Annex 1 - Assessment of Sampling Error

Quantification of the sampling error is based on the continuity-corrected version of the Clopper-Pearson confidence interval [1] on the binomial proportion, [39]. This confidence interval is referred to as a sampling error, and details of its implementation are outlined below.

The probability $C = (1-\alpha)$ for occurrence of proportion $\hat{p} = \frac{x}{n}$ (occurrence of x instances of interest in n number of trials) can be assigned based on an inverse of the regularised incomplete Beta function given by (1).

$$I_\beta(a,b,x) = \frac{\int_0^x t^{a-1} \cdot (1-t)^{b-1} \cdot dt}{\int_0^1 t^{a-1} \cdot (1-t)^{b-1} \cdot dt} \quad (1)$$

Namely, given the desired probability C, an interval that contains the "true" binomial proportion $\hat{p}$ can be found as $\{\hat{p}_{lo}; \hat{p}_{up}\}$, where $\hat{p}_{lo} = \frac{x_{lo}}{n}$ and $\hat{p}_{up} = \frac{x_{up}}{n}$, and where the number of occurrences x_{lo} and x_{up} derive from the inverse solutions to equations (2) and (3), respectively:

$$I_\beta\left(n - x_{lo} + \frac{1}{2},\quad x_{lo} + \frac{1}{2};\quad 1 - \hat{p}\right) = \frac{\alpha}{2} \quad (2)$$

$$I_\beta\left(n - x_{up} + \frac{1}{2},\quad x_{up} + \frac{1}{2};\quad 1 - \hat{p}\right) = 1 - \frac{\alpha}{2} \quad (3)$$

The solutions can be denoted as $x = I_\beta^{-1}$.

In case of cumulative probability function for random variable X, the above proportions refer to the maximum number of occurrences of random variable X up to the specific value x.

[1] Also referred to as an equal-tailed Bayesian interval or Jeffrey's prior interval.

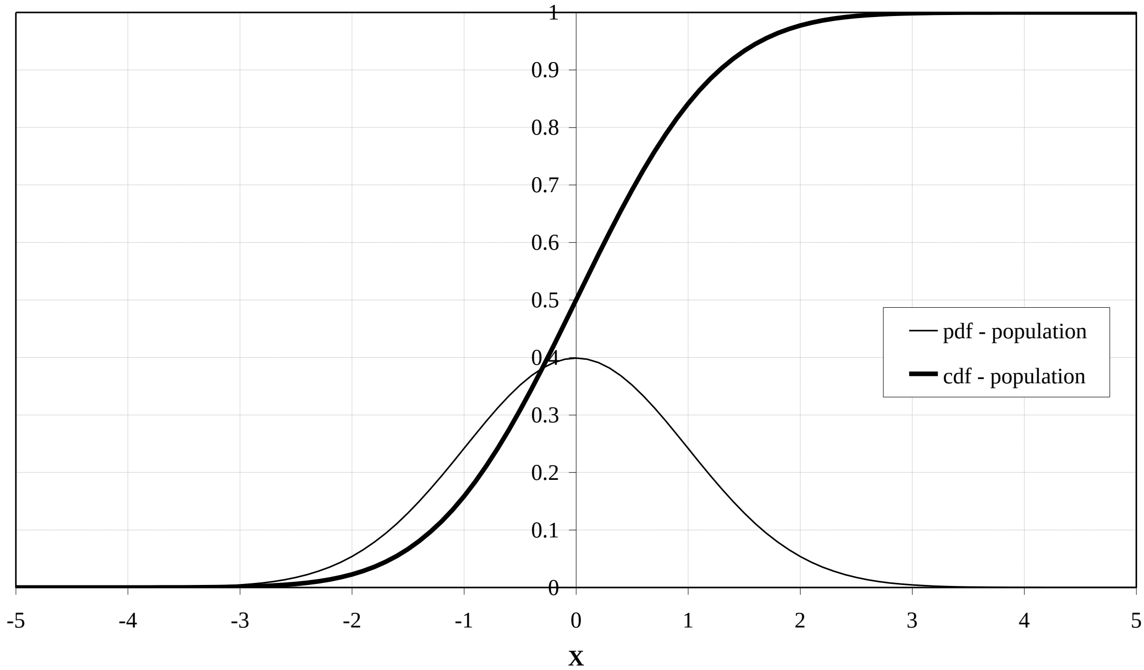

Figure 1 - Example of probability distribution for the population of random variable X, where X~N(0,1)

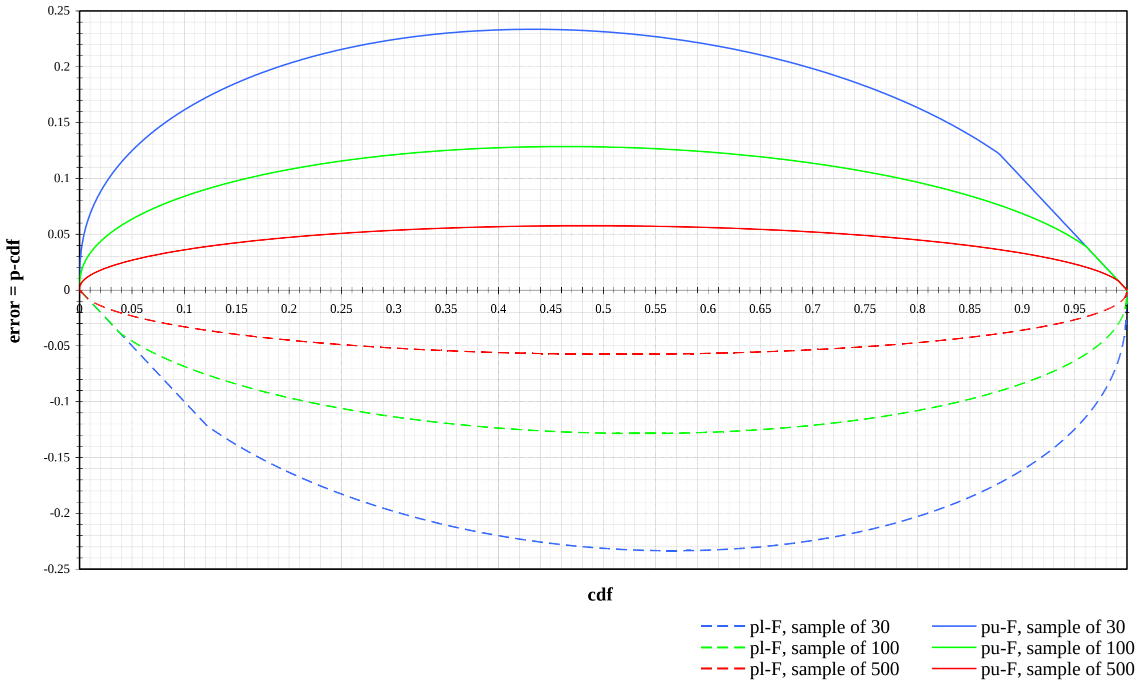

Figure 2 - Continuity-corrected Clopper-Pearson 99% confidence interval (sampling error) on binomial proportion for n = 30, n = 100 and n = 500

