[STAT-14] Confidence limits for the mean

Miguel-Angel Canela
Associate Professor, IESE Business School

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

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

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 is taught as one of the great things of Statistics, it is relevant only for small-sample statistical analysis.

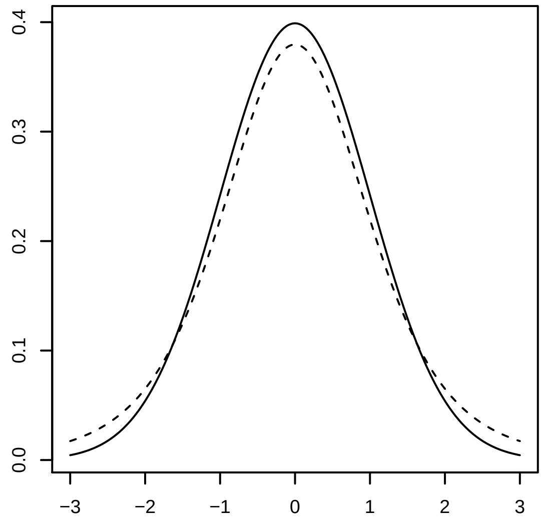

Figure 1. Density curves $\mathcal{N}(0,1)$ and $t(5)$ (dashed line)

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** $1-2\alpha$. 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

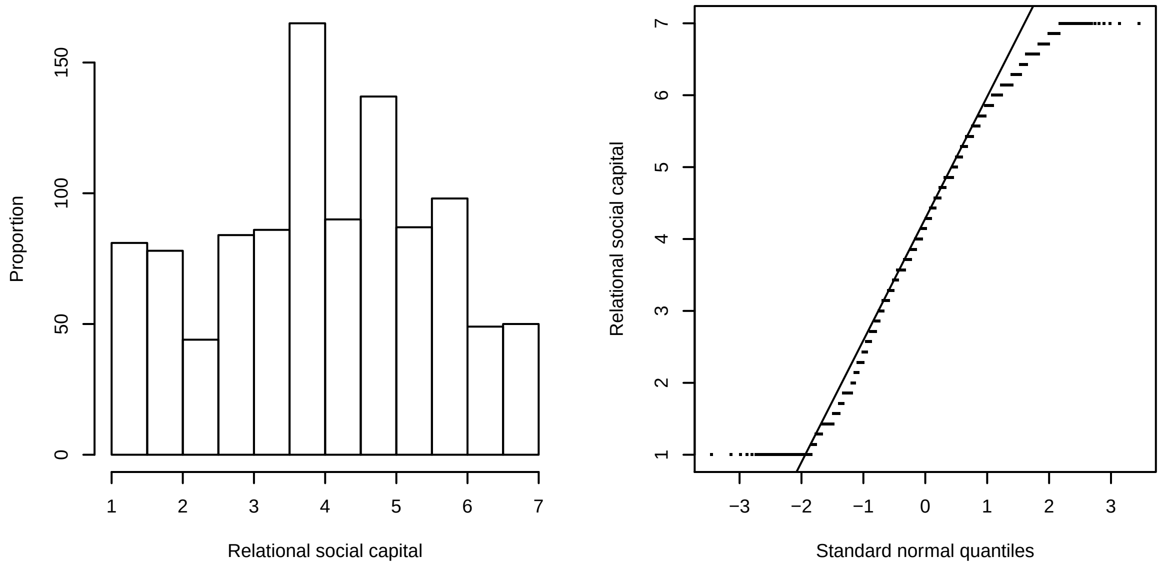

Figure 2. Histogram and normal probability plot (Example 1)

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