Statistical modelling

#1.b Central limit theorem

Dr. Léo Belzile
HEC Montréal

Null distribution

When we perform an hypothesis test, we need to know the behaviour of the statistic under the null hypothesis in order to draw a conclusion (reject/fail to reject $\mathscr{H}_0$)

The test statistic is often

- a Wald statistic (mean or maximum likelihood estimator)

Under regularity conditions and for n sufficiently large, the null distribution is approximately normal. Why?

Central Limit Theorem (informal)

Let $Y_1, \ldots, Y_n$ be a random sample from a distribution with

- expectation μ,
- (finite) variance σ^2.

If n is large, the mean $\overline{Y}_n$ approximately follows a normal distribution with mean μ and variance σ^2/n.

$$\overline{Y}_n \stackrel{\cdot}{\sim} \mathsf{No}(\mu, \sigma^2/n)$$

Central Limit Theorem (formal)

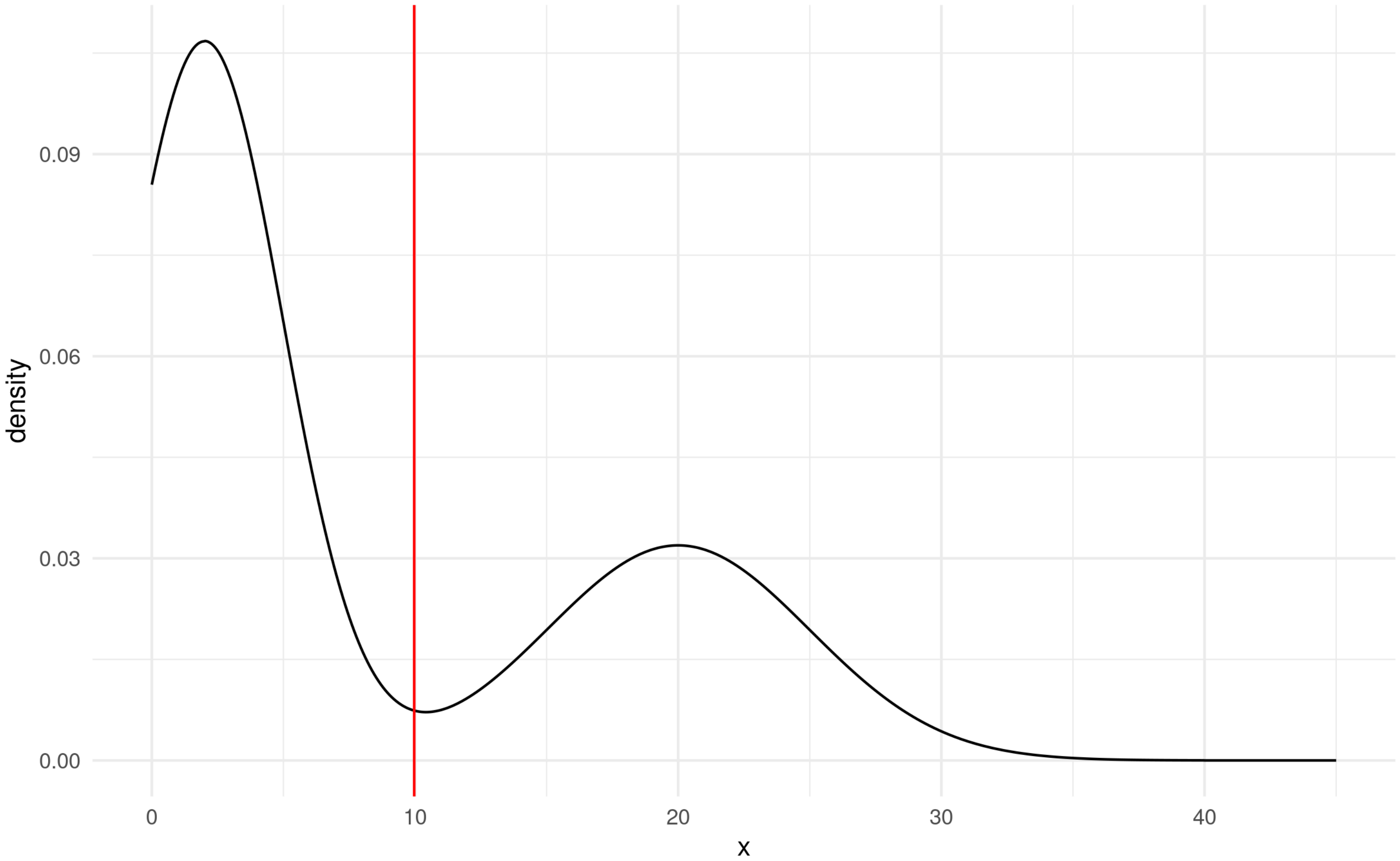
Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables with distribution F and finite variance and let $\overline{Y}_n = n^{-1} \sum_{i=1}^n Y_i$ denote the sample mean.

