Modélisation statistique

#1.b Théorème central limite

Dr. Léo Belzile
HEC Montréal

Loi nulle

Lorsqu'on effectue un test statistique, on doit connaître son comportement sous l'hypothèse nulle afin de tirer une conclusion (rejeter/ne pas rejeter $\mathscr{H}_0$).

La statistique de test est souvent

- une statistique de Wald (test-t, estimateur du maximum de vraisemblance)
- racine du test du rapport de vraisemblance

Dans ces cas, sous des hypothèses de régularité et pour n suffisamment grand, la loi nulle qui sert de référence est approximativement normale. Pourquoi?

Théorème central limite (informel)

Si $Y_1, \ldots, Y_n$ est un échantillon aléatoire simple d'une population

- d'espérance μ,
- de variance σ^2 finie.

Alors la loi de la moyenne empirique $\overline{Y}_n$ est approximativement normale centrée en μ et de variance σ^2/n.

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

Théorème central limite (formel)

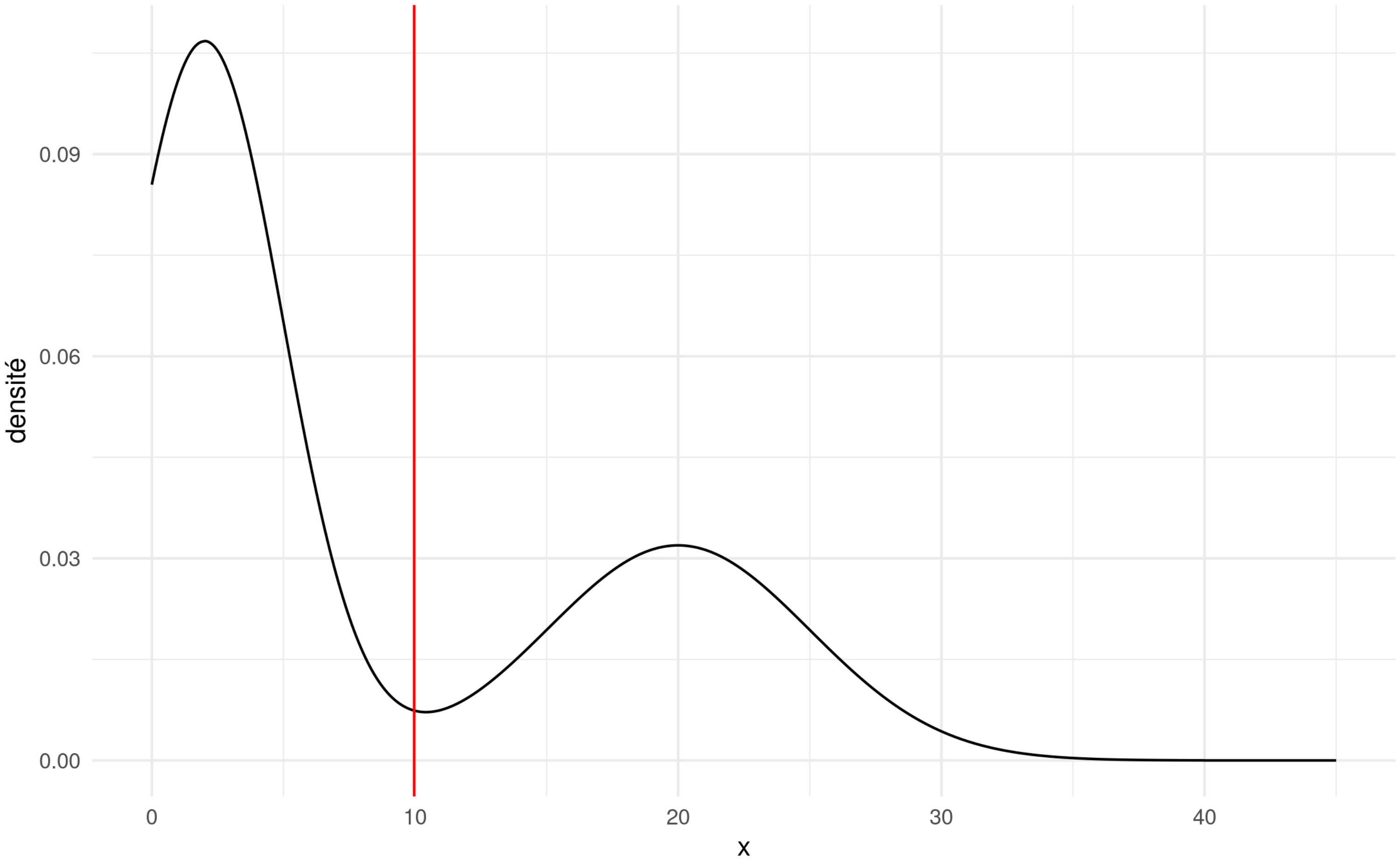
Soit $Y_1, \ldots, Y_n$ des variables aléatoires indépendantes et identiquement distribuées de loi F de variance finie et $\overline{Y}_n = n^{-1} \sum_{i=1}^n Y_i$.

