[STAT-07] 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 θ, accepting that θ 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 a customer. In certain applications, such as modeling 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 such 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. 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 the parameter μ of 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 θ is a parameter, both estimators and estimates of θ 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 μ denotes the mean of a distribution, the usual estimator of μ is the sample mean, although the sample median can also be used (Example 1 of the preceding lecture). 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

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

is the moment estimator of the variance. Unfortunately, the expectation of this statistic does not coincide with σ^2, as shown in the preceding lecture, 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 θ 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**. Typically, these factors are close to 1 for big sample sizes.

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 θ, we say that $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ when $\mathrm{se}[\hat{\theta}_1] \le \mathrm{se}[\hat{\theta}_2]$. For instance, both the sample mean and the sample median can be used as estimators of μ 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, assuming that you trust your statistical package choosing the most efficient estimator for each job.

Consistency

Another approach to the assessment of an estimator 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.

The t distribution

The inference about the mean of a univariate normal distribution is based on the fact that, if X has a $\mathcal{N}(\mu, \sigma^2)$ distribution, then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

If we don't know σ (this is what happens in practice), we can replace σ by S, getting

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}\,.$$

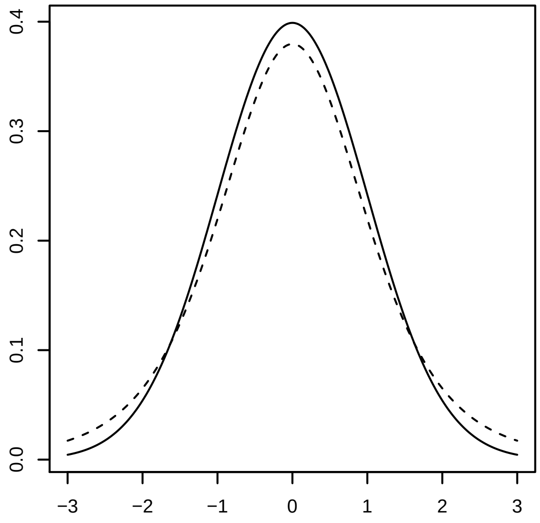

Figure 1. Density curves $\mathcal{N}(0,1)$ and $t(5)$ (dashed line)

The distribution of T is no longer the standard normal, but a different distribution, the **Student t distribution with $n-1$ degrees of freedom**. A Student t is a symmetric distribution, with zero mean and a bell-shaped density curve, similar to the $\mathcal{N}(0,1)$ density (Figure 1). As the χ^2 model, the Student's t is a collection of probability distributions which are specified by the number of degrees of freedom.

The formula for the Student t density with n degrees of freedom, which I denote by $t(n)$, is

$$f(x) = \frac{\Gamma\big((n+1)/2\big)}{(n\pi)^{1/2}\Gamma\big(n/2\big)} \left(1+\frac{x^2}{n}\right)^{(n+1)/2}.$$

The first factor is a normalization constant.

¶ As given here, this formula still makes sense when n is not an integer. Non-integers can be used in certain nonstandard tests.

For an alternative definition, take two independent variables X and Y, and the t Student is obtained as

$$X \sim \mathcal{N}(0,1),\ Y \sim \chi^2(n) \Longrightarrow \frac{X}{\sqrt{Y}} \sim t(n).$$

Because of the symmetry of the t density with respect to zero, the mean and the skewness are null (the skewness converges only for $n>3$). For $n>2$, the variance is $n/(n-2)$ (infinite for $n=1$) and the kurtosis $6/(n-4)$ ($n>4$). This is relevant for a low n, so the Student's t can be used as a model for a distribution which is reasonably bell-shaped but has extra weight at the tails. This trait is exploited in financial analysis.

I denote by t_α, or by $t_\alpha(n)$ if there is ambiguity, the critical values of the Student's t, more specifically, the $(1-\alpha)$-quantile. So, if T has a $t(n)$ distribution, then $\mathrm{p}\big[T > t_\alpha\big] = \alpha$. The Student $t(n)$ converges (in distribution) to the standard normal as $n \to \infty$. This means that, denoting by F_n the CDF of the $t(n)$ distribution, we have

$$\lim_{n\to\infty} F_n(z) = \Phi(z), \qquad \lim_{n\to\infty} t_\alpha(n) = z_\alpha.$$

The practical consequence of this is that, although the Student t is taught as one of the great things of Statistics, it is relevant only for small-sample statistical analysis.

