Analyse de données
Licence Pro "Amélioration Végétale"

Marc Bailly-Bechet

Université Claude Bernard Lyon I – France

marc.bailly-bechet@univ-lyon1.fr

Table des matières

Organisation des enseignements d'analyse de données

- 3 cours "théoriques" de 1h30.
- 16h de TP sur ordinateur.

Pourquoi faire des statistiques en biologie ?

Variabilité : Une expérience en biologie donne rarement un résultat tranché ou parfaitement reproductible.

Quantité : Les nouvelles technologies biologiques permettent de recueillir des quantités pharamineuses de données.

Les statistiques vues de loin

Population $\Rightarrow$ **Échantillon**

p, μ, σ^2 — n individus tirés aléatoirement

$\Uparrow$ $\Downarrow$

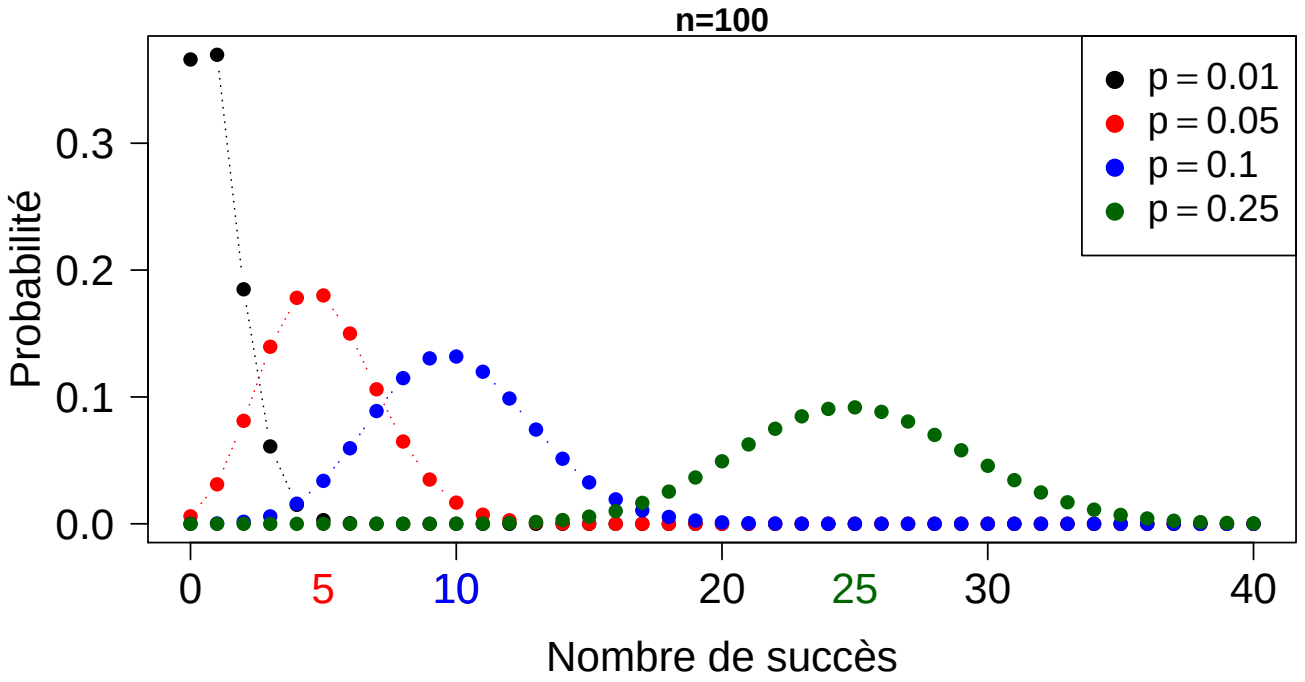
Tests, estimation — $\frac{k}{n}$, $\bar{x}$, s^2