Alors, la moyenne empirique converge en distribution pour tout $y \in \mathbb{R}$,

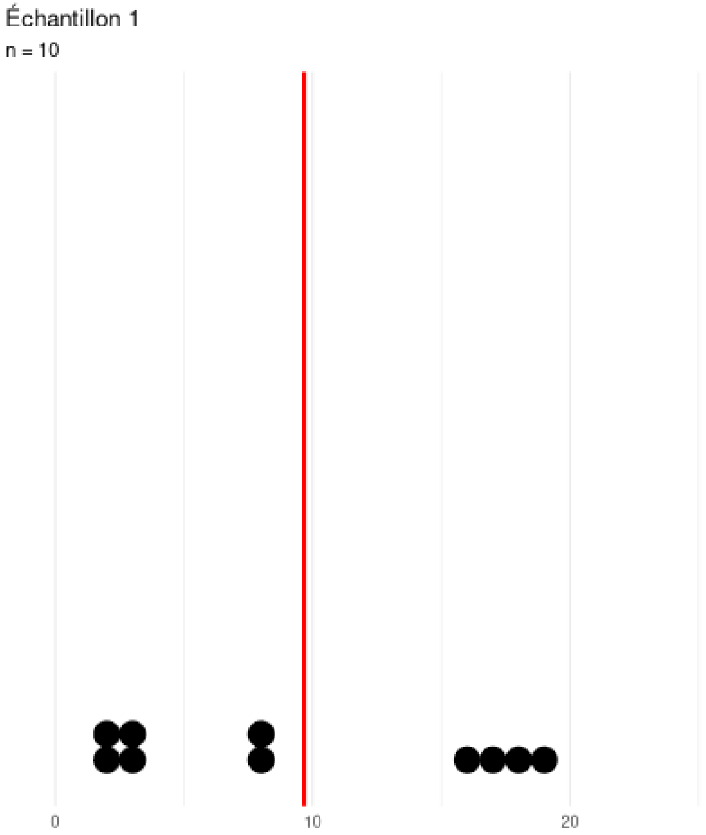
$$\lim_{n \to \infty} \mathsf{P}\left(\sqrt{n}\frac{\overline{Y}_n - \mu}{\sigma} \leq y\right) = \Phi(y)$$

où $\Phi(y)$ est la fonction de répartition de $\mathsf{No}(0, 1)$.

Représentons graphiquement le théorème central limite en tirant des échantillons de la loi suivante (tronquée à gauche, multimodale, etc.)