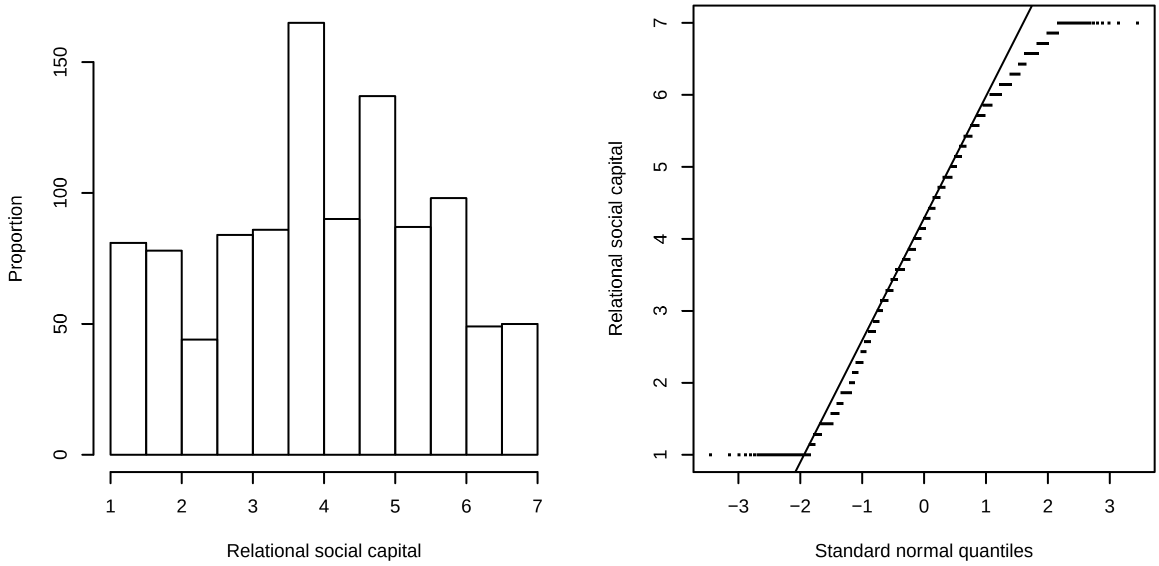

Figure 2. Histogram and normal probability plot (Example 1)

Confidence limits for a mean

Suppose that X has a $\mathcal{N}(\mu, \sigma^2)$ distribution. In the 95% of the cases, we get

$$\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\,.$$

This formula gives limits for μ, called the 95% **confidence limits** for the mean. If X is not normally distributed but n is high (in many cases it suffices with $n > 25$), this formula gives an approximation which, in general, is taken as acceptable. Replacing 1.96 by an adequate critical value z_α, we can switch from the 95% to our probability of choice. Thus, the formula

$$\bar{x} \pm z_\alpha\,\frac{\sigma}{\sqrt{n}}$$

gives the limits for a **confidence level** of $1 - 2\alpha$. If the confidence level is not specified, it is understood that it is 95% ($\alpha = 0.025$). With the confidence limits, we can compare the sample mean $\bar{x}$ to a reference value μ_0. If μ_0 falls out of the limits, we conclude, with the corresponding confidence level, that $\mu \neq \mu_0$. We say then that the difference $\bar{x} - \mu_0$ is **significant**.

With real data, σ is unknown, but, for a big n, it can be replaced by s, obtaining an approximate formula for the confidence limits of the mean. Nevertheless, there is an exact formula, appropriate for a small n, in which z_α is replaced by $t_\alpha(n-1)$. The formula is then

$$\bar{x} \pm t_\alpha(n-1)\,\frac{s}{\sqrt{n}}\,.$$

The difference between these two formulas becomes irrelevant for a big sample. If the normality assumption is not valid, the formula of the confidence limits is still approximately valid for big samples, by virtue of the central limit theorem. In such case, using either z_α or t_α does not matter, since they will be close.

Example 1. Using data collected from 1,817 individuals in 36 business units of 7 multinational firms, a recent study examined the relationships both among the structural, relational and cognitive dimensions of **internal social capital** and between these dimensions and their antecedents. A popular approach decomposes internal social capital into three dimensions: structural, relational and cognitive. The `scapital` data set contains data on these three dimensions, based on 1–7 **Likert scales** with 3, 7 and 4 items, and a dummy for being female.

I average the seven items of relational social capital, to get a unique measure, calculating the 95% confidence limits for the mean in the female group. We have

$$n = 1,049, \qquad \bar{x} = 3.994, \qquad s = 1.552.$$

Based on $t_{0.025}(1048) = 1.962$, we get 3.994 ± 0.094. Of course, for such a sample size, using the t or the $\mathcal{N}(0,1)$ critical value does not matter. Also, we can leave aside the concern about the normality of the distribution, although the diagnostic plots of Figure 2 show that normality is questionable here. For the male group, the limits are 4.325 ± 0.111. So, the two intervals do not overlap, suggesting that there is a real difference between male and female employees on this dimension.

¶ Source: D Pastoriza, MA Ariño, JE Ricart & MA Canela (2015), Does an ethical work context generate internal social capital?, *Journal of Business Ethics* **129**, 77–92.

Homework

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

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

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,

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

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.

C. Give an asymptotic formula for the 95% confidence limits of the mean of a Poisson distribution.