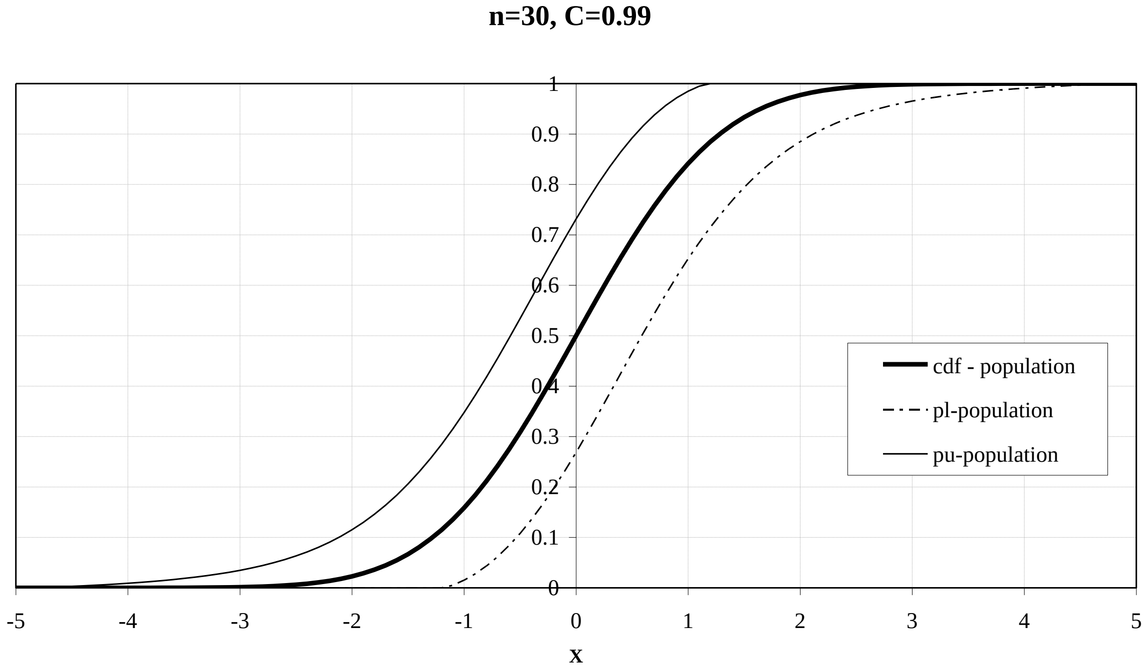

Figure 3 - 99% sampling error on the cdf for the population of random variable X; sample size n = 30

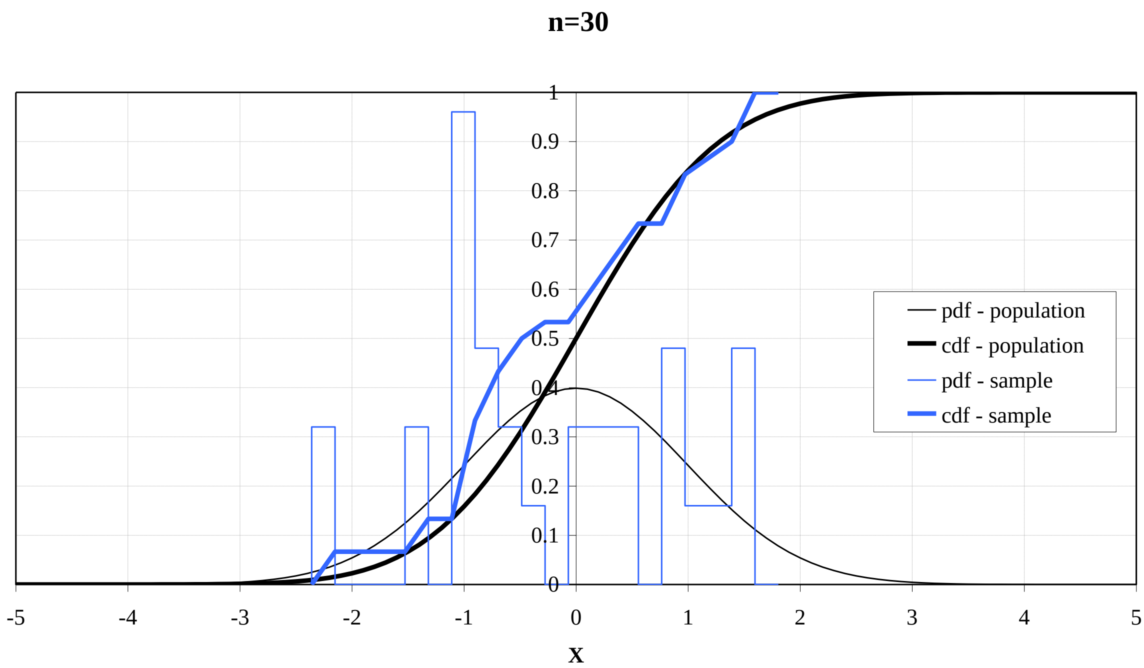

Figure 4 - Example of probability distribution for a randomly generated sample of variable X; sample size n = 30

