# [STAT-13] Parameter estimation

**Miguel-Angel Canela**
**Associate Professor, IESE Business School**

### Statistical inference

Roughly speaking, **statistical inference** is the process of drawing conclusions about a **probability model**, based on a data set. The conclusions of statistical inference are usually related to the values of some **parameters** of the model.

More specifically, inference is concerned with one of the following tasks:

- **Estimation**. In many statistical analyses, we assume that the data have been sampled from a probability distribution that is known, except for the values of one or more parameters. I denote here the parameter by $\theta$, accepting that $\theta$ can be multidimensional (I use boldface in that case). The range of acceptable values of the parameters is called the **parameter space**. Example: for a normal distribution, the two-dimensional parameter is $\boldsymbol{\theta} = (\mu, \sigma^2)$, and the parameter space is $\mathbb{R} \times (0, +\infty)$.
- **Testing**. In hypothesis testing, we are concerned with the values of some unknown parameters of a prespecified model. We set a formal hypothesis about these parameters, such as $\mu_1 = \mu_2$, or $\rho = 0$, and the analysis leads to accepting or rejecting that hypothesis. In academic papers, the statistical hypothesis tested is related to some theoretical hypothesis, usually in a way that the rejection of the statistical hypothesis supports the theoretical hypothesis.
- **Prediction**. Another form of inference deals with the prediction of random variables not yet observed. For instance, we can model the arrival times for customers with an exponential distribution, wishing to predict the arrival time of the next customer. In certain applications, such as modelling stock prices with time series models, prediction is the key issue, and models are evaluated in terms of the accuracy of their predictions.
- **Decision**. In certain contexts, after the data have been analyzed, we must choose within a class of decisions with the property that the consequences of each decision depend on the unknown value of some parameter. For instance, the health authorities may decide about giving the green light to a new drug, based on the results of a clinical trial. **Decision theory** is not covered in this course.
- **Experimental design**. In the experimental sciences, the researcher develops, before the data collection, a detailed plan in which the values of some independent variables are specified. Such a plan is called a **experimental design**. Guidelines for developing experimental designs are usually included in courses for experimental researchers (including psychology and market research). I skip this here, since (most of) you are expected to deal with **observational data**, for which such designs are not feasible. Experimental design and the subsequent data analysis are called **conjoint analysis** in market research and **policy capturing** in organizational research.

The rest of this course is concerned with estimation and testing. This lecture sets the framework for the assessment of estimation methods.

## Estimators

A statistic can be used as an **estimator** of an unknown parameter of a parameter. We distinguish between the estimator and its individual values, called **estimates**. Since there may be many potential estimators for a parameter, e.g. the mean and the median for $\mu$ in the $\mathcal{N}(\mu, \sigma^2)$ distribution, we want to use the estimators with better properties. Thus, textbooks discuss the desirable properties that an estimator may have. For instance, **linear estimators**, based on linear expressions, are usually preferred.

If $\theta$ is a parameter, both estimators and estimates of $\theta$ can be denoted by $\hat{\theta}$. This notation is practical in a theoretical discussion, or when we do not have a specific notation. For instance, if $\mu$ denotes the mean of a distribution, the usual estimator of $\mu$ is the sample mean, although the sample median can also be used (STAT-11). For the sample mean, we have a especial notation, so we write $\bar{X}$ instead of $\hat{\mu}$.

The sample mean is the preferred estimator of the mean, since its sampling distribution has any desirable property. It is also an example of a **moment estimator**. These estimators are obtained by replacing, in the formula of the moment, the expectation operator E for an average of the corresponding powers of the observations. For instance, the statistic

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

is the moment estimator of the variance. Unfortunately, the expectation of this statistic does not coincide with $\sigma^2$, so (most) statisticians prefer the sample variance (with $n-1$ in the denominator).

## Bias of an estimator

This lecture continues with a brief description of the properties that make an estimator adequate, restricting the detail to the estimation of a single (unidimensional) parameter. We are interested in the properties related to the sampling distribution: mean, variance, normality etc. I start with the **bias**. The bias of an estimator $\hat{\theta}$ of an unknown parameter $\theta$ is the mean deviation with respect to the true value of the parameter,

$$\mathrm{B}\big[\hat{\theta}\big] = \mathrm{E}\big[\hat{\theta} - \theta\big].$$

Taking the deviation $\hat{\theta} - \theta$ as the error of our estimate, the bias has a direct interpretation as an average error. An **unbiased estimator** is one for which the bias is null. For instance, from the preceding lecture, we know that the sample mean is unbiased, and, after correcting the denominator, the sample variance is also unbiased. Moment estimators of skewness and kurtosis are sometimes corrected in a similar way. A factor used to correct the bias is called a **bias correction factor**.

