Central Limit Theorem

Anthony Perez Eisenbarth

The Central Limit Theorem (CLT) establishes that, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a bell curve) even if the original variables themselves are not normally distributed. Put another way, the CLT states that the sum of a number of independent and identically distributed random variables with finite variances will tend to a normal distribution as the number of variables grows.

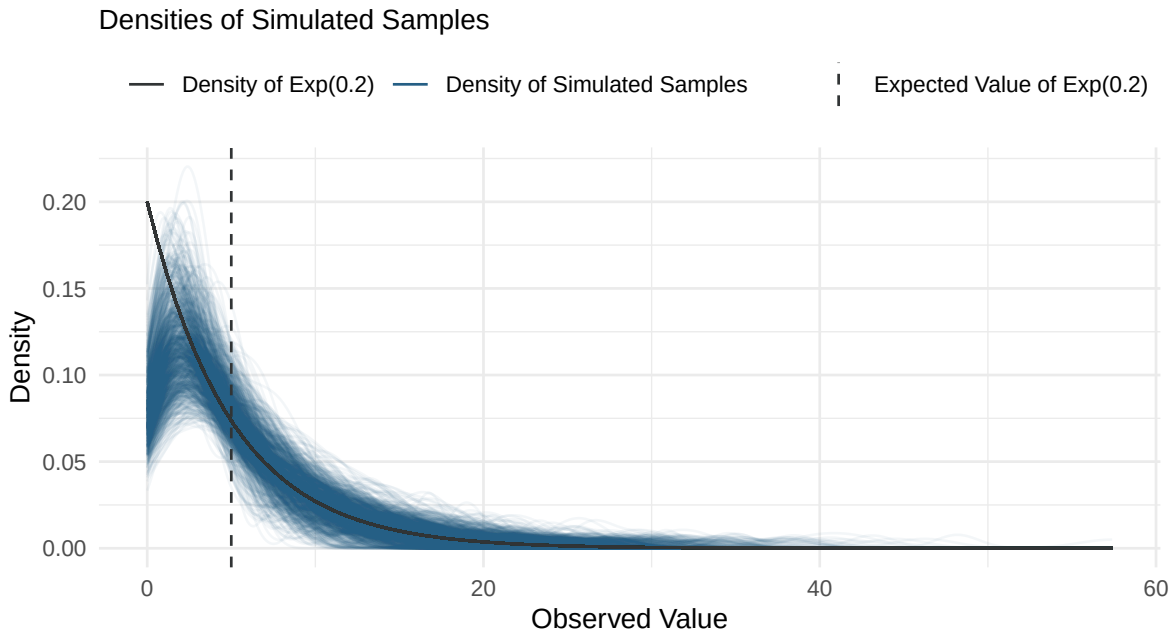
A simple example of this is that if one flips a coin many times, the probability of getting a given number of heads will approach a normal distribution, with the mean equal to half the total number of flips. At the limit of an infinite number of flips, it will equal a normal distribution.

In this experiment, $m = 1000$ random samples are produced, each of which with $n = 40$ observations. All the observations were produced from an exponential distribution $\lambda e^{-\lambda x}$ with rate $\lambda = 0.2$. The goal of the experiment is verify, through simulation, the veracity of the CLT. If the CLT holds, then the probability distribution of the average and the variance of the samples drawn from the above exponential distribution will converge to a standard normal distribution.

Simulated Samples

From the simulation, a 1000 samples $\underline{y}_i$ were obtained, which contain observations that are the realizations of the 40000 i.i.d. random variables $Y_{i,j} \sim Exp(0.2)$ where $i = 1, 2, \ldots, 1000$ and $j = 1, 2, \ldots 40$. A random variable $Y_{i,j} \sim Exp(\lambda)$, where $\lambda \in (0, \infty)$, has an expected value $\mu_Y := E(Y_{i,j}) = \frac{1}{\lambda}$ and variance $\sigma_Y^2 := Var(Y_{i,j}) = \frac{1}{\lambda^2}$. So for $\lambda = 0.2$ it is $\mu_Y = 5$ and $\sigma_Y^2 = 25$.