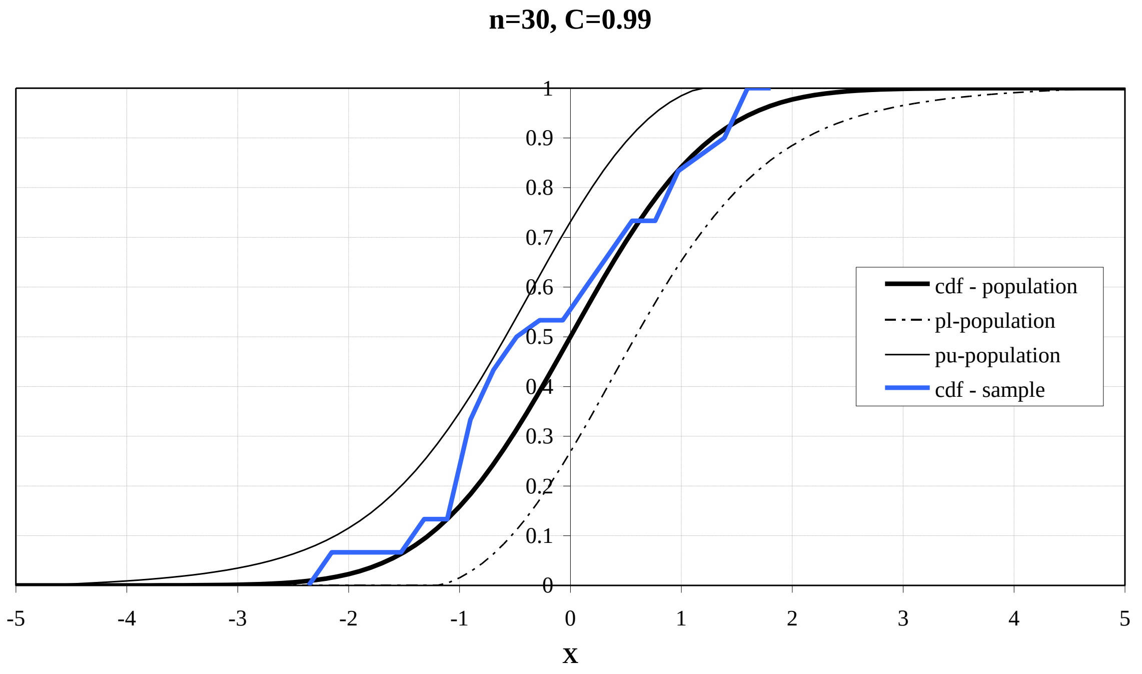

Figure 5 - Assigned cdf on the basis of the sample of 30 will be contained within the 0.5% quintiles around the cdf

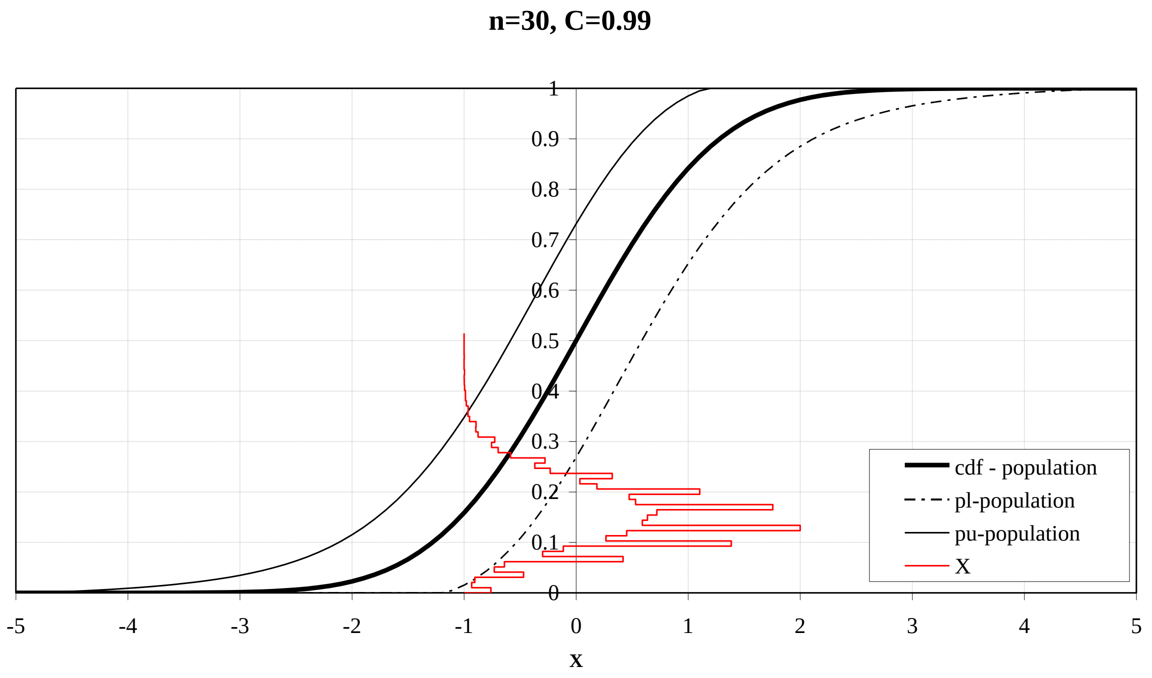

Figure 6 - Distribution of probability density for the binomial proportion $\frac{I_\beta^{-1}(x,\hat{p}=0.15)}{n}$**, Monte-Carlo sampling**

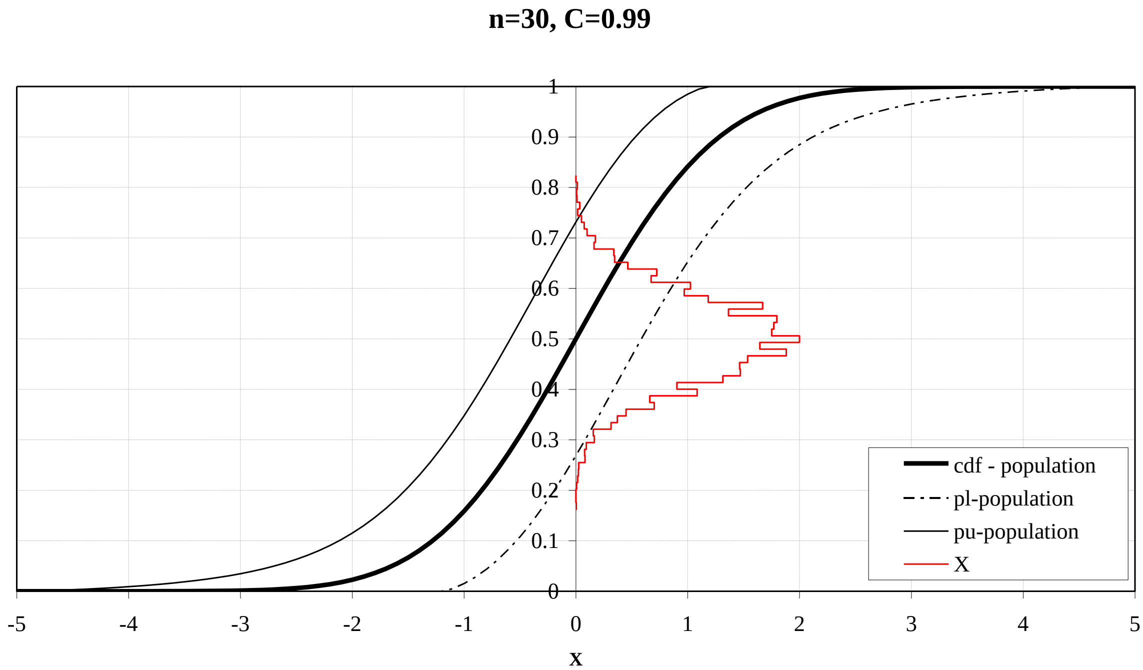

Figure 7 - Distribution of probability density for the binomial proportion $\dfrac{I_{\beta}^{-1}\left(x, \hat{p}=0.5\right)}{n}$**, Monte-Carlo sampling**

