People's Democratic Republic of Algeria
University Med Khider of Biskra
Faculty of SNVSTU

Estimation

Dr. Ben Gherbal Hanane

Lecture: Data Analysis in Biosciences — Level: L3
Email: hanane.bengherbal@univ-biskra.dz

Introduction

In biosciences, accurate estimation of population parameters such as the mean or variance is essential for analyzing experimental data. This lecture introduces the fundamental concepts of statistical estimation, including point estimation and confidence intervals. We explore how to estimate parameters from sample data using unbiased and consistent estimators, and how to construct confidence intervals based on sample statistics under different conditions. Several real-world examples in biology and medicine are provided to illustrate the practical application of these methods.

1. Notations

Let X be a random variable such that $E(X) = \mu$ and $V(X) = \sigma^2$.

Estimation of μ: The sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator of μ (unbiased: $E(\overline{X}) = \mu$; and consistent: $V(\overline{X}) = \frac{\sigma^2}{n} \to 0$).

Estimation of σ^2 when μ is known: When the mean $\mu = E(X)$ of the population is known, then $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$. We have $E(S_n^2) = \sigma^2$, so S_n^2 is an unbiased and consistent estimator of σ^2.

Estimation of σ^2 when μ is unknown: In general, $\mu = E(X)$ is unknown and is replaced by its estimator $\overline{X}$, leading to the empirical variance: $S_e^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

S_e^2 is a biased but consistent estimator of σ^2, and asymptotically unbiased: $E(S_e^2) = \frac{n-1}{n}\sigma^2$.

Theorem 1. *The corrected sample variance is:* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. S^2 *is an unbiased and consistent estimator of* σ^2. *So,* S^2 *is the best estimator of* σ^2. *Or:* $S^2 = \frac{n}{n-1}S_e^2$.

2. Point Estimation

The first approach to estimate a parameter is point estimation. This is based on the notion of an estimator, defined as follows:

Definition 2. *Let* X *be a random variable whose distribution depends on a parameter* θ, *and let* $X_1, X_2, ..., X_n$ *be a random sample of* X. *A point estimator of* θ *is a statistic* $\widehat{\theta}$ *of the form* $\widehat{\theta} = h(X_1, X_2, ..., X_n)$, *satisfying certain properties.*

Some examples:

- To estimate the mean $\mu = E(X)$ of a random variable X, consider the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- To estimate the variance $\sigma^2 = V(X)$, use the sample variance: $S_e^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, and the corrected (unbiased) sample variance: $S^2 = \frac{n}{n-1}S_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Remark 3. *There are several methods for determining point estimators, including: the maximum likelihood method, the method of moments, and the least squares method.*

Theorem 4. *Let* $X_1, X_2, ..., X_n$ *be a random sample of size n from a random variable* X *with mean* $\mu = E(X)$ *and variance* $\sigma^2 = V(X)$. *Then:* $E(\overline{X}) = \mu$, $V(\overline{X}) = \frac{\sigma^2}{n}$, $E(S^2) = \sigma^2$.

3. Confidence Interval Estimation

An estimation is called an interval estimation when we estimate an unknown parameter θ (e.g., a mean) by constructing an interval $[a, b]$, such that: $P(a < \theta < b) = 1 - \alpha$

Where:

- a and b are the confidence bounds.
- $1 - \alpha$ is the confidence level (close to 1, e.g., 0.90 or 0.99) and α is the error risk.

3.1. Confidence Intervals for a Mean

We want to estimate the mean μ of a normal population using a random sample. Let X be a random variable with mean μ and variance σ^2, and let $X_1, X_2, ..., X_n$ be a random sample of size n. For a given confidence level $1 - \alpha$, the confidence intervals (CI) for μ are summarized as follows: Where

Situation	**Distribution used**	**Confidence Interval (Level $1 - \alpha$)**
σ known, $X \sim \mathcal{N}(\mu, \sigma^2)$	$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$	$\mu \in \left[\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$
σ unknown, $X \sim \mathcal{N}(\mu, \sigma^2)$, $n < 30$	$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$	$\mu \in \left[\overline{X} \pm t_{(\alpha/2, n-1)} \cdot \frac{S}{\sqrt{n}}\right]$
σ unknown, $n > 30$	$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim \mathcal{N}(0, 1)$	$\mu \in \left[\overline{X} \pm z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}\right]$

Table 1: Confidence Intervals for the Mean under Different Conditions

$z_{\alpha/2}$ is obtained from the standard normal table such that $F(Z) = 1 - \frac{\alpha}{2}$, and S is the estimator of σ. The value $t_{(\alpha/2,n-1)}$ is read from the Student distribution table with $n - 1$ degrees of freedom.