For any $y \in \mathbb{R}$, the mean converges in distribution to a normal distribution, i.e.,

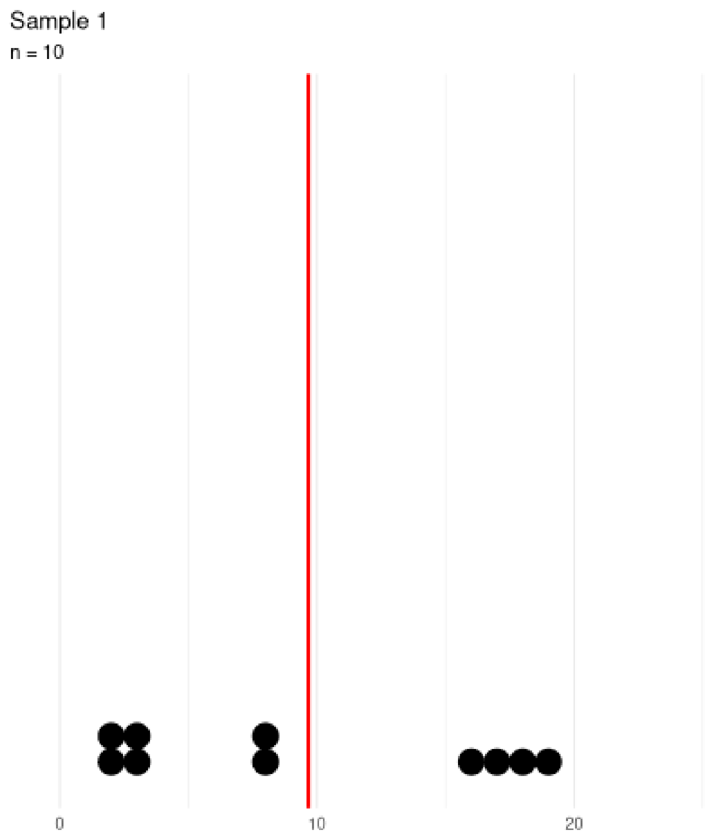
$$\lim_{n \to \infty} \mathsf{P}\left(\sqrt{n}\frac{\overline{Y}_n - \mu}{\sigma} \leq y\right) = \Phi(y)$$

where $\Phi(y)$ is the distribution function of $\mathsf{No}(0,1)$.

Let's represent graphically the central limit theorem by drawing samples repeatedly from the following distribution (left truncated, multimodal, etc.)

