Standard Error of the Sampling Distribution

Rémi Viné

2022

1 Statement

The standard error is derived from the sampling distribution. It estimates the standard deviation of this *theoretical* distribution. For σ the population standard deviation and s the sample standard deviation, denoting $\sigma_{\bar{x}}$ the standard error,and with n the sample size, we have

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

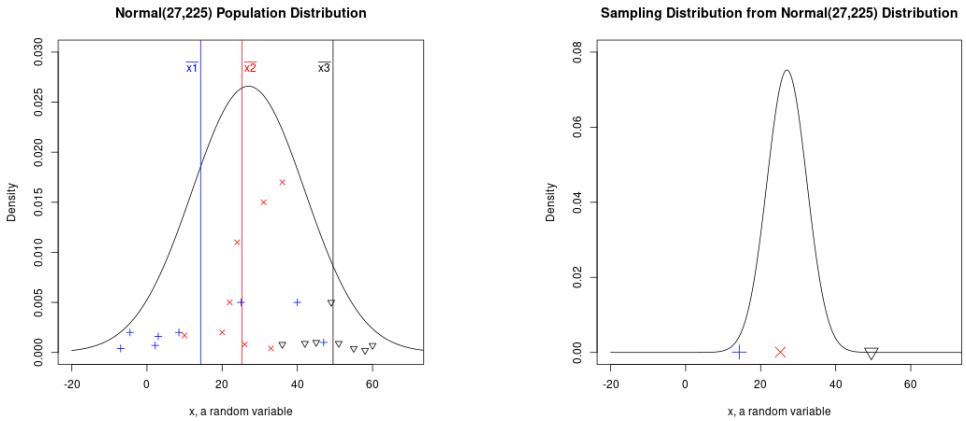
which is often estimated using

$$s_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

At the limit, if one constructed an incomplete sampling distribution using only samples of 1 unit, then the standard error would equal the population standard deviation as $n = 1$. On the contrary, for $n \to N$ (with N the population size), the standard error shrinks and would even equate zero for $n = N$ as the sample is identical to the population and there is therefore no *error* due to sampling.

1.1 Illustration of Sampling Distribution's construction

Assume there is a population of a given random variable that follows a Normal distribution with parameters $\mu = 27$ and $\sigma^2 = 225$. Since the population distribution is Normal, so is the sampling distribution, whatever the sample size. Here, samples of size $n = 8$ are taken and only three are depicted. The sample composed of red crosses is centered and does not have observation outside the $\pm 1\sigma$ interval. On the contrary, the sample composed of blue crosses observations is more scattered and its sample average is slightly lower than the population average. Last, the sample composed of black triangles is only composed of samples located among the highest values of the random variable. All samples are possible in the population but with different probabilities. Clearly, the dark triangle sample is less likely to happen, as the sampling distribution shows. More precisely, because the mean of the dark triangles sample (composed of 8 elements) is 49.5 units, and that the sampling distribution can be considered as being a Normal distribution, it is easy to find the probability of a sample to have a sample average of at least 49.5 units: $P(\bar{x} \geq 49.5) = P(z \geq \frac{49.5-27}{\frac{15}{\sqrt{8}}}) = P(\bar{x} \geq 4.243) = 0.00001$.[1] Clearly, the occurrence of such set is rare!



[1] On the other hand, if one wanted to find the probability, in the population to observe a datapoint equivalent to 49.5 units, one would obtain $P(x \geq 49.5) = P(z \geq \frac{49.5-27}{15} = P(z \geq 1.5) = 0.0668$.

1.2 Bienaymé formula

This formula is useful for the present proof along with many applications in statistics that involve the variance. For X_i an uncorrelated sequence of random variables, Bienaymé formula states that

$$Var\left(\sum_i X_i\right) = \sum_i Var(X_i)$$

If X_i random variables are correlated, then Bienaymé formula is

$$Var\left(\sum_i X_i\right) = \sum_i Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j) = \sum_{i,j} Cov(X_i, X_j)$$

2 Proof

Assume we have n samples with sample average $\bar{x}$ and standard deviation σ. The variance of the sampling distribution is

$$Var(x_1 + x_2 + x_3 + ... + x_n) = Var\left(\sum_i x_i\right)$$

These samples are independent, so that Bienaymé formula (see subsection 1.2) can be directly applied.

$$Var(x_1 + x_2 + x_3 + ... + x_n) = \sum_i Var(x_i)$$

Since x_i are derived from the population distribution whose standard deviation is σ, it is as if all sample standard deviation, on average could be approximated by σ (see the Law of Large Number for example in wikipedia). Hence

$$Var(x_1 + x_2 + x_3 + ... + x_n) = n\sigma^2$$

$$Var(\frac{1}{n}\sum_i x_i) = Var(\bar{x}) = n\frac{\sigma^2}{n^2}$$

Hence,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

□