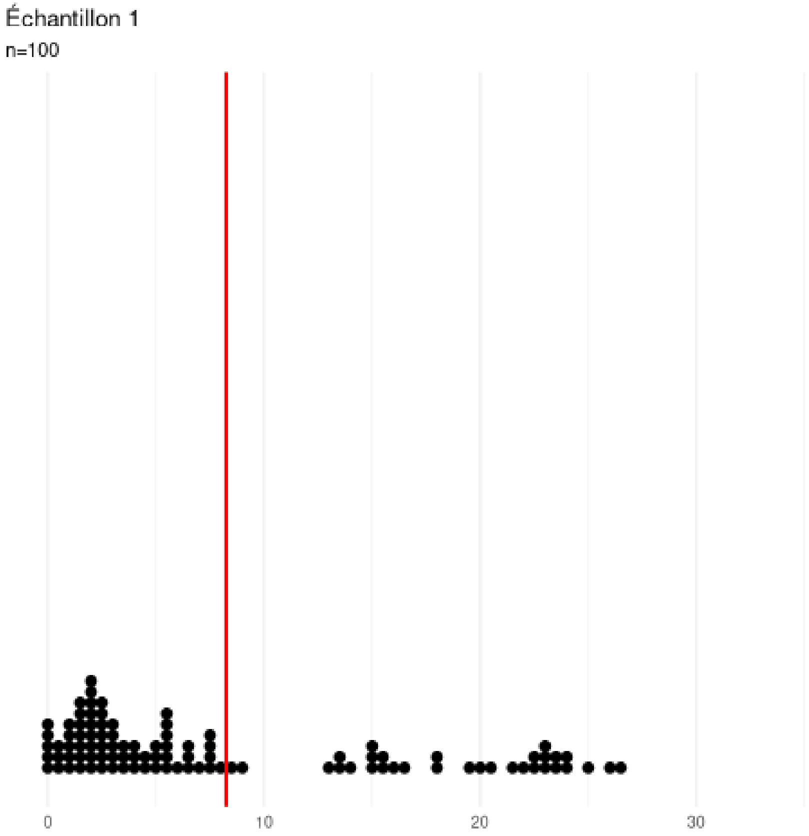


Tirons 20 échantillons aléatoires de taille $n = 10$ de cette loi.



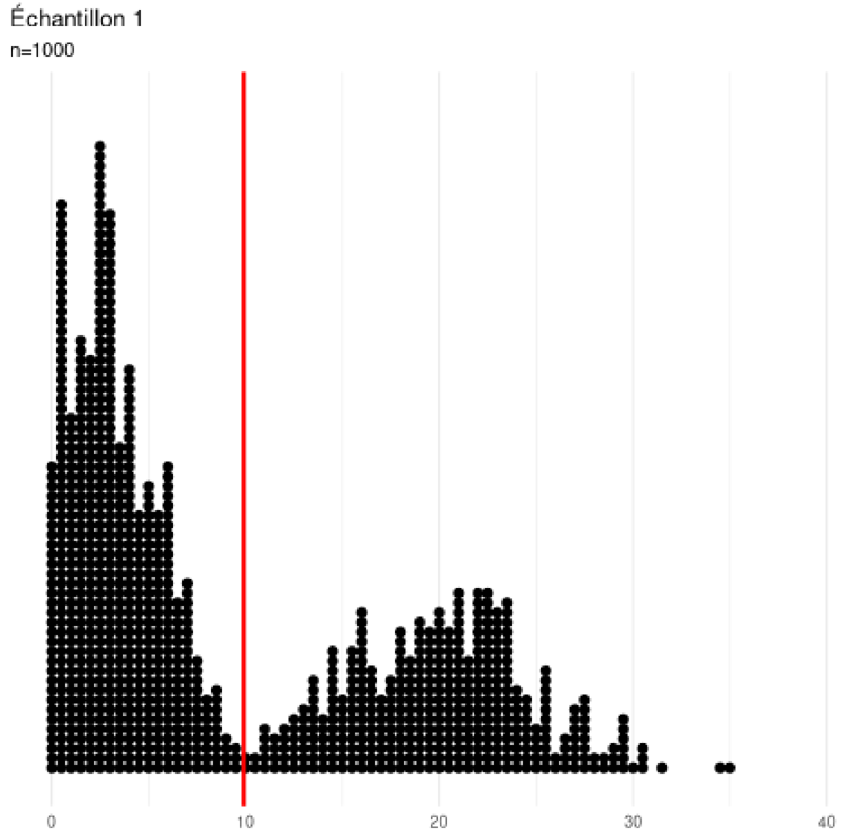
Répartition des $n = 10$ observations et moyenne empirique (trait rouge)

Si on augmente la taille de l'échantillon à $n = 100$, la variabilité de la moyenne diminue.