Figure 1



A comparison of the densities from the 1000 simulated samples
versus the Exponential distribution from which their observations were sampled.

Sample of Sample Means

For each random sample $\underline{Y}_i = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,40})$, the random variable $\bar{Y}_i = \frac{1}{40} \cdot \sum_{j=1}^{40} Y_{ij}$,is the sampling mean. The sampling means $\bar{Y}_i$, of the the 1000 random samples, are i.i.d. random variables because they are linear transformations of the i.i.d. random variables $Y_{i,j}$.

Define $X_i := \bar{Y}_i$, so the random variable X_i has expected value $\mu_X = E(X_i) = E(\bar{Y}_i)$ and variance $\sigma_X^2 = Var(X_i) = Var(\bar{Y}_i)$. The sample $\underline{x} = (x_1, x_2, \ldots, x_{1000}) \equiv (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{1000})$ where $x_i = \bar{y}_i = \frac{1}{40} \cdot \sum_{j=1}^{40} y_{i,j}$, consists of the observed sample means of 1000 original samples $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_{1000}$

Mean of Sample Means

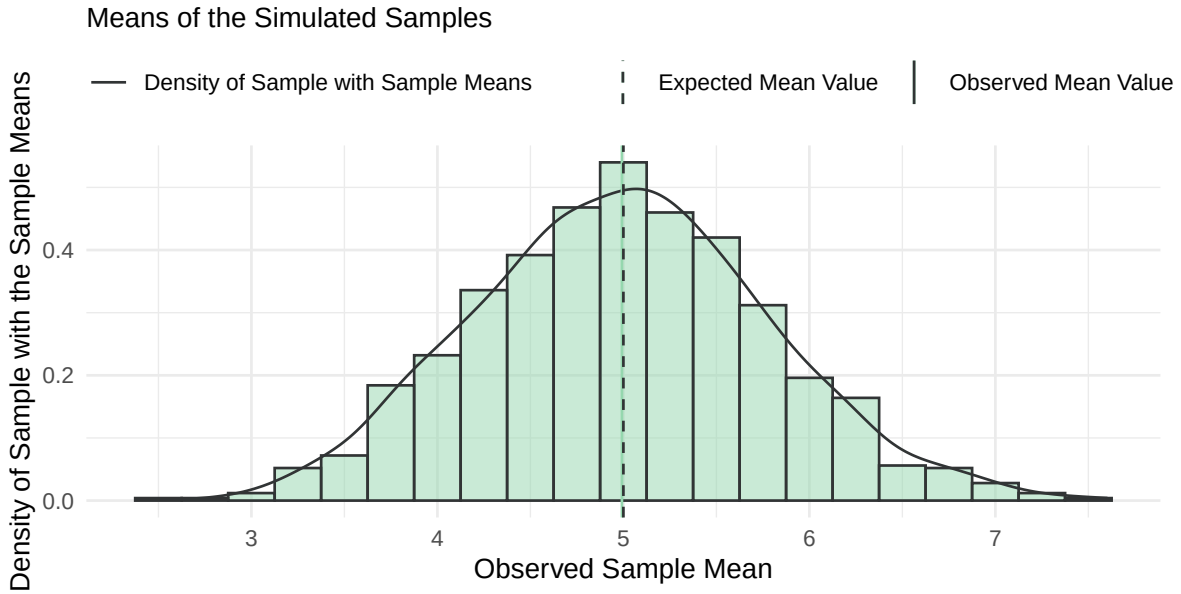
According to the proposition (P1), the random variable X_i defined as equivalent to the sampling mean $\bar{Y}_i$ of the i-th random sample with n observations, has an expected value equal to the expected value of the distribution from which the sample was taken.

So $\mu_X = E(X_i) = E(\bar{Y}_i) = E(Y_{ij}) = \mu_Y = 5$.

