Mathematics for Data Science

Lecture 3: Point Estimation

Mohamad GHASSANY

EFREI Paris

Introduction to Statistical Inference

- **Statistics** is the science of collecting, processing and analyzing data derived from the observation of random phenomena.
- Data analysis is used to **describe** the phenomena studied, **make predictions** and **make decisions** about them. In this way, statistics is an essential tool for understanding and managing complex phenomena.
- The data studied can be of any nature, which makes statistics useful in all disciplinary fields.

The fundamental point is that the data present uncertainties and **variations**.

Statistical methods are divided into two classes:

- **Descriptive statistics**, **exploratory statistics** or **data analysis**, aims to summarize the information contained in the data in a synthetic and efficient way. Probabilities play only a minor role here.
- **Inferential statistics** goes beyond the simple description of data. Its purpose is to **make predictions** and **make decisions** based on observations. In general, it is necessary to propose **probabilistic models** of the studied random phenomenon and to know how to manage the risks of errors. Probabilities play a fundamental role here.
- **Probability** can be considered as a branch of pure mathematics, based on the theory of measurement, abstract and completely disconnected from reality.
- **Applied probability** proposes **probabilistic models** of the course of concrete random phenomena. One can then, **prior to any experiment**, make predictions about what will happen.

Example: it is usual to model the duration of the good functioning or life of a system, let's say a light bulb, by a random variable X of exponential law of parameter λ. Having adopted this probabilistic model, we can perform all the calculations we want. For example:

- The probability that the bulb has not yet failed at date t is $P(X > t) = e^{-\lambda t}$.
- The average lifetime is $E(X) = 1/\lambda$.
- If n identical light bulbs are turned on at the same time, and they work independently of each other, the number N_t of light bulbs that will fail before a time t is a random variable of binomial distribution $\mathcal{B}(n, P(X \leq t)) = \mathcal{B}(n, 1 - e^{-\lambda t})$. Thus we expect that, on average, $E(N_t) = n(1 - e^{-\lambda t})$ bulbs will fail between 0 and t.

In practice, if we want to use the theoretical results stated above, we have to make sure that we have chosen a good model, i.e.that the life span of these bulbs is a random variable with an exponential law, and, on the other hand, we have to be able to calculate the value of the parameter λ in some way. It is statistics that will allow us to solve these problems. To do this, we need to do an **experiment**, **collect data** and **analyze** them.

We therefore set up what we call a **test** or an **experiment**. We run $n = 10$ identical bulbs in parallel and independently of each other, under the same experimental conditions, and we record their lifetimes. Let's say that we obtain the following lifetimes, expressed in hours: $91.6, 35.7, 251.3, 24.3, 5.4, 67.3, 170.9, 9.5, 118.4, 57.1$

Let us note $x_1, \dots, x_n$ these observations. We will therefore consider that $x_1, \dots, x_n$ are the ***samples*** of random variables $X_1, \dots, X_n$.

This means that after the experiment, the lifetime has been observed. We say that x_i is a sample (a realization) of X_i on the test performed.

Since the bulbs are identical, it is natural to suppose that X_i have the same law. This means that the same random phenomenon is observed several times.

We can also assume that the X_i are independent random variables. We can then ask the following questions:

1. With respect to these observations, is it reasonable to assume that the lifetime of a light bulb is a random variable with an exponential distribution? If not, what other law would be more appropriate? This is a **fit test** (Chi-square test) problem.
2. If the exponential distribution model has been chosen, how can we propose a good value (or set of values) for the parameter λ? This is a parametric **estimation** problem.
3. In this case, can we guarantee that λ is less than a fixed value λ_0? This will guarantee that $E(X) = 1/\lambda \geq 1/\lambda_0$, in other words that the bulbs will be sufficiently reliable. This is a **parametric hypothesis testing** problem.
4. If we have 100 light bulbs, how many failures can we expect in less than 50 hours? This is a **prediction** problem.

Definition: Random sample

The random variables $X_1, X_2, \dots, X_n$ are a random sample of size n if

a. the X_i's are independent random variables
b. every X_i has the same probability distribution.

An observation (realization) of the sample is $(x_1, \dots, x_n)$.

Definition: Statistic

A **statistic** is any function of the observations in a random sample.

$$T(X) = T(X_1, \dots, X_n)$$

For example, each of $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, X_1^2 or $(X_1, X_3 + X_4, 2\ln X_6)$ is a statistic.

Since a statistic is a random variable, it has a probability distribution.

The probability distribution of a statistic is called a *sampling distribution*.

If $X_1, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2 , and if $\overline{X}_n$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$, is the **standard normal distribution** $\mathcal{N}(0,1)$.

If we have two independent populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, and if $\overline{X}_1$ and $\overline{X}_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations, then the sampling distribution of

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is **approximately standard normal** if the conditions of the central limit theorem apply. If the two populations are normal, the sampling distribution of Z is exactly standard normal.

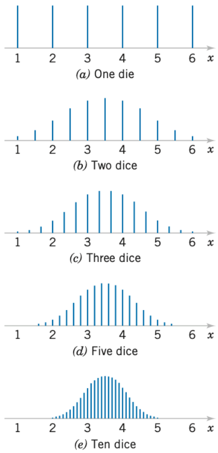

Figure 1: Distributions of average scores from throwing dice.

Definition: Point estimator

A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic T_n. The statistic T_n is called the point estimator.