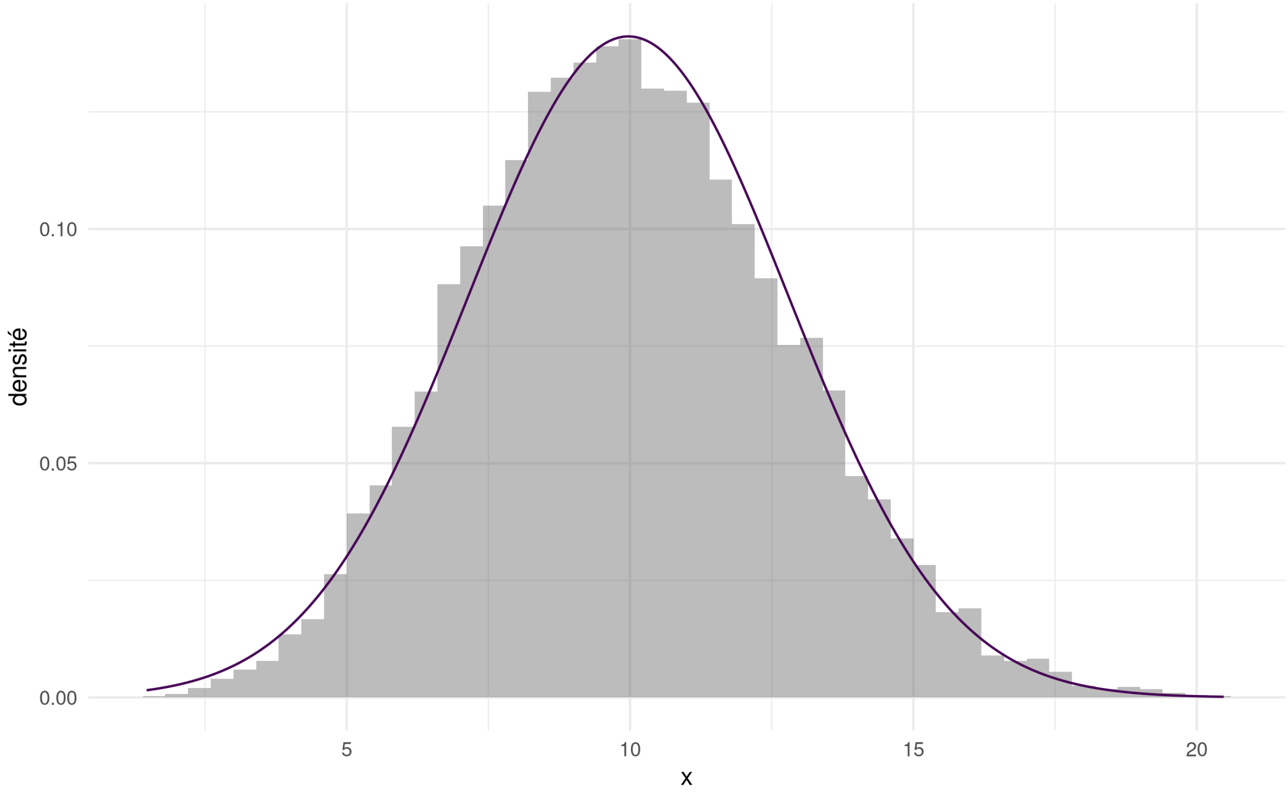
Répartition des $n = 100$ observations et moyenne empirique (trait rouge)

La même chose, avec $n = 1000$ observations par échantillon.



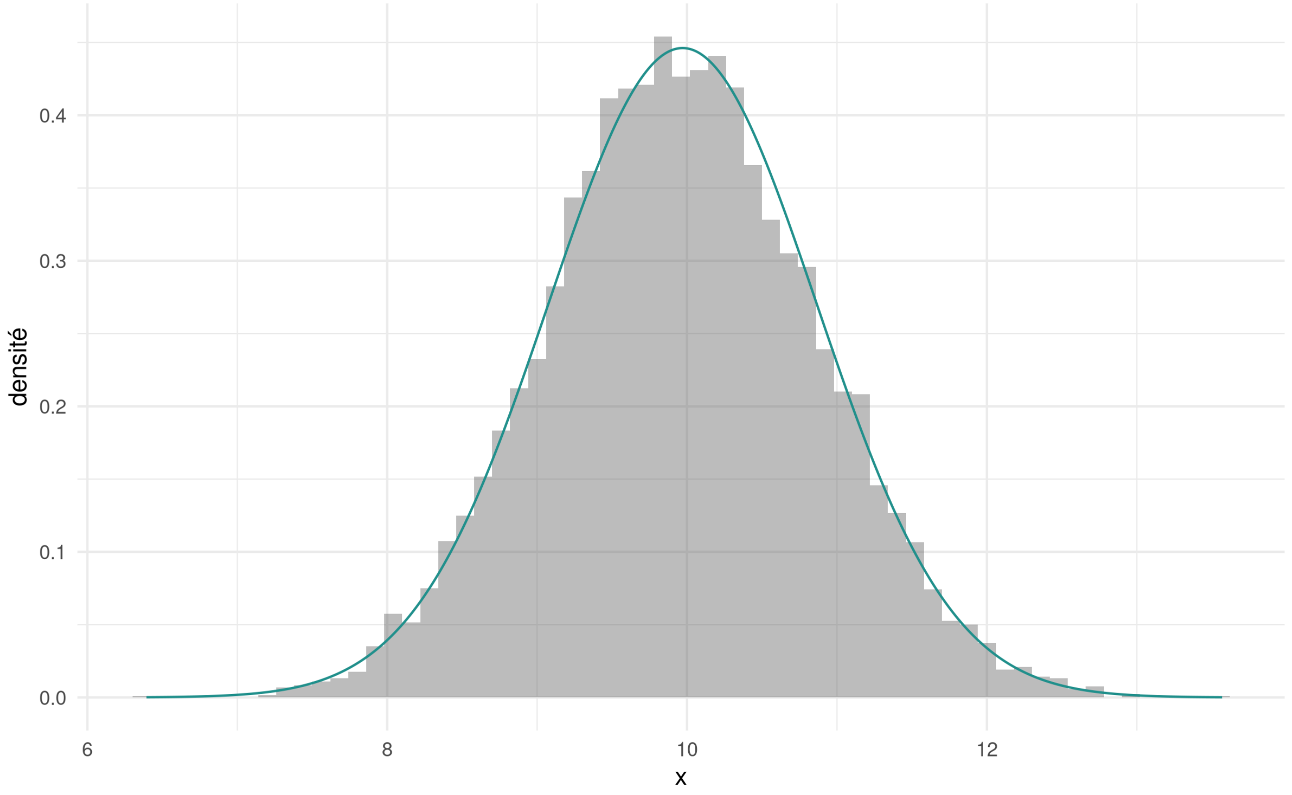
Répartition des $n = 1000$ observations et moyenne empirique (trait rouge)