Indeed the mean $\bar{x} = 4.992119$, of the sample $\underline{x}$, is 'very close' to the theoretical expected value $\mu_X = 5$.

```
## [1] 4.992119
```

Figure 2

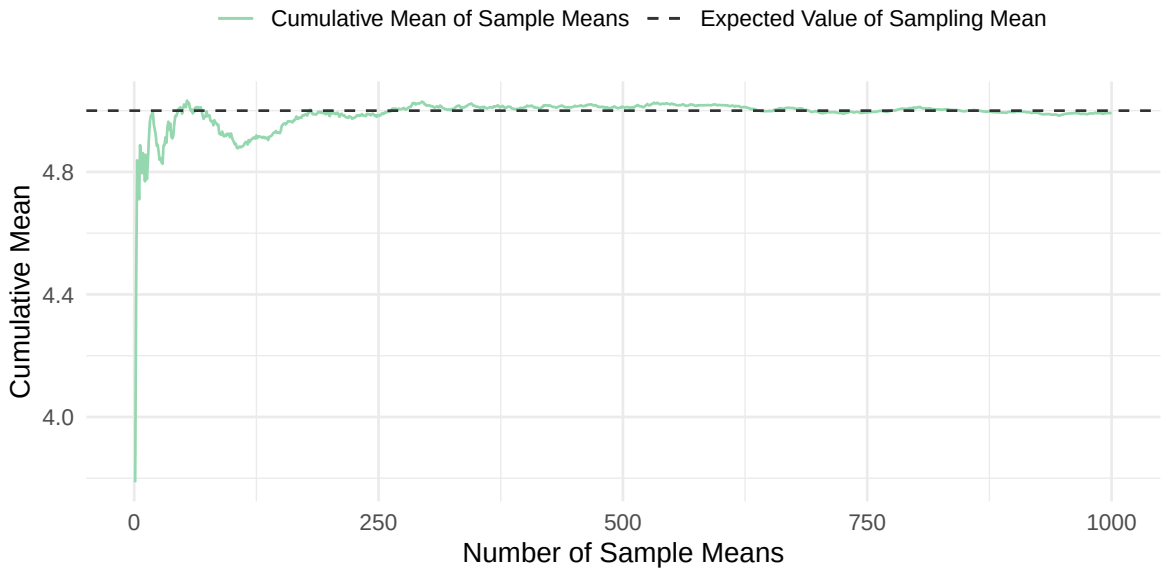


The Histogram and the Density of Sample with the Means of the Simulated Samples.
The Mean of all Sample Means is 'close' to the Expected Value
of the Exponential Distribution from which the samples were produced
but the density curve is not similar.

Furthermore as the number of means from sample means increases the cumulative mean converges to the expected value $\mu_x = 5$ (Figure 3).

Figure 3

Cumulative Mean of Sample Means.



The cumulative mean of i.i.d. sample means, converges to the theoretical expected value of the sampling mean, as the number of the observed means increases.

Variance of Sample Means

According to the proposition (P2), the random variable X_i defined as equivalent to the sampling mean $\bar{Y}_i$ of the i-th random sample with n observations, has a variance equal to the variance of the distribution from which the sample was taken divided by the sample size n.

So $\sigma_X^2 = Var(X_i) = Var(\bar{Y}_i) = \frac{Var(Y_{ij})}{n} = \frac{\sigma_Y^2}{n} = 0.625$.