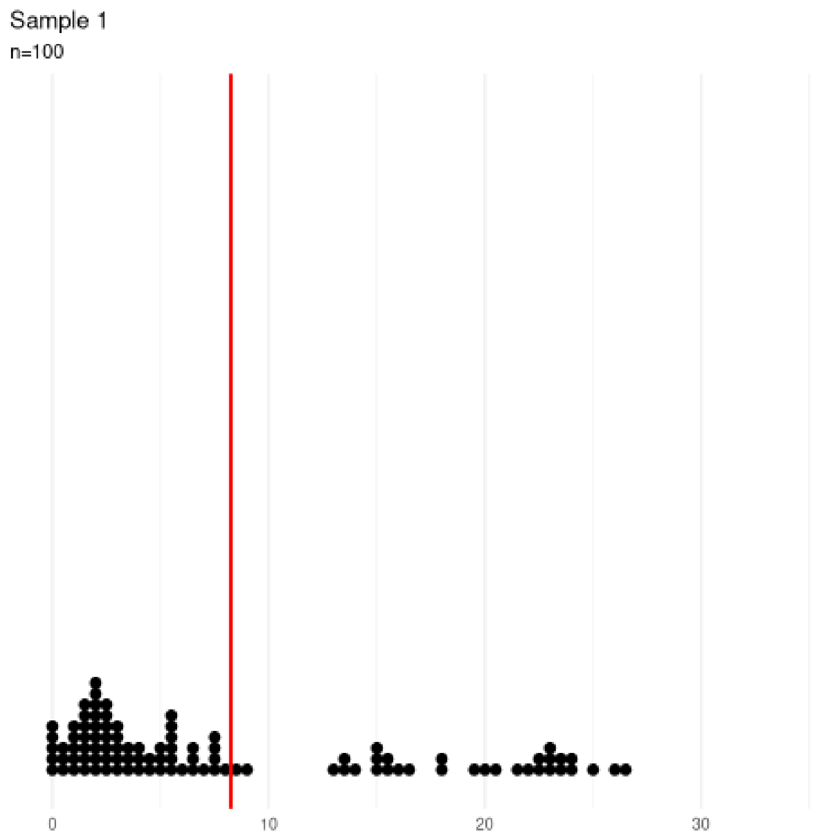


Let's draw $\mathbf{20}$ random samples of size $\boldsymbol{n} = \mathbf{10}$.



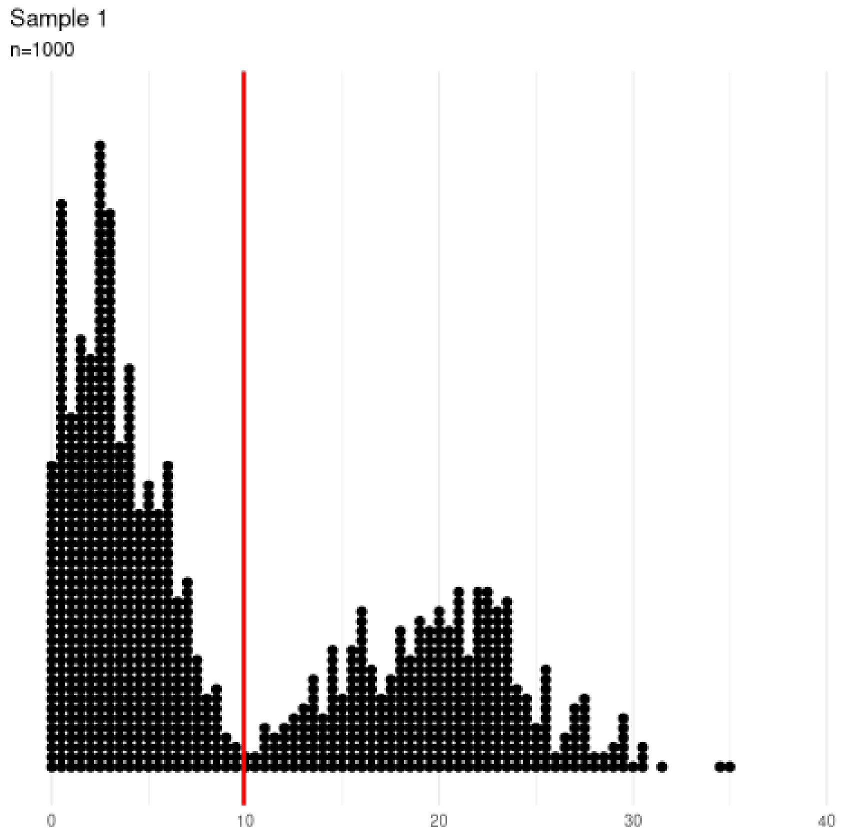
Dot plot of random sample of size $\boldsymbol{n} = \mathbf{10}$ and sample mean (vertical red line)

If we increase the sample size to $n = 100$, the variability of the sample mean decreases.