Si on fait un histogramme des moyennes (traits rouges), qu'est-ce qu'on obtient?



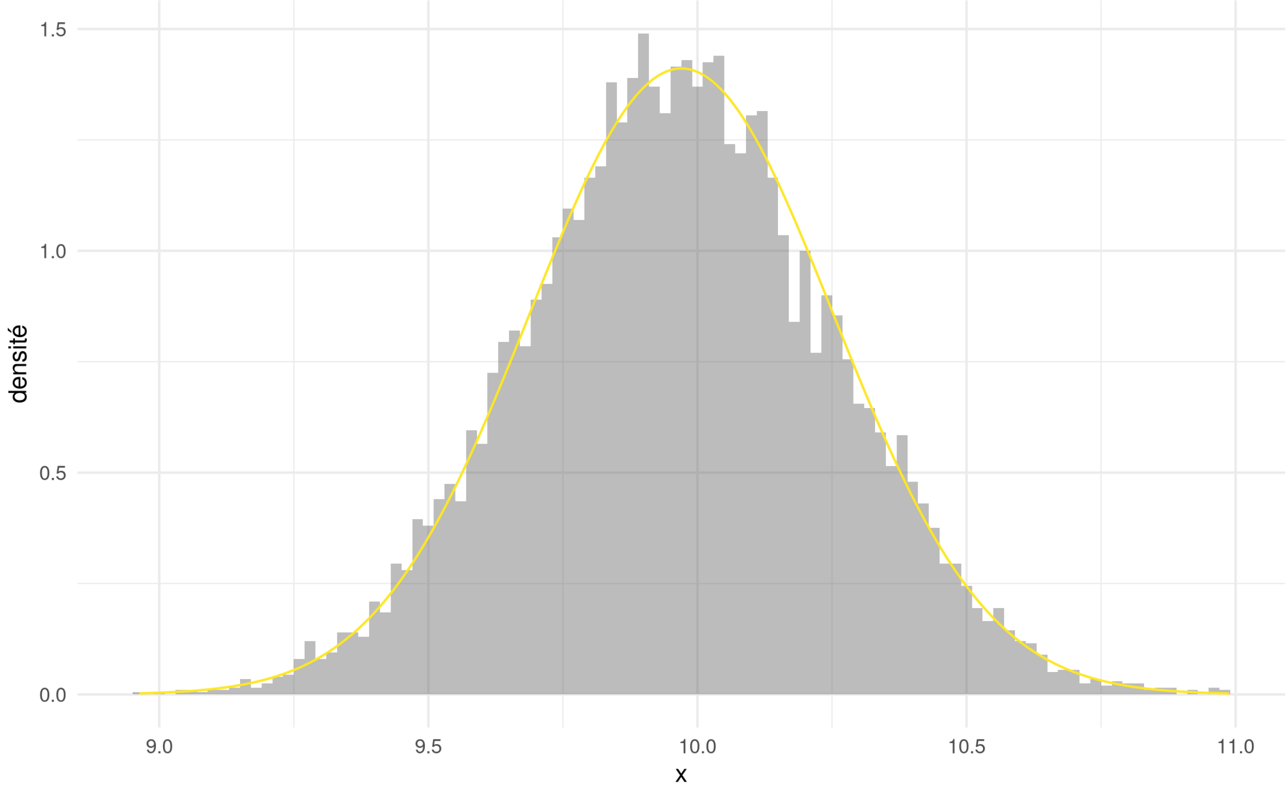
Distribution empirique et approximation normale de la moyenne de $n = 10$ observations.