Example 5. *The weight X of containers of a product follows a normal distribution with standard deviation $\sigma = 0.3\,g$. A random sample of 100 containers yields a sample mean of $\overline{X} = 49.7\,g$. Compute a 95 % confidence interval for μ.*

Solution 6. *Since σ is known and n is large, the CI is:* $\mu \in \left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

Given $1-\alpha = 0.95$, $\alpha = 0.05$, *and from the standard normal table,* $z_{\alpha/2} = 1.96$*:*

$$\mu \in \left[49.7 - 1.96 \cdot \frac{0.3}{\sqrt{100}}, 49.7 + 1.96 \cdot \frac{0.3}{\sqrt{100}}\right] = [49.64, 49.76]$$

Example 7. *A random sample of 20 students of the same age and gender has a sample mean height of 1.73m and a sample standard deviation of 0.1m. Compute a 95% CI for the population mean height.*

Solution 8. *Here,* σ *is unknown and* $n < 30$.

With $n = 20$, $\overline{X} = 1.73\,m$,

$$S^2 = \frac{n}{n-1}S_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{20}{19}(0.1)^2 = 0.0105 \approx 0.011$$

From the Student table: $t_{(0.05,19)} = 2.093$, *then:*

$$\mu \in \left[1.73 - 2.093 \cdot \sqrt{\frac{0.011}{20}}, 1.73 + 2.093 \cdot \sqrt{\frac{0.011}{20}}\right] = [1.681, 1.779]$$

The average height of students in the population is therefore within the interval $[1.68\,m, 1.78\,m]$ *with a probability of 0.95.*

Additional Examples

Example 1: Estimating Bacterial Growth Rate

In an experiment studying the effect of temperature on *E. coli* growth, the optical density (OD600) was measured after 24 hours for 10 cultures incubated at 37°C:

$$0.42, 0.39, 0.45, 0.41, 0.43, 0.44, 0.40, 0.42, 0.43, 0.41$$

- Estimate the mean and standard deviation of the optical density.
- Calculate the 95% confidence interval for the mean growth rate.

Example 2: Seed Germination Under Salt Stress

30 barley seeds were grown in a saline environment. The number of germinated seeds after 7 days for 10 trays was:

$$23, 25, 24, 22, 26, 24, 25, 23, 22, 24$$

- Compute the average number of germinated seeds.
- Determine the 95% confidence interval.
- Compare with a control group mean of 27 seeds.

Example 3: Hemoglobin Concentration in Lab Mice

The hemoglobin concentration (g/dL) in 12 mice was recorded as:

$$13.5, 13.7, 14.1, 13.9, 13.8, 14.0, 13.6, 13.8, 14.2, 13.7, 13.9, 14.0$$

- Estimate the mean and standard deviation.
- Determine the 95% confidence interval.

Example 4: Drug Effect on Mouse Weight

Eight mice were given a drug, and their weights (g) were recorded:

$$22.5, 23.1, 22.8, 23.0, 23.2, 22.9, 23.3, 23.0$$

- Estimate the mean and confidence interval at 95%.
- Compare with untreated mice average weight: 22.6g.

Example 5: Frog Body Temperature

Body temperatures (°C) of 6 frogs were measured after relocation:

$$23.4, 23.6, 23.5, 23.7, 23.8, 23.5$$

- Calculate the average body temperature and its standard deviation.
- Provide a 95% confidence interval and interpret it biologically.

Example 6:

A machine fills bottles with a mean μ and known standard deviation $\sigma = 0.2$ L. A sample of 40 bottles gave a mean fill of 1.02 L.

- Construct a 95% confidence interval for the true mean fill level.
- Since σ is known and $n = 40$, use the Z-distribution.

Example 7:

A study records the weights (in kg) of 50 individuals. The sample mean is 70 kg and the sample standard deviation is 8.5 kg.

- Estimate the population mean weight.
- Construct a 95% confidence interval using the sample statistics.
- Since $n = 50$ and σ is unknown, use normal approximation (Z).

Example 8: Cholesterol Levels in a Clinical Trial

A clinical study was conducted to evaluate the effect of a dietary supplement on cholesterol levels. After 30 days, the total cholesterol (mg/dL) of 35 patients was measured. The sample mean was 198.4 and the sample standard deviation was 14.2.

- Estimate the true average cholesterol level in the population.
- Construct a 95% confidence interval for the mean.