[STAT-11] Sampling distributions

Miguel-Angel Canela
Associate Professor, IESE Business School

Sample mean

Let $X_1, \ldots, X_n$ be a (statistical) sample from a distribution, that is, a collection of independent variables following that distribution. A function $G = g(X_1, \ldots, X_n)$ is called a **statistic** (singular). The obvious example is the **sample mean**, defined as

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

Note that, here, $\bar{X}$ is not a number, but a random variable, with a probability distribution. To distinguish this distribution from the distribution from which we draw the sample, we call it a **sampling distribution**. In general, statistics are interesting when their sampling distributions have nice properties. I'll be more specific about this later.

In the case of the sample mean, we note, that

$$\mathrm{E}\big[\bar{X}\big] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \mu, \qquad \mathrm{var}\big[\bar{X}\big] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}[X_i] = \frac{\sigma^2}{n}\,.$$

The first formula means that, on the average, the sample mean is right as an approximation of the population mean μ. Also, since the variance is a measure of the variation about the expectation, the second formula tells us that the approximation improves as the sample size increases.

Note that the independence of the observations has been used here but not in the preceding argument. When sampling from a $\mathcal{N}(\mu, \sigma^2)$ distribution, we know something more on the sampling distribution of the mean, that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. If the distribution is not normal, we still have the central limit theorem, discussed in the next lecture.

Limit theorems

From the expression of the variance, we see that it tends to zero as $n \to \infty$. This means that the variation of $\bar{X}$ becomes very smal for big samples. We say that $\bar{X}$ converges to μ as $n \to \infty$. This statement, called the **law of large numbers**, is one of the great theorems of Mathematical Statistics.

Although the idea of the law of large numbers is clear enough, a comment on **limit theorems** is worth here. In Statistics textbooks, a lot effort is put on explaining the distinctions among the different types of convergence. Why? First, because the proofs of limit theorems can be more or less difficult depending on that. Second, because, although the definition of the limit of a sequence of numbers has nothing to hide, because numbers are simple things, but a random variable carries a lot on its back. What does converge, the values of the variables, the densities, parameters like means and variances?

There are different types of convergence, and developing them here will take more space than allowed. So, this discussion is quite short. The law of large numbers is easily proved if we formulate it in terms of **convergence in probability**. It can be stated as: for every number $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{p}\big[|\bar{X} - \mu| > \epsilon\big] = 0.$$

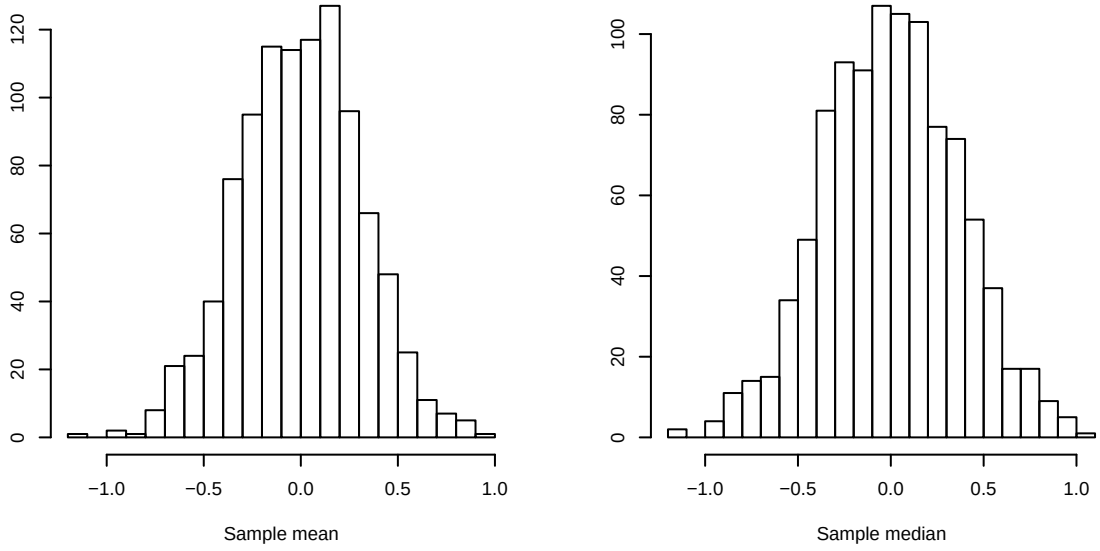

Figure 1. Sampling distributions of the mean and median

This is expressed, in short, as $\text{plim}\bar{X} = \mu$. The proof is based on the **Chebyshev inequality**,

$$\text{p}\big[|X - \mu| > \epsilon\big] \leq \frac{\sigma^2}{\epsilon^2}\,,$$

which is valid for any probability distribution with moments of first and second order, and not hard to prove.