A point estimator is a random variable. An estimation is a value.

Estimation problems occur frequently in engineering. We often need to estimate

- The mean μ of a single population
- The variance σ^2 (or standard deviation σ) of a single population
- The proportion p of items in a population that belong to a class of interest
- The difference in means of two populations, $\mu_1 - \mu_2$
- The difference in two population proportions, $p_1 - p_2$

Unbiased Estimators

An estimator should be "close" in some sense to the true value of the unknown parameter.

Definition: Bias of an Estimator

The point estimator T_n is an unbiased estimator for the parameter θ if

$$E(T_n) = \theta$$

If the estimator is not unbiased, then the difference $E(T_n) - \theta$ is called the bias of the estimator T_n .

Formally, we say that T_n is an unbiased estimator of θ if the expected value of T_n is equal to θ. This is equivalent to saying that the mean of the probability distribution of T_n (or the mean of the sampling distribution of T_n) is equal to θ.

When an estimator is unbiased, the bias is zero; that is, $E(T_n) - \theta = 0$.

The bias measures a systematic error of estimation. If $E(T_n) - \theta < 0$, T_n tends to under-estimate θ.

Sample Mean and Variance

Suppose that X is a random variable with mean μ and variance σ^2. Let $X_1, \ldots, X_n$ be a random sample of size n from the population represented by X.

- Show that the sample mean $\overline{X}_n$ is an unbiased estimator of μ.
- Suggest an an unbiased estimator of σ^2.

Minimum Variance Unbiased Estimator

If we consider all unbiased estimators of θ, the one with the smallest variance is called the **minimum variance unbiased estimator** (MVUE).

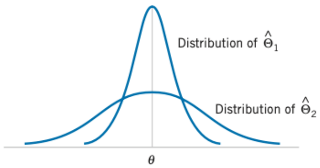

Figure 2: The sampling distributions of two unbiased estimators.

If $X_1, \dots, X_n$ is a random sample of size n from a normal distribution with mean μ and variance σ^2, the sample mean $\overline{X}_n$ is the MVUE for μ.

Mean Squared Error of an Estimator

Definition: Mean Squared Error of an Estimator

The mean squared error of an estimator T_n of the parameter θ is defined as

$$MSE(T_n) = E[(T_n - \theta)^2]$$

The mean squared error can be rewritten as follows:

$$\begin{aligned} MSE\left(T_n\right) &= E\left[\left(T_n-\theta\right)^2\right] = E\left[\left(T_n-E\left(T_n\right)+E\left(T_n\right)-\theta\right)^2\right] \\ &= E\left[\left(T_n-E\left(T_n\right)\right)^2\right]+2E\left[T_n-E\left(T_n\right)\right]E\left[E\left(T_n\right)-\theta\right]+E\left[\left(E\left(T_n\right)-\theta\right)^2\right] \\ &= \operatorname{Var}\left(T_n\right)+\left[E\left(T_n\right)-\theta\right]^2 \\ &= \text{ Variance of the estimator } + \text{ squarred bias} \end{aligned}$$

Methods of Point Estimation

Methods of Point Estimation

In this section, we discuss methods for obtaining point estimators: **the method of moments** and **the method of maximum likelihood**.

- Maximum likelihood estimates are generally preferable to moment estimators because they have better efficiency properties.
- However, moment estimators are sometimes easier to compute.
- Both methods can produce unbiased point estimators.

Method of Moments

The general idea behind the method of moments is to **equate** population moments, which are defined in terms of expected values, to the corresponding sample moments.
The population moments will be functions of the unknown parameters. Then these equations are solved to yield estimators of the unknown parameters.

If $E(X) = \phi(\theta)$, where ϕ is an invertible function, the moment estimator of θ is $\hat{\theta}_n = \phi^{-1}(\overline{X}_n)$.

For example, if the parameter to estimate is the expected value of X_i, the **moment estimator** of $E(X)$ is the sample mean $\overline{X}_n$.

Examples:

- Exponential distribution
- Normal distribution
- Gamma distribution

Suppose that X is a random variable with probability distribution depending on a single unknown parameter θ. Let $x_1, \dots, x_n$ be the observed values in a random sample of size n. Then the likelihood function of the sample is

$$\mathcal{L}(\theta; x_1, \dots, x_n) = \begin{cases} P(X_1 = x_1, \dots, X_n = x_n; \theta) & \text{if } X_i \text{ are discrete} \\ f_{X_1,\dots,X_n}(x_1, \dots, x_n; \theta) & \text{if } X_i \text{ are continuous} \end{cases}$$

After supposing that all the X_i are independant:

$$\mathcal{L}(\theta; x_1, \dots, x_n) = \begin{cases} \prod_{i=1}^{n} P(X_i = x_i; \theta) = \prod_{i=1}^{n} P(X = x_i; \theta) & \text{if } X_i \text{ are discrete} \\ \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \prod_{i=1}^{n} f(x_i; \theta) & \text{if } X_i \text{ are continuous} \end{cases}$$

Maximum likelihood estimator

Note that the likelihood function is now a function of only the unknown parameter θ. The **maximum likelihood estimator (MLE)** of θ is the value of θ that maximizes the likelihood function $\mathcal{L}(\theta)$ (or its $\ln$).

Examples:

- Bernoulli distribution
- Exponential distribution
- Normal distribution