of s_n^2:

1. **Computation:**

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} \xi_i^2 - (\overline{\xi})^2.$$

The quantity $\frac{1}{n}\sum_{i=1}^{n}(\xi_i)^2 := m_2$ is called the sample second moment.
Hence $s_n^2 = m_2 - (\overline{\xi})^2$.

2. **Steiner formula:** for all $a \in \mathbb{R}$:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - (\overline{x} - a)^2.$$

3. If $\xi_1, \xi_2, \ldots, \xi_n$ are pairwise uncorrelated, then

$$\mathbb{E}(s_n^2) = \frac{n-1}{n}\sigma^2.$$

Proof of the Steiner formula

Let $a \in \mathbb{R}$ be arbitrary. Then the "camel rule" (shifting rule) must be applied.

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}((x_i - a) + (a - \overline{x}))^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left((x_i - a)^2 + 2(x_i - a)(a - \overline{x}) + (a - \overline{x})^2\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 + \underbrace{2(a - \overline{x})\frac{1}{n}\sum_{i=1}^{n}(x_i - a)}_{-2(\overline{x}-a)^2} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(a - \overline{x})^2}_{(\overline{x}-a)^2} =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - 2(\overline{x} - a)^2 + (\overline{x} - a)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - (\overline{x} - a)^2.$$

Remark: From the Steiner formula, choosing $a = 0$ gives the first statement.

Proof of $\mathbb{E}(s_n^2) = \frac{n-1}{n}\sigma^2$

Let $\xi_1, \xi_2, \ldots, \xi_n$ be a pairwise uncorrelated sample, with the theoretical expected value m and the theoretical variance σ^2. To prove the result, we apply the Steiner formula with $a = m$. Based on statement 2 and the additivity and homogeneity of expectation, we get:

$$\mathbb{E}(s_n^2) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\xi_i - m)^2 - \mathbb{E}(\overline{\xi} - m)^2.$$

It is obvious that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\xi - m)^2 = \frac{1}{n}n\sigma^2 = \sigma^2.$$

Now we show that $\mathbb{E}(\overline{\xi} - m)^2 = \frac{1}{n}\sigma^2$. In the proof, we use the identity for the square of an n-term sum, that is,

$$\left(\sum_{i=1}^{n} a_i\right)^2 = \sum_{i=1}^{n} a_i^2 + 2\sum_{i<j} a_i a_j.$$

Using the pairwise uncorrelatedness, we obtain that

$$\begin{aligned}
\mathbb{E}(\overline{\xi} - m)^2 &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i - \frac{1}{n}\sum_{i=1}^{n} m\right)^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i - m)\right)^2 = \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}(\xi_i - m)\right)^2 = \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}(\xi_i - m)^2 + 2\sum_{i<j}(\xi_i - m)(\xi_j - m)\right) = \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}(\xi_i - m)^2 + 2\sum_{i<j}\underbrace{\mathbb{E}(\xi_i - m)(\xi_j - m)}_{0}\right) = \\
&= \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2.
\end{aligned}$$

From the previous steps, we obtain that

$$\mathbb{E}(s_n^2) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\xi_i - m)^2 - \mathbb{E}(\overline{\xi} - m)^2 = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2.$$

Theorem

The sample variance is an asymptotically unbiased estimator of the theoretical variance.

Definition

Corrected sample variance:

$$s_n^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \overline{\xi})^2,$$

however, it can also be calculated as

$$s_n^{*2} = \frac{n}{n-1} s_n^2.$$

Theorem

s_n^{*2} is an unbiased estimator of the theoretical variance.

Proof.

It is evident by the fact that $\mathbb{E}(s_n^*) = \frac{n-1}{n}\sigma^2$. □

Definition

If ξ is an absolutely continuous random variable, then the (theoretical) **median** of ξ is the number $\nu \in \mathbb{R}$ such that $\mathbb{F}(\nu) = \frac{1}{2}$.

Definition

Let the ordered sample be: $\xi_1^* \leqslant \xi_2^* \leqslant \cdots \leqslant \xi_n^*$.
Sample median:

$$\operatorname{med}(\xi_i) := \begin{cases} \dfrac{\xi_k^* + \xi_{k+1}^*}{2}, & \text{if } n = 2k, \\ \xi_k^*, & \text{if } n = 2k-1. \end{cases}$$

Definition

Let med denote the sample median of $\xi_1, \xi_2, \ldots, \xi_n$. Then the median absolute deviation is defined by

$$\operatorname{MAD}(\xi_1, \ldots, \xi_n) = \operatorname{med}(|\xi_1 - \operatorname{med}|, \ldots, |\xi_n - \operatorname{med}|)$$

Definition

Empirical distribution function: Let $\xi_1^* < \xi_2^* < \cdots < \xi_n^*$ be an ordered sample. Then the empirical distribution function $\mathbb{F}_n^* : \mathbb{R} \to \mathbb{R}$ defined by

$$\mathbb{F}_n^*(x) := \begin{cases} 0, \text{ ha } x \leqslant \xi_1^*, \\ \dfrac{k}{n}, \text{ ha } \xi_k^* < x \leqslant \xi_{k+1}^*, \\ 1, \text{ ha } x > \xi_n^*. \end{cases}$$

Remark

It is easy to see that the funciton $\mathbb{F}_n^*$ is a monotonically increasing, left-continuous step function that jumps by $\frac{1}{n}$ at each element of the ordered sample.
The empirical distribution function can also be introduced by the definition

$$\mathbb{F}_n^*(x) := \frac{1}{n} \sum_{i=1}^{n} I_{]\xi_i, +\infty[}(x),$$

where the function $I_H : \mathbb{R} \to \mathbb{R}$ is defined by

$$I_H(x) := \begin{cases} 1, \text{ ha } x \in H; \\ 0, \text{ egyébként.} \end{cases}$$

for easy set $H \subseteq \mathbb{R}$. The function I_H is **the indicator function** of the set H.
This second definition is preferable, as the first one because the second definition does not requires the elements of sample to be distinct.