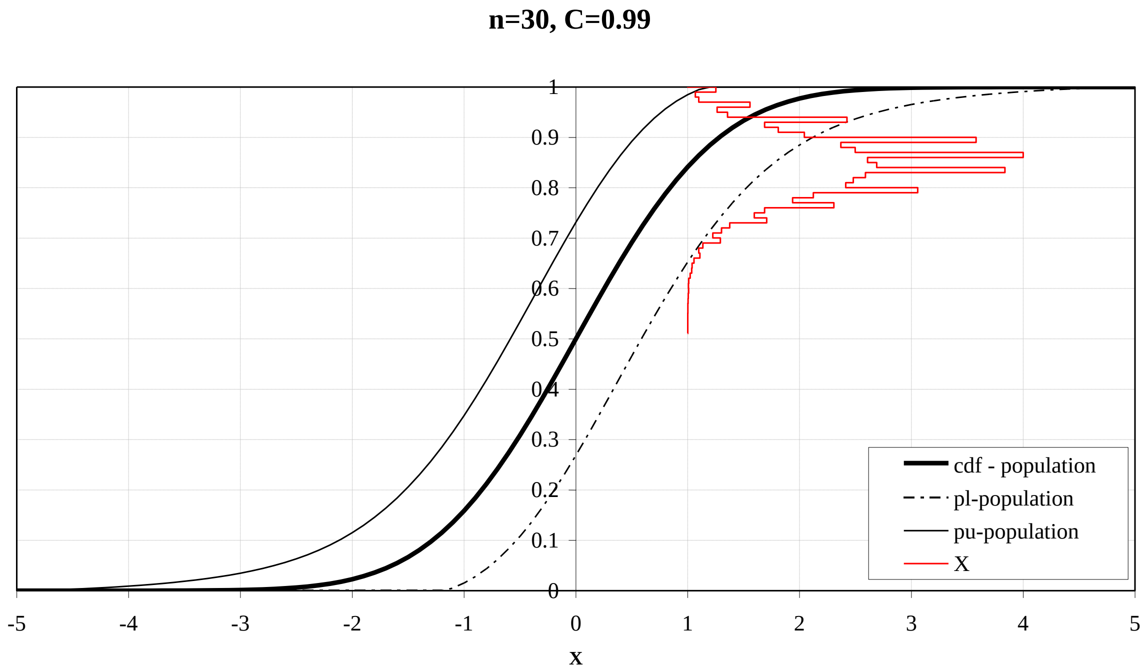

Figure 8 - Distribution of probability density for the binomial proportion $\dfrac{I_{\beta}^{-1}\left(x, \hat{p}=0.85\right)}{n}$**, Monte-Carlo sampling**

n=30, C=0.99, 100,000 Monte Carlo (MC) trials

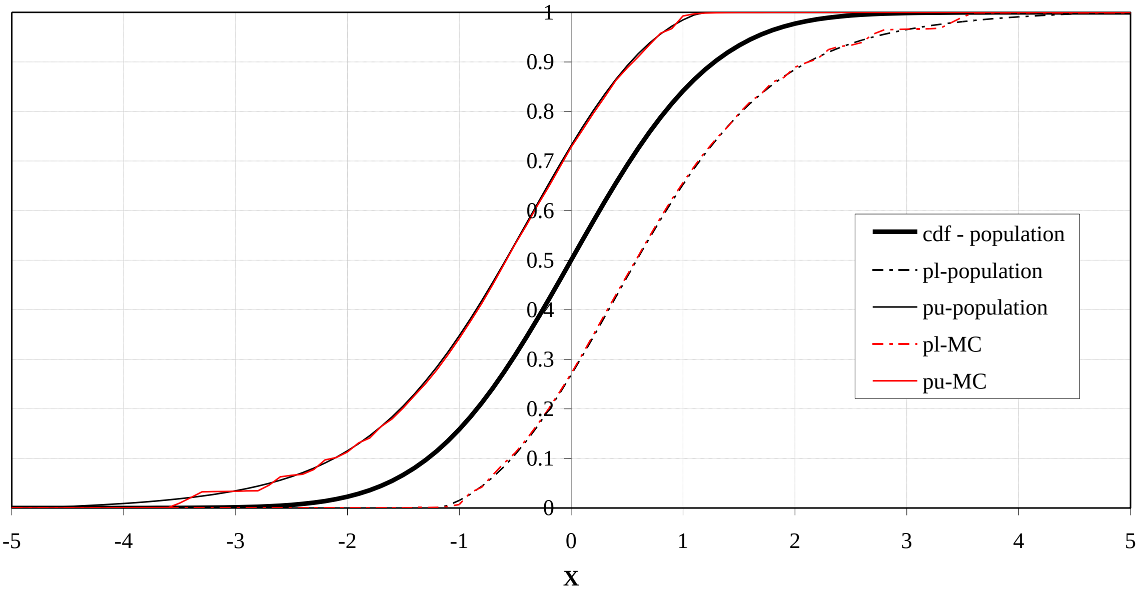

Figure 9 - n=30, c=0.99, 100000, Monte Carlo trials

A Monte-Carlo experiment confirms that only in about 1,000 occasions out of 100,000 samples of 30 elements drawn randomly from population N(0,1), would the cdf for any of the samples be beyond the 0.5% quintiles off any value of the cdf for the population. Sample size n = 30.

n=30, C=0.99

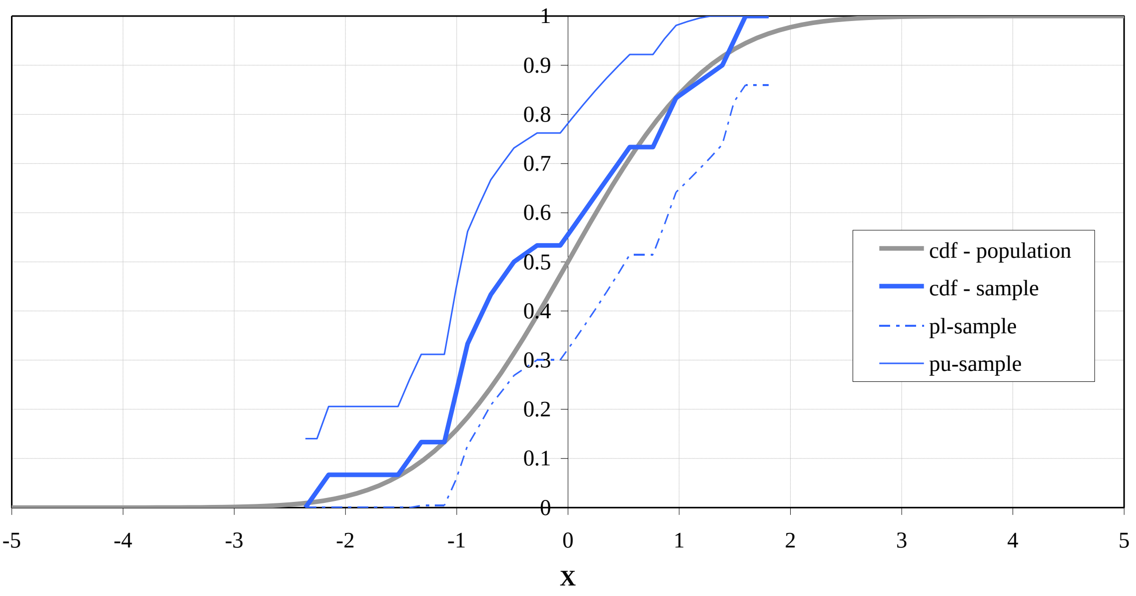

Figure 10 - Sampling error, n = 30, c = 0.99

Figure 10 shows that since the cdf of a population is never known, the sampling error allows deriving the interval around the sample cdf within which the population cdf can be expected with C probability; sample size n = 30.