## Standard errors

The **mean square error**, defined as

$$\mathrm{MSE}\big[\hat{\theta}\big] = \mathrm{E}\Big[(\hat{\theta} - \theta)^2\Big],$$

also has a direct interpretation. It can be shown to have two components,

$$\mathrm{MSE}\big[\hat{\theta}\big] = \mathrm{B}\big[\hat{\theta}\big]^2 + \mathrm{var}\big[\hat{\theta}\big].$$

The standard deviation of an estimator is called the **standard error**. We denote it by $\mathrm{se}[\hat{\theta}]$. Among unbiased estimators, the one with the lower standard error is preferred. This is called

**efficiency**. More explicitly, if $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased estimators of $\theta$, we say that $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ when $\mathrm{se}[\hat{\theta}_1] \leq \mathrm{se}[\hat{\theta}_2]$. For instance, both the sample mean and the sample median can be used as estimators of $\mu$ for an $\mathcal{N}(\mu, \sigma^2)$ distribution, but the mean is more efficient. We have found this in the simulation of lecture STAT-11, but a mathematical proof is more difficult. In many cases maximum efficiency is sought among linear estimators. This leads to the concept of **best linear unbiased estimators** (BLUE). "Best" means here minimum variance.

These definitions can be extended to an estimator of a multidimensional parameter without pain, assuming that you are familiarized with matrix and vector formulas. If we take a (multimensional) estimator $\hat{\boldsymbol{\theta}}$ of a parameter vector $\boldsymbol{\theta}$, the bias is a vector and the MSE a matrix, given by

$$\mathrm{MSE}[\hat{\boldsymbol{\theta}}] = \mathrm{E}\Big[\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)^{\mathsf{T}}\Big] = \mathrm{B}\big[\hat{\boldsymbol{\theta}}\big]\mathrm{B}\big[\hat{\boldsymbol{\theta}}\big]^{\mathsf{T}} + \mathrm{cov}\big[\hat{\boldsymbol{\theta}}\big].$$

Also, the variance is replaced by the covariance matrix in the efficiency comparisons: $\hat{\boldsymbol{\theta}}_1$ is more efficient than $\hat{\boldsymbol{\theta}}_2$ when $\mathrm{cov}[\hat{\boldsymbol{\theta}}_2] - \mathrm{cov}[\hat{\boldsymbol{\theta}}_1]$ is positive semidefinite. Although the definitions are so easily extended, handling efficiency becomes a bit involved. I leave this here.

## Consistency

Another approach is based on the convergence of an estimator as the sample size tends to infinity. A common requirement for an estimator is **consistency**. Let me use the notation $\hat{\theta}_n$, to emphasize the dependence of the estimator on the sample size $n$.

We say that the sequence $\hat{\theta}_n$ is consistent when $\operatorname{plim} \hat{\theta}_n = \theta$. Based on the Chebychev inequality, it can be shown that $\lim \mathrm{se}[\hat{\theta}_n] = 0$ implies consistency. Thus, those estimators whose variance have an $n$ in the denominator, as the sample mean and variance, are consistent.

## Normality

Another desirable property is asymptotic normality. A sequence of estimators $\hat{\theta}_n$ is asymptotically normal when the CDF of $(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n])/\mathrm{sd}[\hat{\theta}_n]$ converges to the standard normal CDF as $n \to \infty$. This means, in practice, that certain estimators are taken, for big samples, as if they were normally distributed. For instance, owing to the central limit theorem, the sample mean is asymptotically normal. Also, the maximum likelihood estimation method, which I leave for the Econometrics course, produces asymptotically normal estimators in many situations, making the inference from the estimates much simpler.

## Homework

**A.** The standard deviation is usually estimated by the square root of the sample variance,

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \bar{X}\big)^2},$$

called the sample standard deviation. It is a biased estimator. Positively or negatively?

**B.** The **mean absolute deviation** (MAD), defined as $\mathrm{E}\big[|X - \mu|\big]$, is used sometimes as a measure of dispersion. The sample version,

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n}\big|x_i - \bar{x}\big|$$

is used then as an estimator.

(a) Prove that, in a normal distribution, $\mathrm{E}\big[|X-\mu|\big] = \sqrt{2/\pi}\,\sigma$.

(b) Part (a) suggests using the estimator $\hat{\sigma} = \sqrt{\pi/2}\,\mathrm{MAD}$. Generate 10000 independent samples of size 5 of the standard normal and calculate the corresponding estimates of $\sigma = 1$ based on the mean absolute deviation. Check that this estimator is more biased than the sample standard deviation, but the variance is similar.