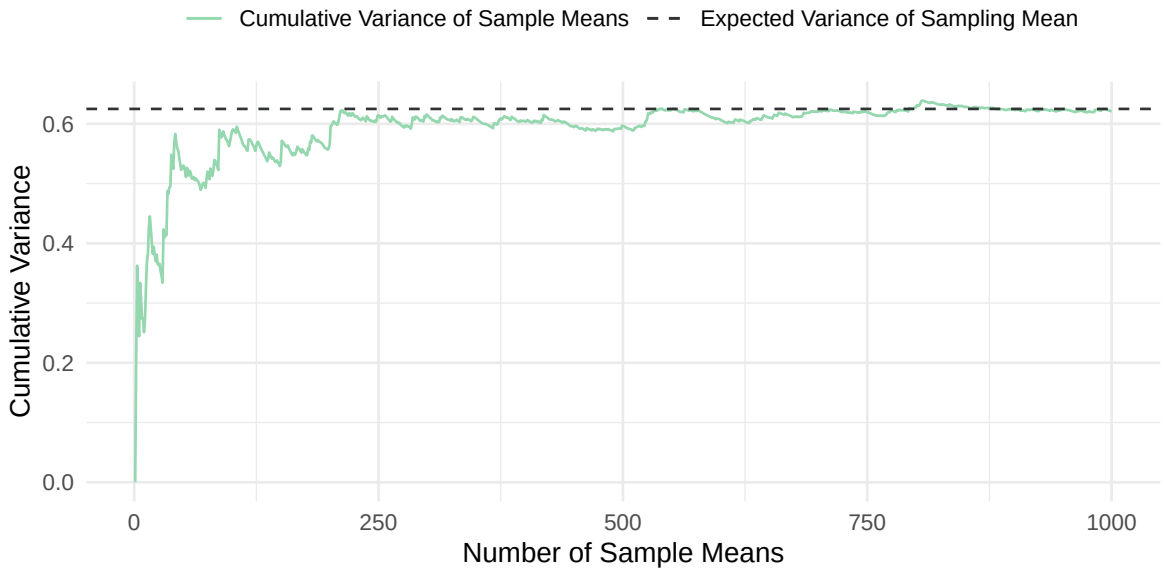
Indeed the sample variance $s^2 = 0.625151$, of the sample $\underline{x}$, is 'very close' to the theoretical variance $\sigma_X^2 = 0.625$.

```
## [1] 0.625151
```

Furthermore as the number of means from sample means increases the cumulative variance converges to the expected value $\sigma_X^2 = 5$ (Figure 4).

Figure 4

Cumulative Variance of Sample Means.



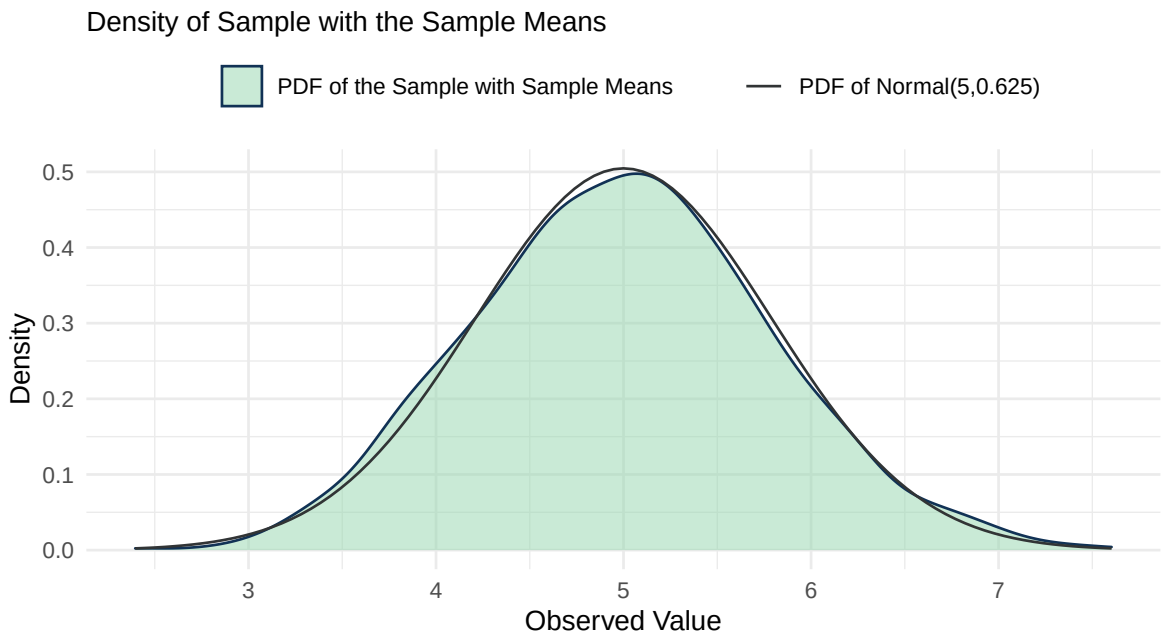
The cumulative variance of i.i.d. sample means, converges to the theoretical variance of the sampling mean, as the number of the observed means increases.