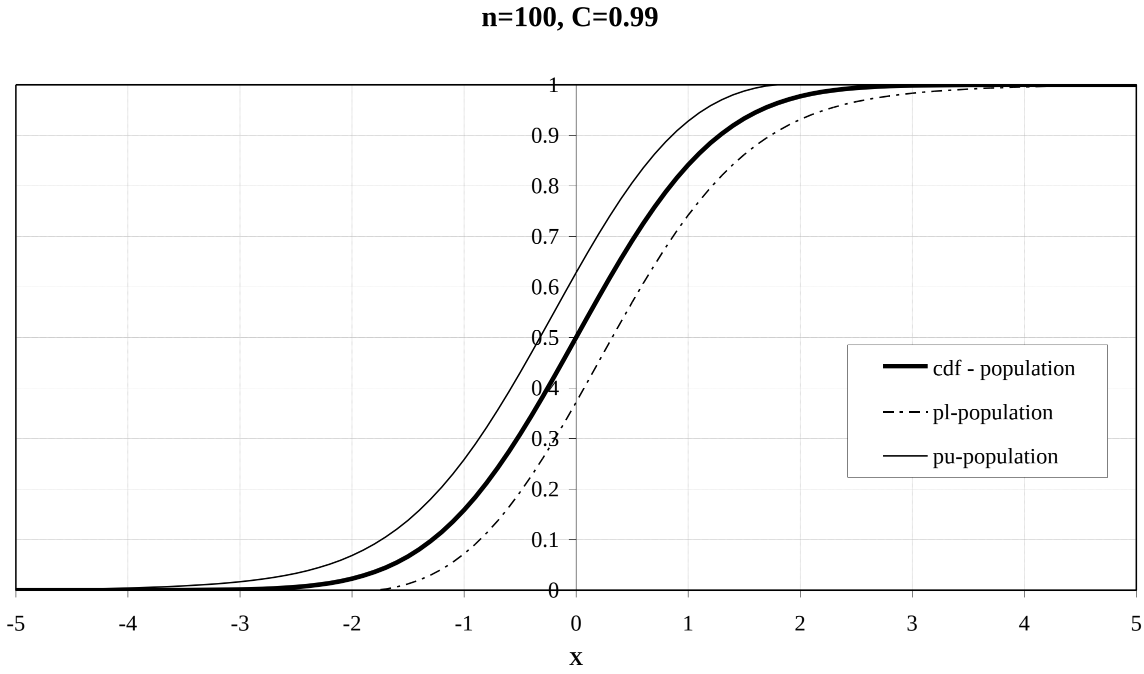

Figure 11 - 99% sampling error on the cdf for the population of random variable X; sample size n = 100

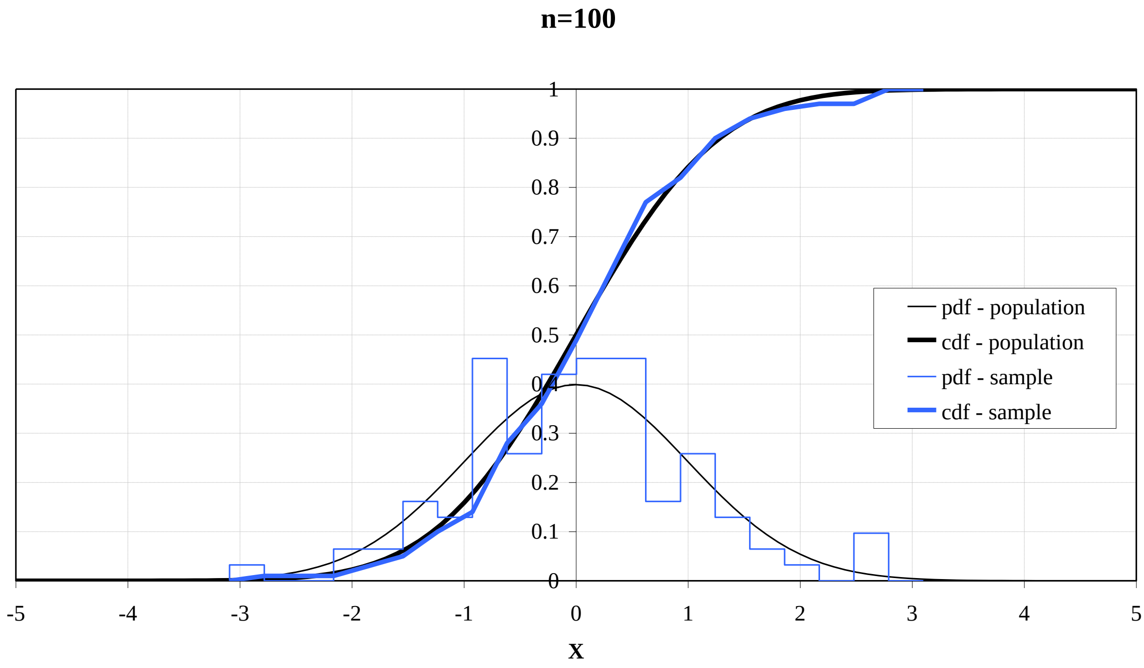

Figure 12 - Example of probability distribution for a randomly generated sample of variable X; sample size n = 100