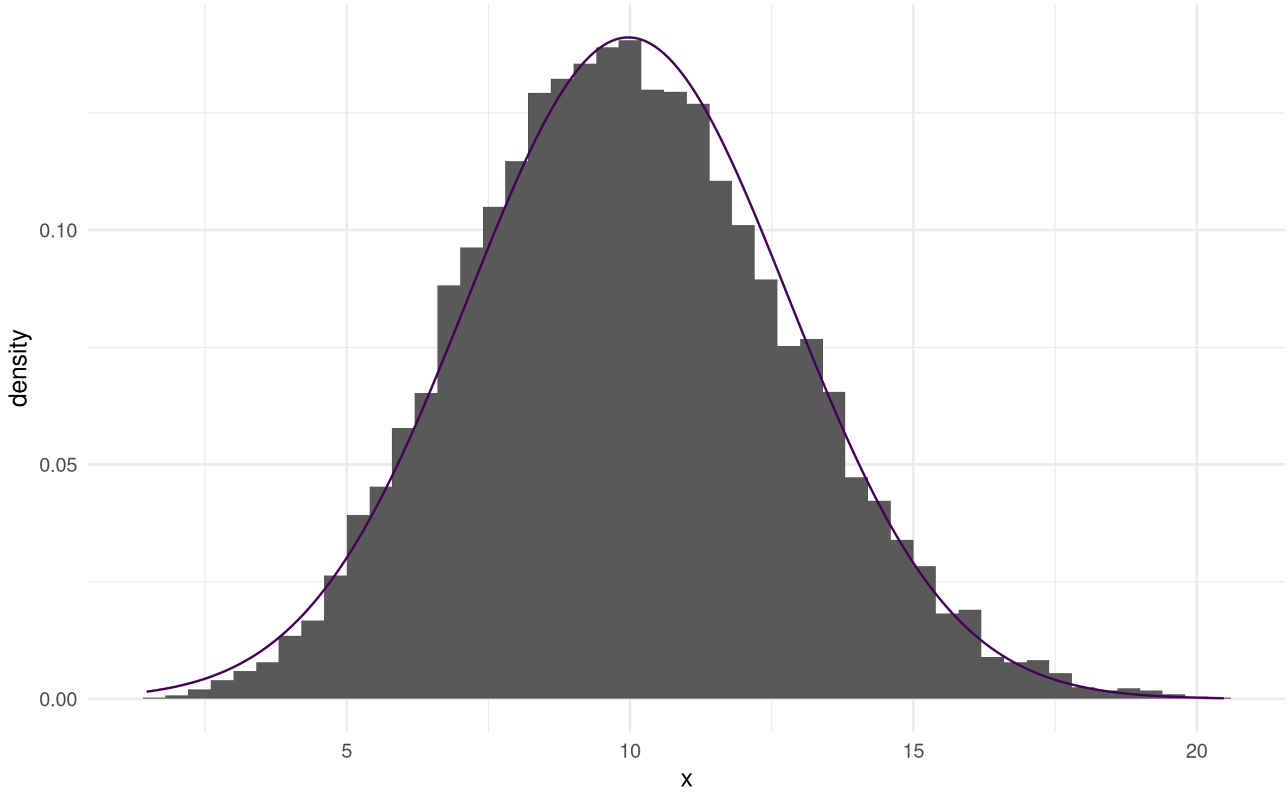
Dot plot of random sample of size $n = 100$ and sample mean (vertical red line)

Same thing, this time with $n = 1000$ observations per sample.



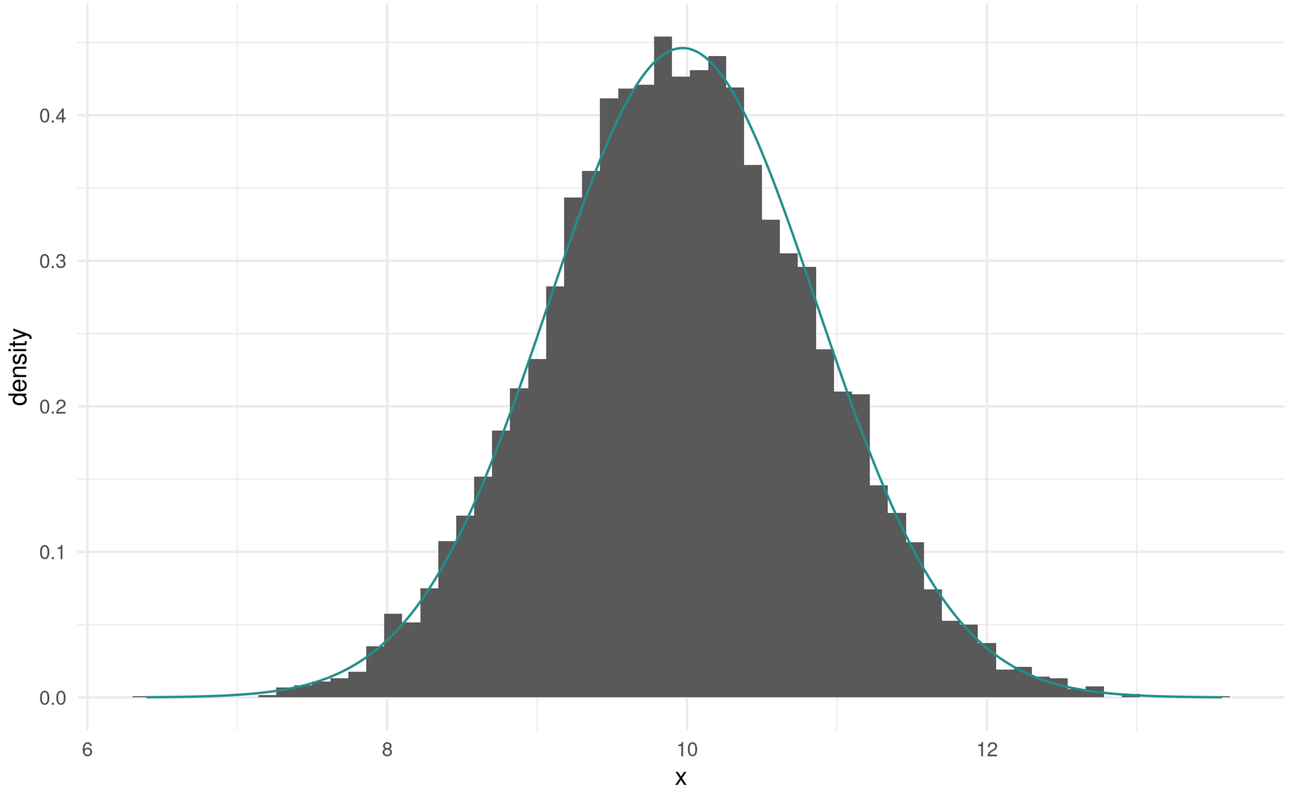
Dot plot of random sample of size $n = 1000$ and sample mean (vertical red line)

