Point estimation

by Marco Taboga, PhD

Point estimation is a type of statistical inference which consists in producing a guess or approximation of an unknown parameter.

In this lecture we introduce the theoretical framework that underlies all point estimation problems.

At the end of the lecture, we provide links to detailed examples of point estimation, in which we show how to apply the theory.

Sample and data-generating distribution

The main elements of a point estimation problem are those found in any statistical inference problem:

- we have a sample that has been drawn from a probability distribution whose characteristics are at least partly unknown;
- the sample ξ is regarded as the realization of a random vector Ξ;
- the joint distribution function of Ξ, denoted by $F_{\Xi}(\xi)$, is assumed to belong to a set of distribution functions Φ, called statistical model.

phi

Parametric model

When the model Φ is put into correspondence with a set $\Theta \subseteq \mathbb{R}^p$ of real vectors, then we have a parametric model.

The set Θ is called the parameter space and its elements are called parameters.

Denote by θ_0 the parameter that is associated with the data-generating distribution $F_{\Xi}(\xi)$ and assume that θ_0 is unique. The vector θ_0 is called the **true parameter**.

Estimate and estimator

parameter space

Point estimation is the act of choosing a vector $\hat{\theta} \in \Theta$ that approximates θ_0. The approximation $\hat{\theta}$ is called an **estimate** (or point estimate) of θ_0.

When the estimate $\hat{\theta}$ is produced using a predefined rule (a function) that associates a parameter estimate $\hat{\theta}$ to each ξ in the support of Ξ, we can write

$$\hat{\theta} = \hat{\theta}(\xi)$$

The function $\hat{\theta}(\xi)$ is called an **estimator**.

Often, the symbol $\hat{\theta}$ is used to denote both the estimate and the estimator. The meaning is usually clear from the context.

Estimation error

According to the decision-theoretic terminology introduced previously, making an estimate $\hat{\theta}$ is an act, which produces consequences.

Among these consequences, the most relevant one is the estimation error

$$e = \hat{\theta} - \theta_0$$

The statistician's goal is to commit the smallest possible estimation error.

Loss

The preference for small errors can be formalized with a **loss function**

$$L(\hat{\theta}, \theta_0)$$

[WOW! It's so fucking easy! intuitive!!]

that quantifies the loss incurred by estimating θ_0 with $\hat{\theta}$.

Examples of loss functions are:

1. the absolute error:
$$L(\hat{\theta}, \theta_0) = \|\hat{\theta} - \theta_0\|$$
where $\| \ \|$ is the Euclidean norm (it coincides with the absolute value when $\Theta \subseteq \mathbb{R}$);
2. the squared error:
$$L(\hat{\theta}, \theta_0) = \|\hat{\theta} - \theta_0\|^2$$

So different ways to estimate the error [we want to minimize this!]

Risk

When the estimate $\hat{\theta}$ is obtained from an estimator, it is a function of the random vector Ξ and the loss

$$L(\hat{\theta}(\Xi), \theta_0)$$

Why?

is a random variable.

The expected value of the loss

$$R(\hat{\theta}) = \mathrm{E}\left[L(\hat{\theta}(\Xi), \theta_0)\right]$$

is called the **statistical risk** (or, simply, the risk) of the estimator $\hat{\theta}$.

What does this mean? Taking avg of what?

$\hat{\theta}$ from different samples?

Estimates of risk

The expected value in the definition of risk is computed with respect to the true distribution function $F_{\Xi}(\xi)$.

Therefore, we can compute the risk $R(\hat{\theta})$ only if we know the true parameter θ_0 and $F_{\Xi}(\xi)$.

When θ_0 and $F_{\Xi}(\xi)$ are unknown, the risk needs to be estimated.

For example, we can approximate the risk with the quantity

$$\mathrm{E}\left[L(\hat{\theta}(\Xi), \hat{\theta})\right]$$

where:

- we pretend that the estimate $\hat{\theta}$ is the true parameter;
- we denote the estimator of $\hat{\theta}$ by $\hat{\theta}$;
- we compute the expected value with respect to the estimated distribution function $F_{\Xi}(\xi; \hat{\theta})$.

how does this work. Did the estimator not spit the estimate?

Even if the risk is unknown, the notion of risk is often used to derive theoretical properties of estimators.

Risk minimization

Point estimation is always guided, at least ideally, by the principle of risk minimization, that is, by the search for estimators that minimize the risk.

Common risk measures

Depending on the specific loss function we use, the statistical risk of an estimator can take different names:

1. when the absolute error is used as a loss function, then the risk
$$R(\hat{\theta}) = \mathrm{E}\left[\|\hat{\theta} - \theta_0\|\right]$$
is called the Mean Absolute Error (MAE) of the estimator.
2. when the squared error is used as a loss function, then the risk
$$R(\hat{\theta}) = \mathrm{E}\left[\|\hat{\theta} - \theta_0\|^2\right]$$
is called Mean Squared Error (MSE). The square root of the mean squared error is called root mean squared error (RMSE).

Other criteria to evaluate estimators

In this section we discuss other criteria that are commonly used to evaluate estimators.

Unbiasedness

If an estimator produces parameter estimates that are on average correct, then it is said to be unbiased.

The following is a formal definition.

Definition Let θ_0 be the true parameter. An estimator $\hat{\theta}$ is an unbiased estimator of θ_0 if and only if

$$\mathrm{E}[\hat{\theta}] = \theta_0$$

If an estimator is not unbiased, then it is called a biased estimator.

If an estimator is unbiased, then the estimation error is on average zero:

$$\begin{aligned} \mathrm{E}[e] &= \mathrm{E}[\hat{\theta} - \theta_0] \\ &= \mathrm{E}[\hat{\theta}] - \theta_0 \\ &= \theta_0 - \theta_0 = 0 \end{aligned}$$

Consistency

If an estimator produces parameter estimates that converge to the true value when the sample size increases, then it is said to be consistent.

n → sequence of sample

The following is a formal definition.

Definition Let $\{\xi_n\}$ be a sequence of samples such that all the distribution functions $F_{\Xi_n}(\xi_n)$ are put into correspondence with the same parameter θ_0. A sequence of estimators $\{\hat{\theta}_n(\xi_n)\}$ is said to be consistent (or weakly consistent) if and only if

$$\operatorname*{plim}_{n \to \infty} \hat{\theta}_n = \theta_0$$

(do the numbers of sample increase

In any order?

where plim indicates convergence in probability. The sequence of estimators is said to be strongly consistent if and only if

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_0$$

Does it mean sample size also has to increase?

where $\xrightarrow{a.s.}$ indicates almost sure convergence. A sequence of estimators which is not consistent is called inconsistent.

When the sequence of estimators is obtained using the same predefined rule for every sample ξ_n, we often say, with a slight abuse of language, "consistent estimator" instead of saying "consistent sequence of estimators". In such cases, what we mean is that the predefined rule produces a consistent sequence of estimators.

Examples

You can find detailed examples of point estimation in the lectures on:

- Point estimation of the mean;
- Point estimation of the variance.

How to find a point estimator

The methods to find point estimators are called estimation methods.

You can read about these methods here:

- Estimation methods;
- Maximum likelihood estimation.

Point vs interval estimation

There is another kind of estimation, called set estimation or interval estimation.

While in point estimation we produce a single estimate meant to approximate the true parameter, in set estimation we produce a whole set of estimates meant to include the true parameter with high probability.

How to cite

Please cite as:

Taboga, Marco (2021). "Point estimation", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. Online appendix. https://www.statlect.com/fundamentals-of-statistics/point-estimation.