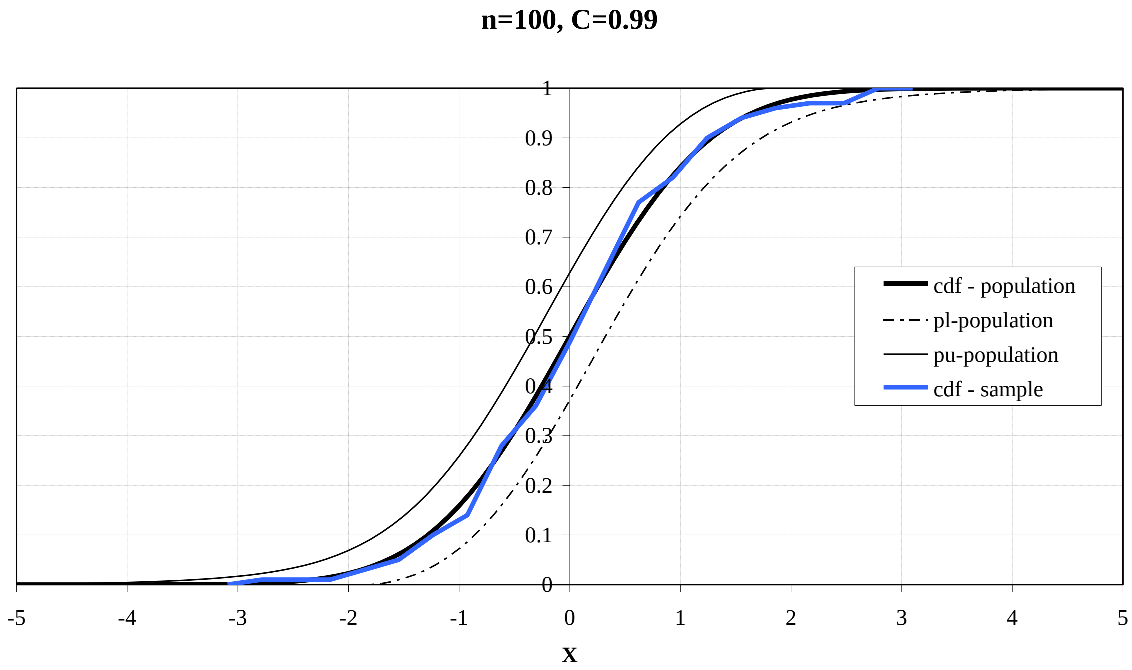

Figure 13 - Assigned cdf on the basis of the sample of 100 will be contained within the 0.5% quintiles around the cdf

n=100, C=0.99, 100,000 Monte Carlo (MC) trials

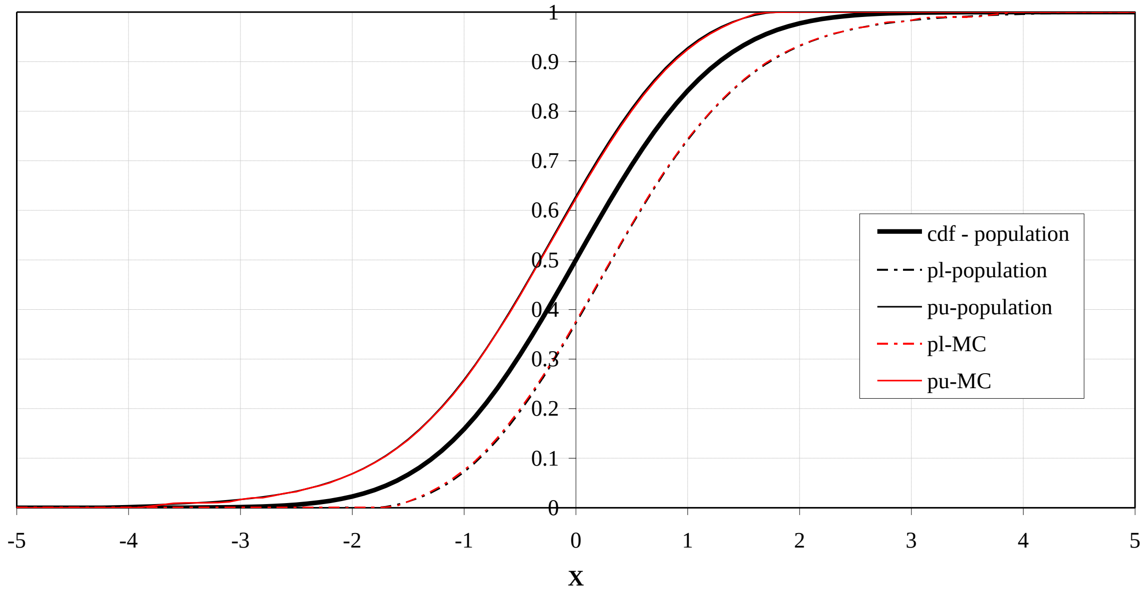

Figure 14 - n=100, c=0.99, 100000, Monte Carlo trials

Figure 14 shows that a Monte-Carlo experiment confirms that only in about 1,000 occasions out of 100,000 samples of 100 elements drawn randomly from population N(0,1), would the cdf for any of the samples be beyond the 0.5% quintiles off any value of the cdf for the population. The sample size is n=100.

n=100, C=0.99

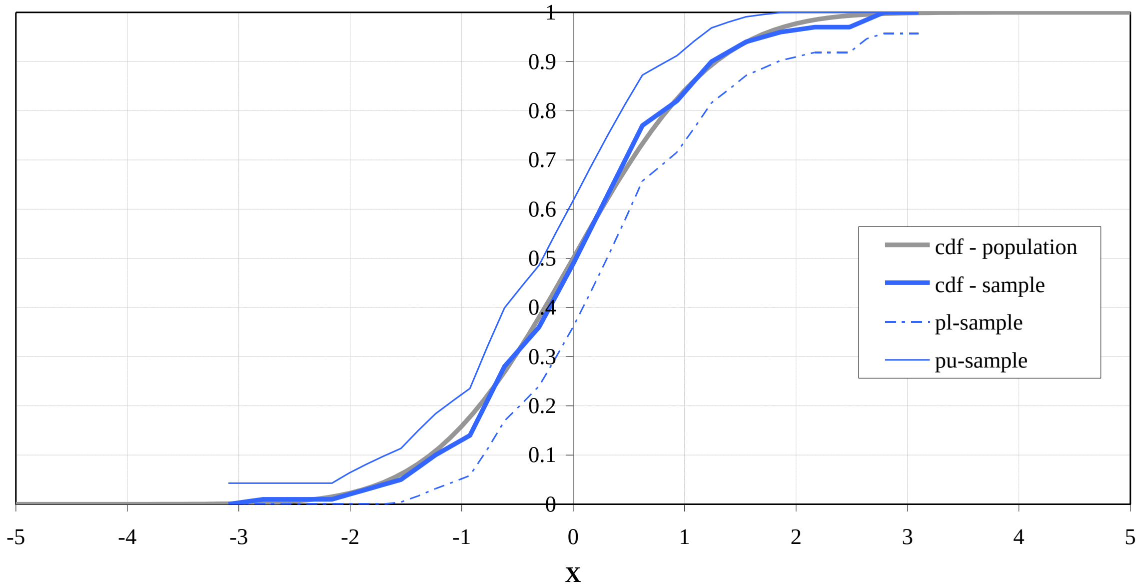

Figure 15 - Sampling error, n=100, c=0.99

Figure 15 shows that since the cdf of a population is never known, the sampling error allows deriving the interval around the sample cdf within which the population cdf can be expected with C probability. The sample size is n=100.