If we draw an histogram of the means (vertical red lines), what do we obtain?



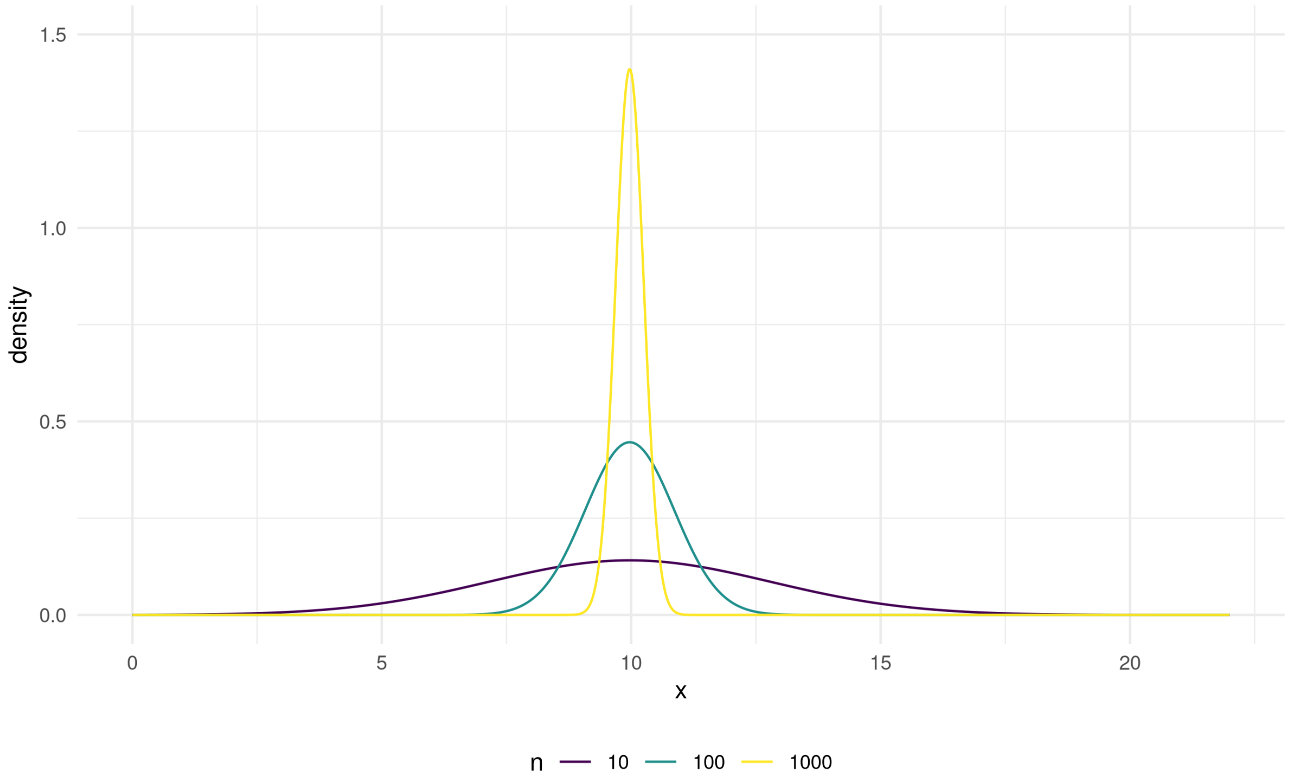
Histogram of the empirical distribution of sample means of $n = 10$ observations and CLT normal approximation.