Distribution of Sample Means

According to the proposition (P3), the distribution of the random variable X_i defined as equivalent to the sampling mean $\bar{Y}_i$ of the i-th random sample with n observations, is approximately Normal with expected value μ_X and variance σ^2_X.

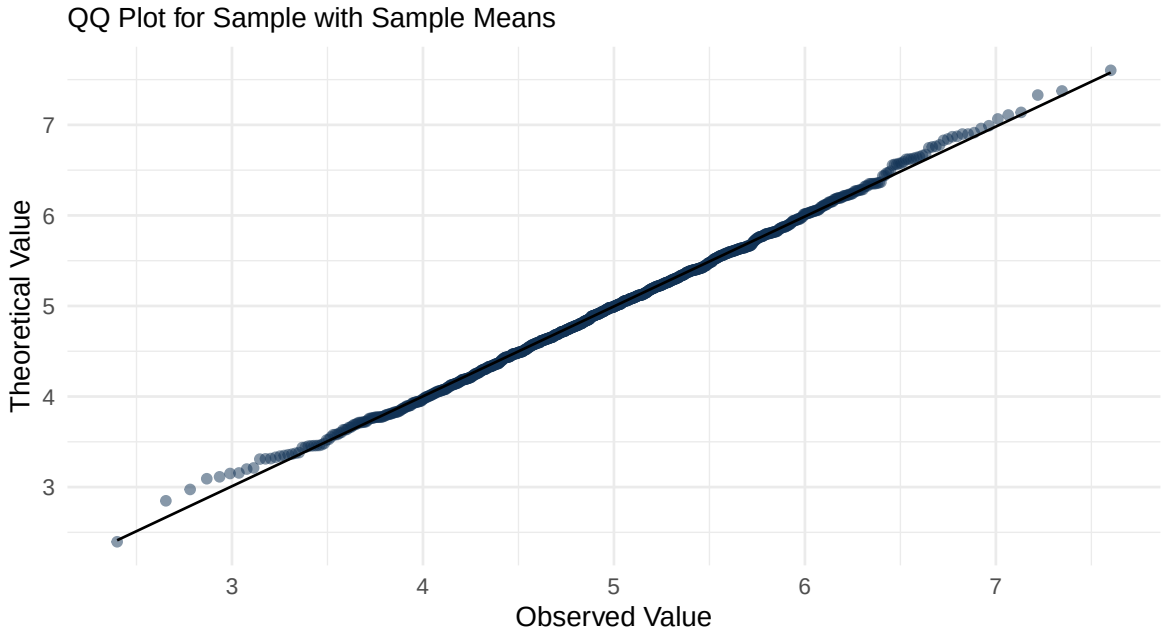
From a visual examination of the density (Figure 5), and the QQ plot (Figure 6) of the sample $\underline{x}$ with the means of the simulated samples it is clear that the observed values fit or approximate those from a normal distribution with expected value $\mu_X = 5$ and variance $\sigma^2_X = 0.625$.

Figure 5



The Density of the Sample with Sample Meansseems to fit very well
the PDF of the Normal Distribution with expected value 5 and variance 0.625.

Figure 6



The Observed Values of the Sample with Sample Means seems to correspond very well
to the Theoretical values from the Normal Distribution with expected value 5 and variance 0.625.

The normality assumption was also verified with the Shapiro-Wilk test, from which a p-value 0.680 was obtained, indicating that there are not enough evidence to discard the null hypothesis, that the sample comes from a Normal Distribution.

```
##
##  Shapiro-Wilk normality test
##
## data:  sample_means$value
## W = 0.99869, p-value = 0.6803
```

Remarks

It is quite possible, especially for the proposition (P3), to obtain samples that may not pass the Shapiro-Wilk test for normality. By increasing the number of simulated samples and/or the sample size of each of them, the observations will eventually conform with the claims of Central Limit Theorem (CTL) and the Law of Large Numbers (LLN).