Example 1. The properties of estimators can be understood by simulating their sampling distributions. Let me sample from the $\mathcal{N}(0,1)$ distribution, with size $n = 10$. I generate 1000 samples, saving the means, medians and variances in a data set that contains 1,000 observations of each of these three statistics.

The histograms of the sample mean and the sample median are shown in Figure 1. The means are –0.0039 and 0.0012, respectively, close to the zero population mean. The standard deviation of the sample mean is 0.3112, close to the theoretical value $1000^{-1/2} = 0.3162$. The standard deviation of the sample median is 0.3652, a bit higher. This agrees with the theory, and supports the preference for the mean.

Chi square distribution

Let $Z_1, \ldots, Z_n$ be independent $\mathcal{N}(0,1)$ variables. We say then that the variable $X = Z_1^2 + \cdots + Z_n^2$ has a **chi square distribution with n degrees of freedom**, in short $X \sim \chi^2(n)$. The χ^2 distribution is used in many tests in practical statistical analysis, specially with maximum likelihood estimation.

It follows directly from the properties of the expectation that the mean and the variance of the $\chi^2(n)$ distribution are n and $2n$, respectively. We see in Figure 1 three χ^2 density curves. The formula of the PDF is not a friendly one,

$$f(x) = \frac{x^{n/2-1}\, e^{-x/2}}{2^{n/2}\, \Gamma(n/2)}\,, \quad x > 0,$$

but the calculations can be easily managed in the computer. The denominator comes from the normalization condition. The notation $\chi^2_\alpha(n)$ is consistent with z_α.

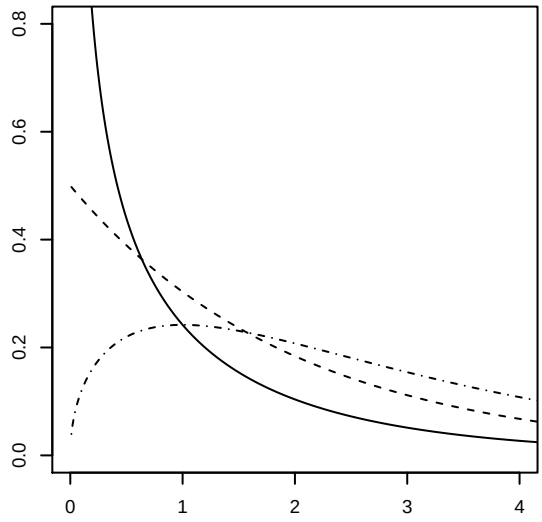

Figure 2. $\chi^2(1)$, $\chi^2(2)$ **and** $\chi^2(3)$ **density curves**

Γ denotes the Gamma function, ubiquitous in probability formulas. It is a positive, increasing function on $(0, +\infty)$, satisfying $\Gamma(n+1) = n!$ for any integer n. Although it is defined by an (almost) unmanageable integral formula, as the standard normal CDF, it is available in the computer. Note that, since $\Gamma(1) = 1$, the case $\chi^2(2)$ is the exponential distribution.

Sample variance

The **sample variance** is defined as

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

The expectation of the sampling distribution is σ^2. Indeed, since all the terms of the sum on the right side of this formula have the same expectation (I assume here $\mu = 0$, to shorten the equations),

$$\begin{aligned} \mathrm{E}\big[S^2\big] &= \frac{n}{n-1}\,\mathrm{E}\big[\big(X_1 - \bar{X}\big)^2\big] = \frac{n}{n-1}\Big(\mathrm{E}\big[X_1^2\big] + \mathrm{E}\big[\bar{X}^2\big] - 2\,\mathrm{E}\big[X_1\bar{X}\big]\Big) \\ &= \frac{n}{n-1}\left(\sigma^2 + \frac{\sigma^2}{n} - \frac{2}{n}\sum_{i=1}^{n}\mathrm{E}\big[X_1 X_i\big]\right) = \frac{n}{n-1}\left(\sigma^2 + \frac{\sigma^2}{n} - 2\,\frac{\sigma^2}{n}\right) = \sigma^2. \end{aligned}$$