Glivenko-Cantelli Theorem

Theorem

Glivenko-Cantelli theorem, or ***the fundamental theorem of mathematical statistics:***

$$\mathbb{P}\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|\mathbb{F}_n^*(x)-\mathbb{F}_n(x)|=0\right)=1$$

where $\mathbb{F}$ *is the theoretical distribution function, which means that the empirical distribution function converges to the theoretical distribution function of the population.*

4. Estimation Methods

Maximum likelihood estimator

Assume that $\xi_1, \xi_2 \ldots, \xi_n$ is an independent sample with unknown parameter ϑ.

$$L(\xi_1, \ldots, \xi_n, \vartheta) = \begin{cases} \prod_{i=1}^{n} p(\xi_i, \vartheta) & \text{in the case of a discrete sample,} \\ \prod_{i=1}^{n} f(\xi_i, \vartheta) & \text{in the case of an absolutely continuous} \end{cases}$$

where $p(\cdot, \vartheta)$ is the theoretical discrete distribution, and $f(\cdot, \vartheta)$ is the theoretical densitiy function. The above cases ϑ is the unknown parameter. The function L is **the likelihood function**.
Let $\widehat{\vartheta}$ denote the maximum likelihood estimator of the unknown parameter $\vartheta \in \Theta$.

Determination of $\widehat{\vartheta}$:

The $\widehat{\vartheta}$ is the maximum point of function $L(\cdot)$, that is

$$L(\widehat{\vartheta}) = \max_{\vartheta \in \Theta} L(\vartheta).$$

For the technical procedure, we use the theorem from analysis that if a differentiable function has a local maximum or minimum at a point, then its first derivative vanishes in this point. The points where the first derivative of a differentiable function vanishes are called the stationary points of the function. Therefore, we need to find the stationary points of the function $\vartheta \to L(\vartheta)$.

Since the function $\vartheta \to L(\vartheta)$ is of product form, and differentiating a product is not easy, we take the logarithm of the function $\vartheta \to L(\vartheta)$, i.e.

$$\vartheta \to l(\vartheta) := \ln(L(\vartheta))$$

which is called the log-likelihood function and denoted by $l(\vartheta)$. The log-likelihood function has several useful properties:

- since the function $\ln$ is strictly increasing, the function $\vartheta \to l(\vartheta)$ has the same maximum points as the function $\vartheta \to L(\vartheta)$;
- the logarithm turns a product into a sum, making differentiation easier.

In summary: We need to solve the equation $\dfrac{\partial}{\partial \vartheta} l(\vartheta) = 0$, and then check that the obtained $\widehat{\vartheta}$ is indeed a maximum of the function $\vartheta \to l(\vartheta)$. (Which will be omitted in the sequel.)

Properties of the Maximum Likelihood Method:

- It does not always provide an unbiased estimator;
- It always provides a consistent estimator;
- It gives the most efficient estimator (if one exists).

k-th Theoretical, and Empirical Moment

If ξ is a random variable, then the k-**th theoretical moment** of ξ is defined by

$$\nu_k := \mathbb{E}(\xi^k)$$

(if it exists).
If $\xi_1, \xi_2, \ldots, \xi_n$ is an independent sample such that the theoretical distribution it comes from has a k-th moment, then this is approximated by the k-**th empirical moment**, which is defined by

$$m_k := \frac{1}{n} \sum_{i=1}^{n} \xi_i^k.$$

Method of Moments

Given an independent sample $\xi_1, \xi_2, \ldots, \xi_n$ with unknown parameters, and our purpose is to determine the unknown parameters.
The unknown parameters can be expressed using theoretical moments, although this often requires solving a nonlinear system of equations.
In estimating the unknown parameters, the empirical moments are used instead of the theoretical moments.

For example:

- The independent sample comes from a $\mathrm{Poiss}(\lambda)$ distribution, and the purpose is to estimate the unknown parameter λ. Since $\mathbb{E}(\xi) = \lambda$, the estimator of the unknown parameter λ is $\widehat{\lambda} = \overline{\xi}$. In this case, because $\mathbb{D}^2(\xi) = \lambda$, the unknown parameter could also be estimated by $\widehat{\lambda} = m_2 - (\overline{\xi})^2$, but in this case the estimator $\widehat{\lambda} = \overline{\xi}$ should be chosen, because it is of lower order.
- If the independent sample comes from an $\mathrm{Exp}(\lambda)$ distribution, then $\mathbb{E}(\xi) = \frac{1}{\lambda}$, so the unknown parameter λ can be estimated by $\widehat{\lambda} = \frac{1}{\overline{\xi}}$.
- If the unknown sample comes from a $\mathcal{N}(m, \sigma^2)$ distribution, then the unknown parameter m can be estimated by $\widehat{m} = \overline{\xi}$, and the unknown parameter σ^2 can be estimated by $\widehat{\sigma^2} = m_2 - (\overline{\xi})^2$.

End of Lecture 10