The quality of the CLT approximation improves when the sample size n increases



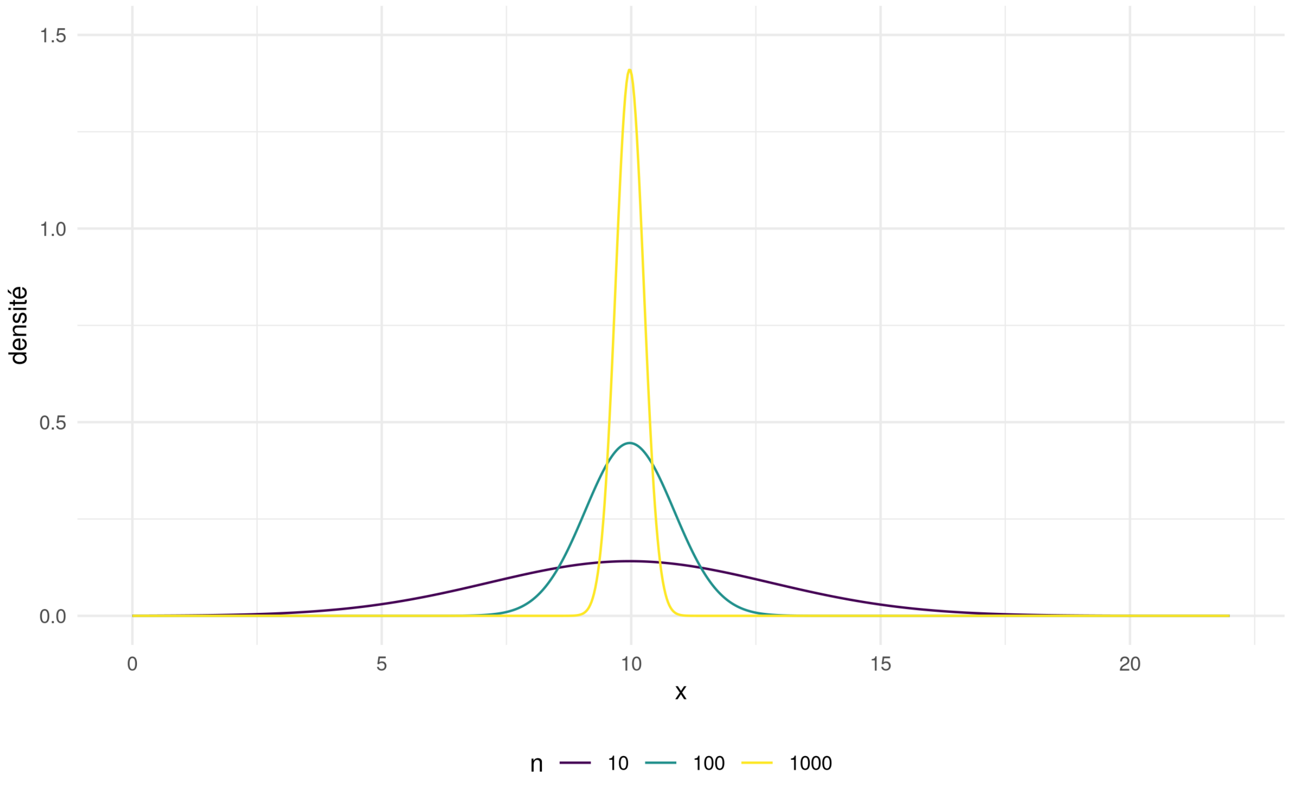
Histogram of the empirical distribution of sample means of $n = 100$ observations and CLT normal approximation.

Convergence is faster near the mean than in the tails of the distribution.



Histogram of the empirical distribution of sample means of $n = 1000$ observations and CLT normal approximation. observations.

The variance of the sample mean $\overline{Y}_n$ when $\mathsf{Va}(Y_i) = \sigma^2 (i = 1, \ldots, n)$ is roughly σ^2/n.



Normal approximation of the mean for different sample sizes.