This explains why the definition of S^2 with $n-1$ in the denominator is favoured by most statisticians. The formula for $\mathrm{var}[S^2]$ is more complex, involving the kurtosis. Nevertheless, under normality, the distribution of the sample variance can be related to the χ^2 model (the proof involves some matrix algebra). More specifically,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

In particular, the variance is

$$\mathrm{var}\big[S^2\big] = \frac{2\,\sigma^4}{n-1}\,.$$

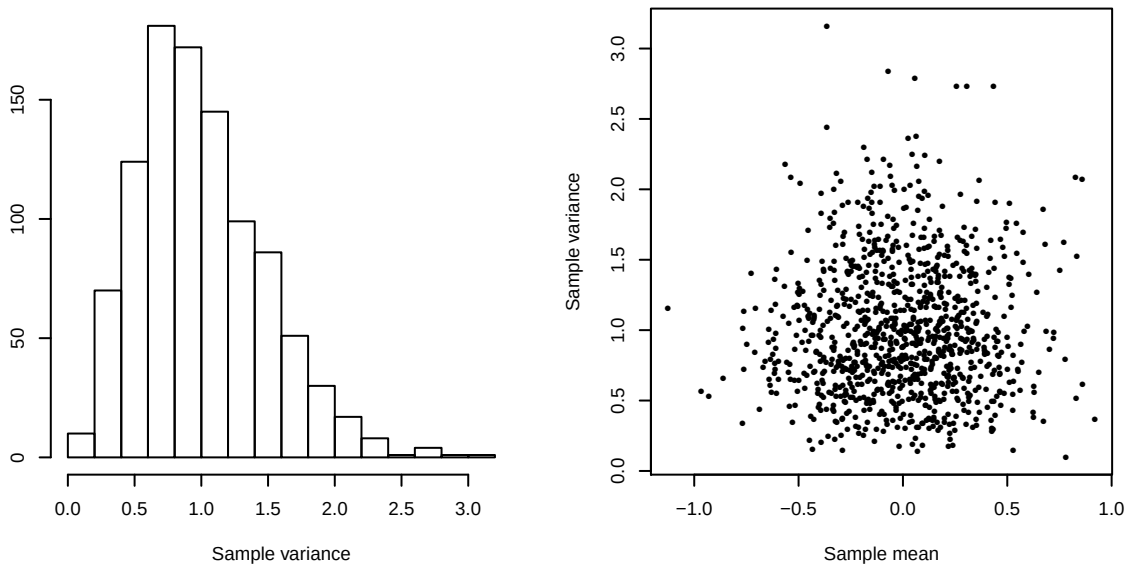

Figure 3. Distribution of the sample variance

Under a normality assumption, the mean and sample variance are independent. I omit the details, but you can find support for this assertion in the simulation below.

Example 1 (continuation). The left panel of Figure 3 is a histogram of the sample variance, consistent with the expected χ^2 profile. The mean is 1.0021 and the standard deviation 0.4732. Note that, according to the theory, the sample variance is one ninth of an observation of a $\chi^2(9)$. The correlation between the sample mean and variance is 0.0167, also in agreement with the theory. This is illustrated by the right panel of the figure. □

Sample covariance and correlation

We can also consider the **sample covariance**,

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} \big(X_i - \bar{X}\big)\big(Y_i - \bar{Y}\big),$$

and the **sample correlation**

$$R = \frac{S_{XY}}{S_X\, S_Y} = \frac{\sum \big(X_i - \bar{X}\big)\big(Y_i - \bar{Y}\big)}{\Big[\sum \big(X_i - \bar{X}\big)^2 \sum \big(Y_i - \bar{Y}\big)^2\Big]^{1/2}}\,.$$

The covariance is rarely tested, but the correlation is a key statistic. The sampling distribution is a bit more difficult. For the moment being, we will be satisfied with a simulation (exercise A).

Homework

A. Simulate the sampling distribution of the correlation for two independent exponential distributions with $\lambda = 1$.