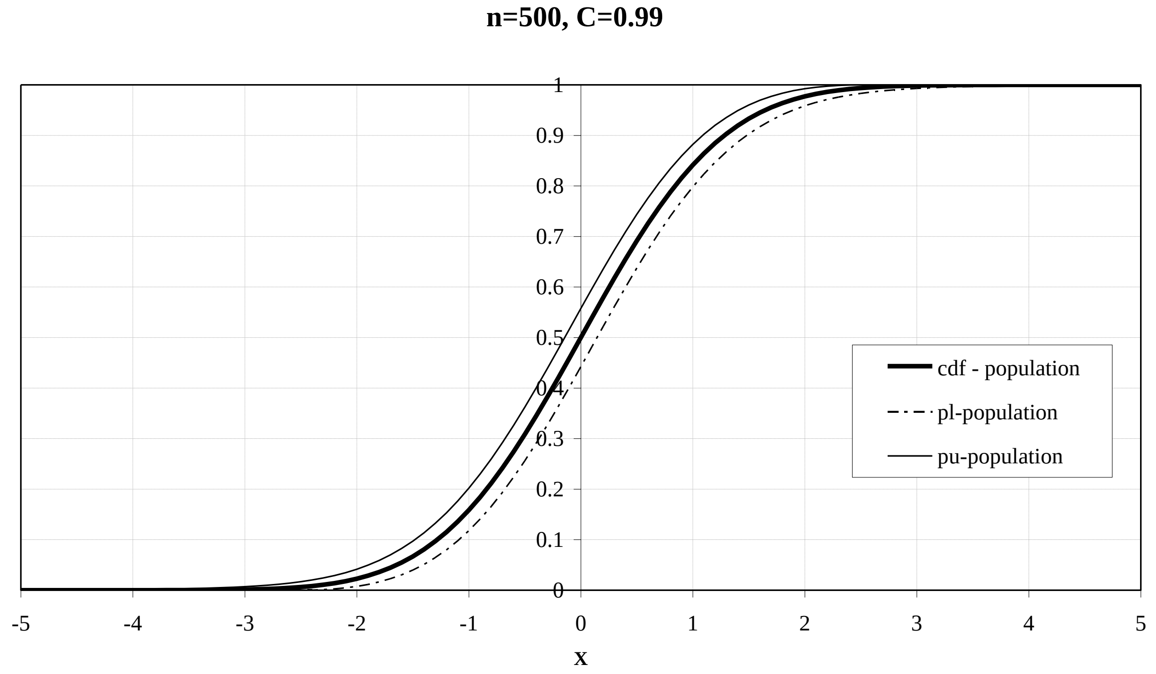

Figure 16 - 99% sampling error on the cdf for the population of random variable X; sample size n=500

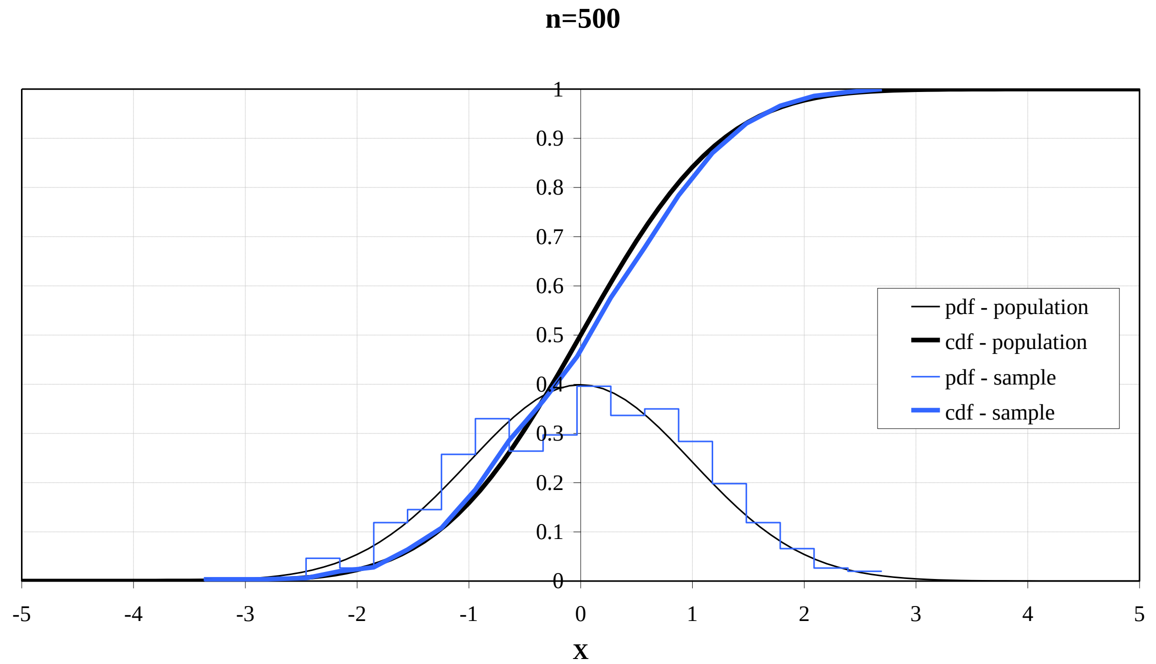

Figure 17 - Example of probability distribution for a randomly generated sample of variable X; sample size n=500