L'approximation fournie par le théorème central limite est meilleure quand la taille de l'échantillon n augmente.



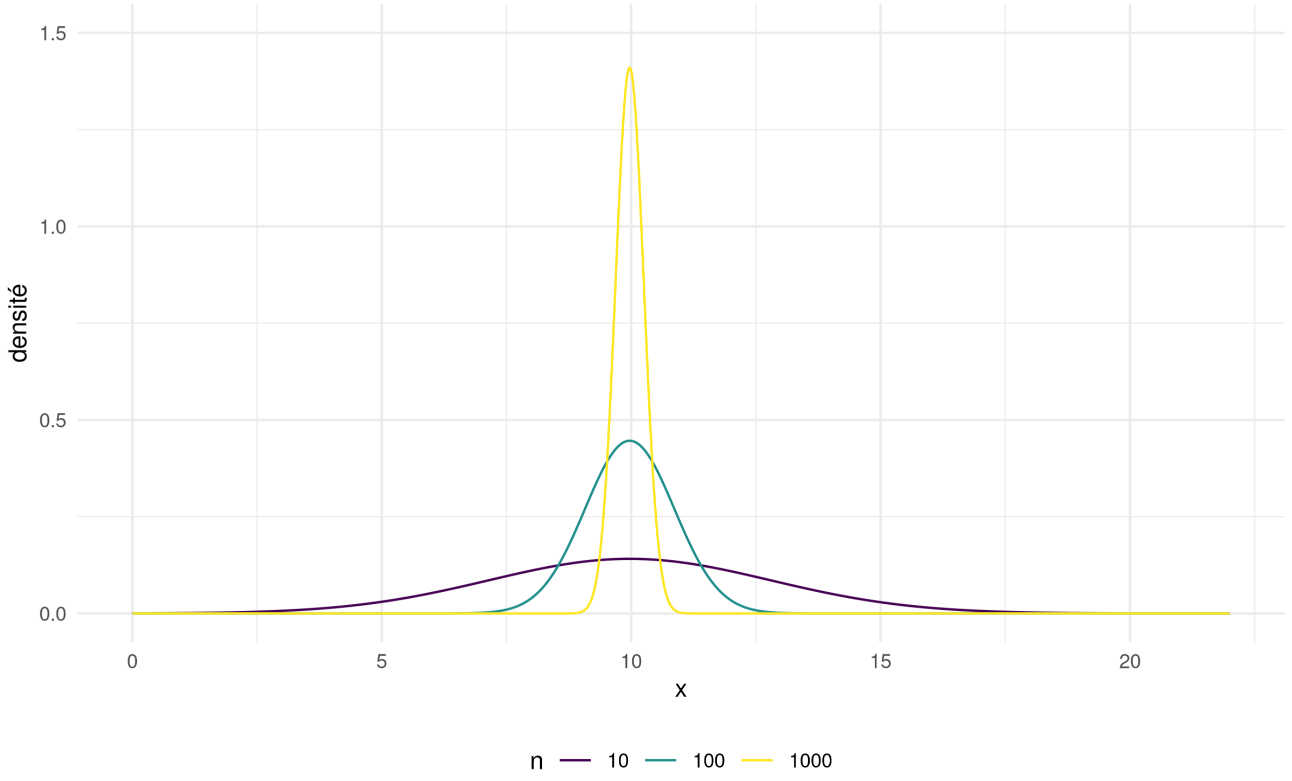
Distribution empirique et approximation normale de la moyenne de $n = 100$ observations.

La convergence est plus rapide au centre de la loi que dans la queue.



Distribution empirique et approximation normale de la moyenne de $n = 1000$ observations.

La variance de la moyenne $\overline{Y}_n$ quand $\mathsf{Va}(Y_i) = \sigma^2$ est σ^2/n.



Approximation normale pour différentes tailles d'échantillons.