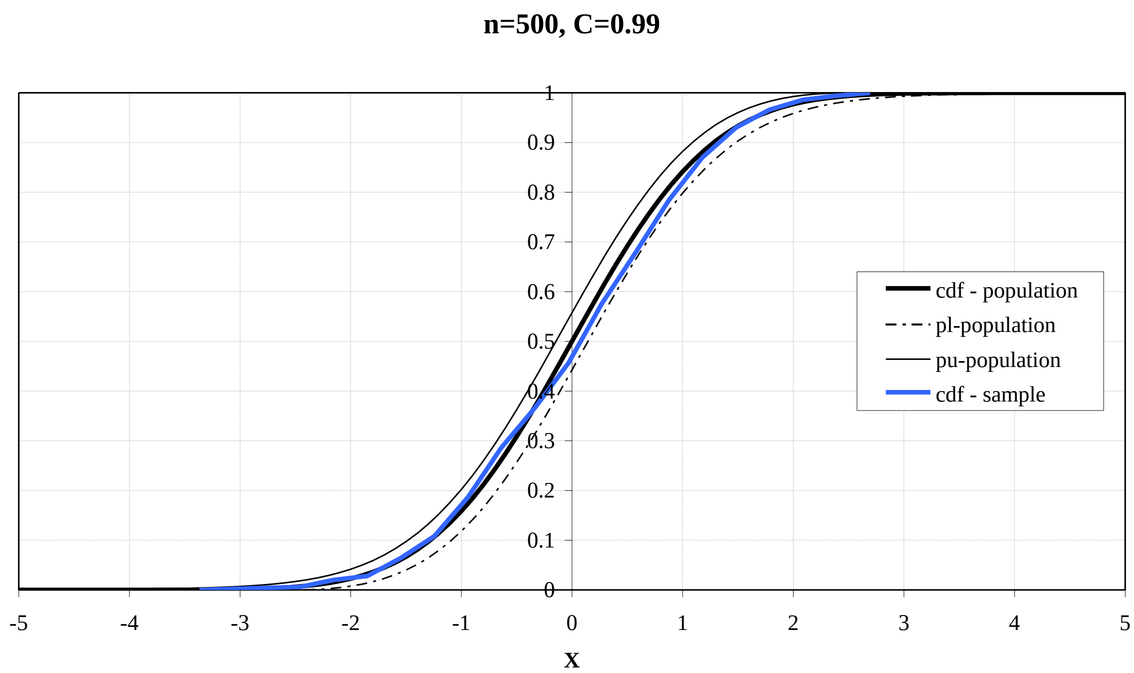

Figure 18 - Cdf population, n=500, c=0.99

Figure 18 shows that the cdf assigned based on the sample of 500 will be contained within the 0.5% quintiles around the cdf for the population, if it is known.

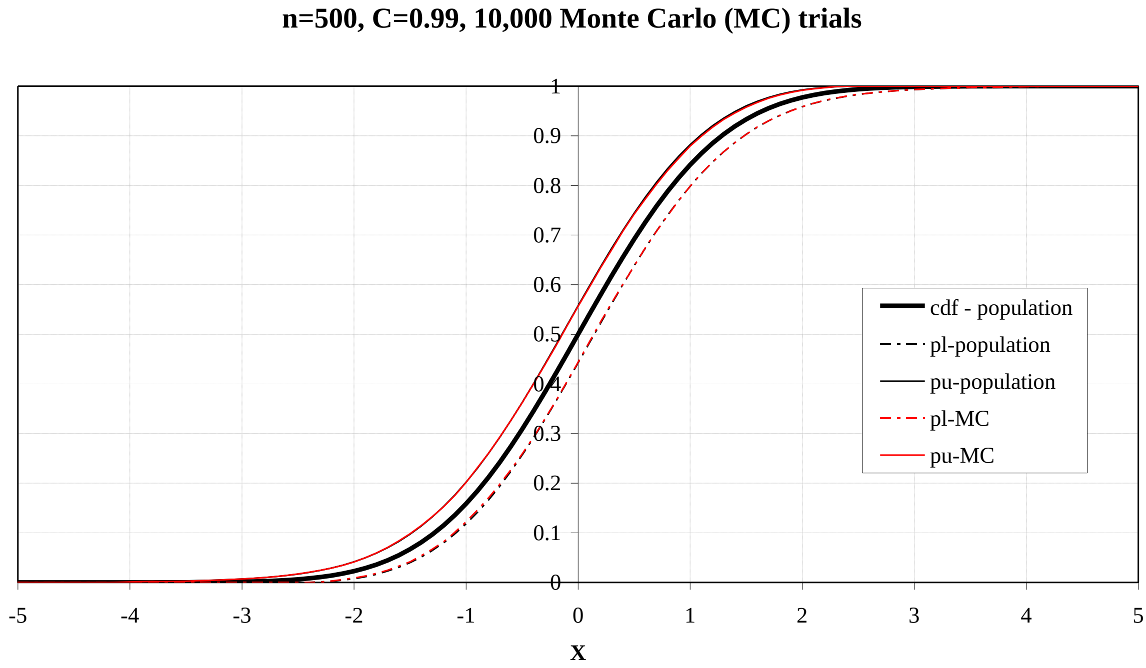

Figure 19 - n=500, c=0.99, 100000, Monte Carlo trials

Figure 19 shows that a Monte-Carlo experiment confirms that only in about 1,000 occasions out of 100,000 samples of 500 elements drawn randomly from population N(0,1), would the cdf for any of the samples be beyond the 0.5% quintiles off any value of the cdf for the population. The sample size is n=500.

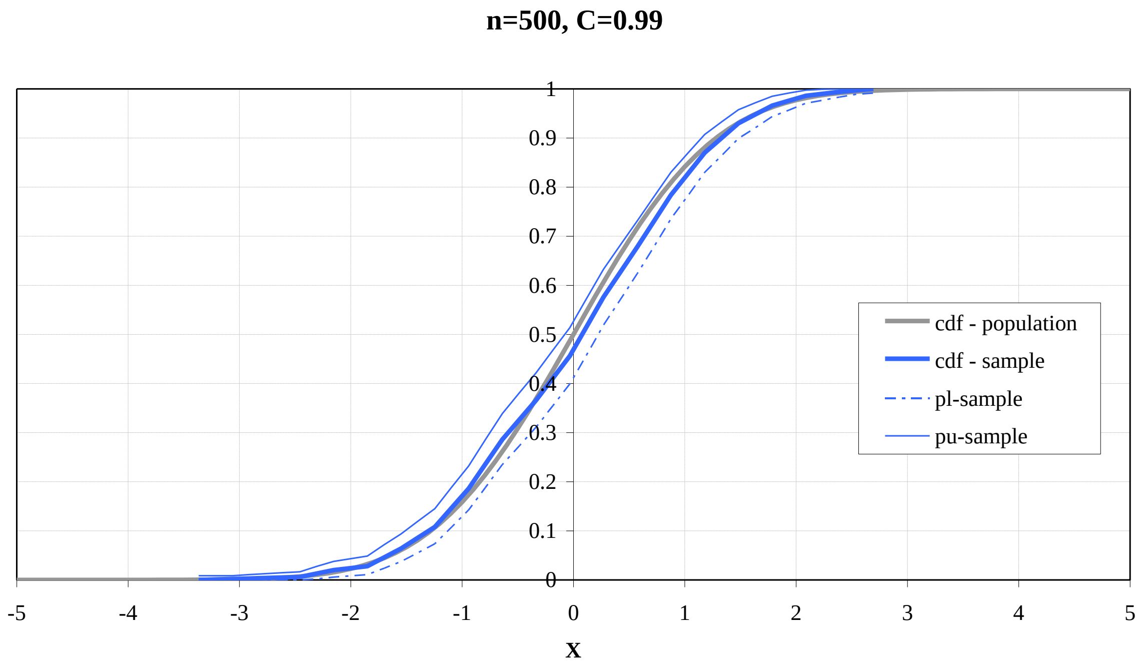

Figure 20 - Sampling error, n=500, c=0.99

Figure 20 shows that since the cdf of a population is never known, the sampling error allows deriving the interval around the sample cdf within which the population cdf can be expected with C probability. The sample size is n=500.