Statistique inférentielle $\Leftarrow$ *Statistiques descriptives*

Table des matières

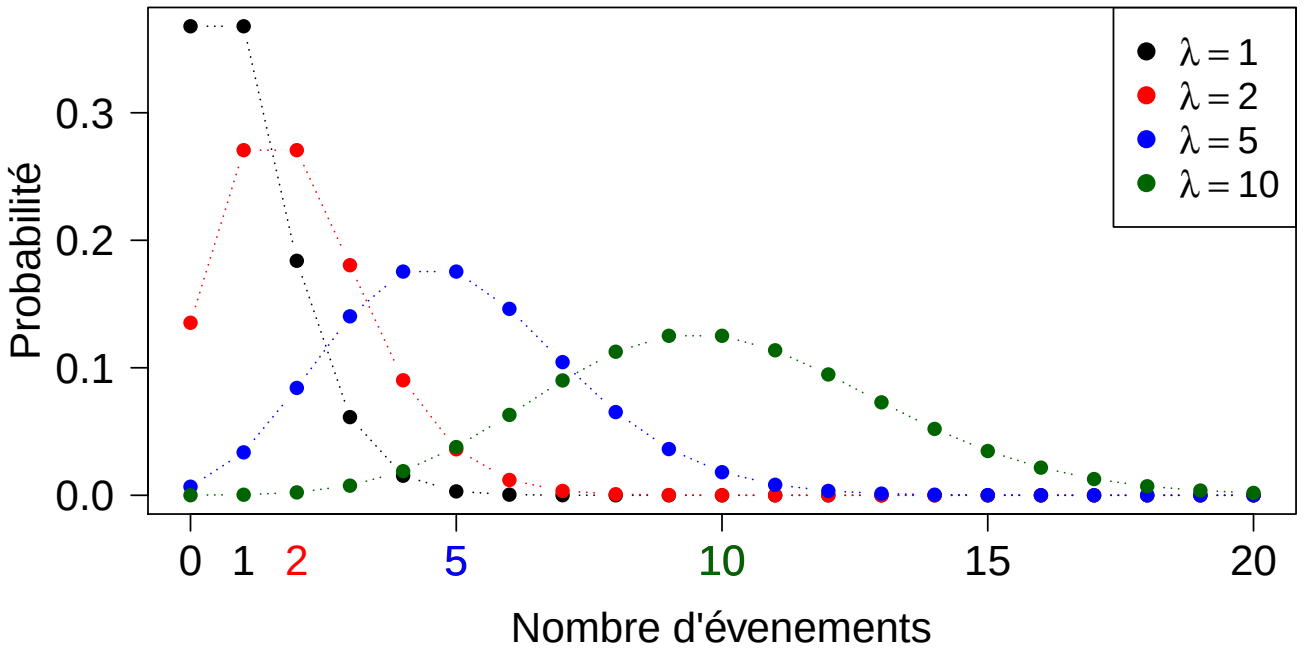
Loi binomiale

La loi binomiale est la loi de probabilité décrivant le *nombre* de réussites parmi un ensemble de tirages aléatoires et indépendants. Elle se note $\mathcal{B}(n, p)$ avec n le nombre de tirages et p la probabilité de réussite à chaque tirage.

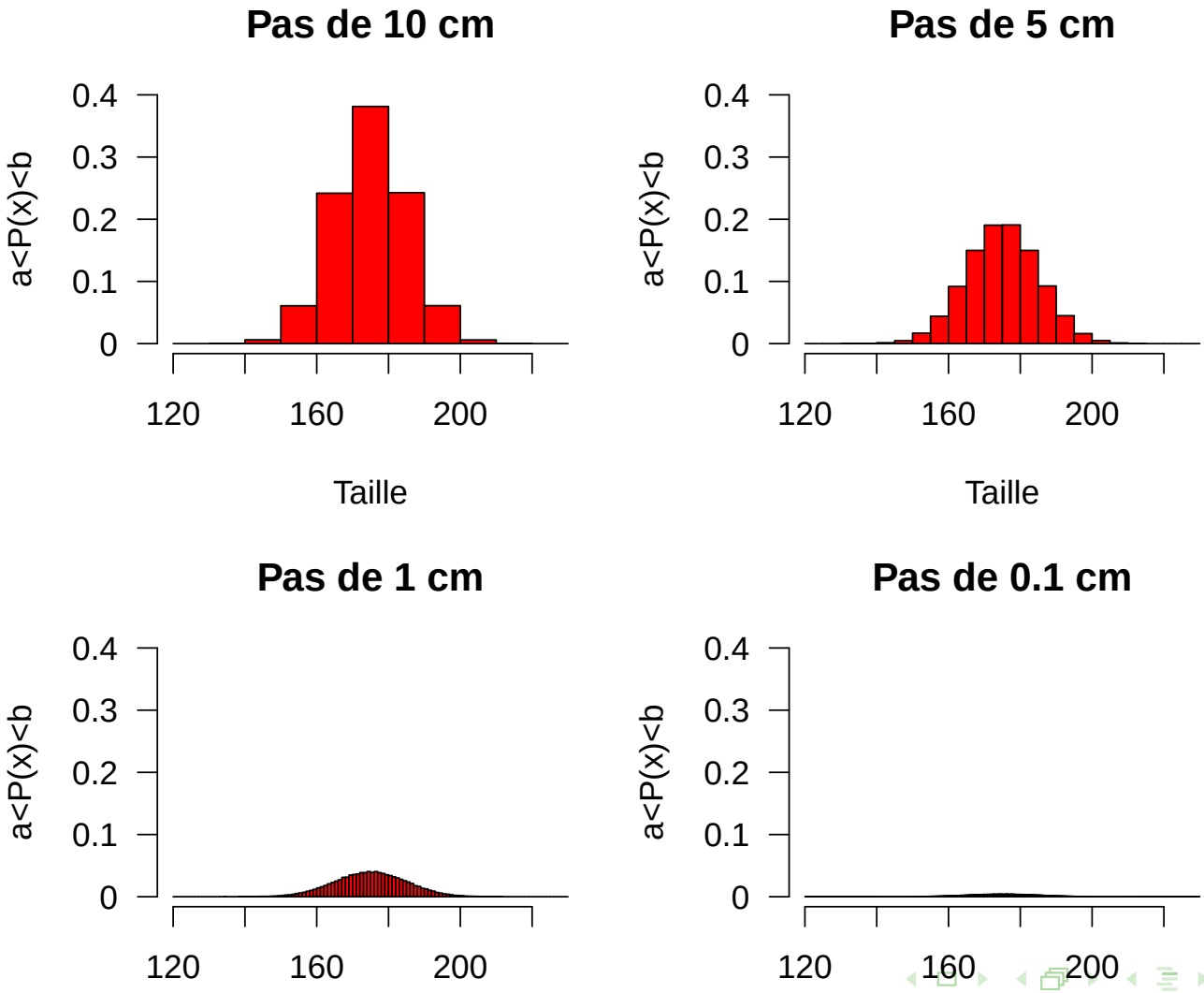


Loi de Poisson

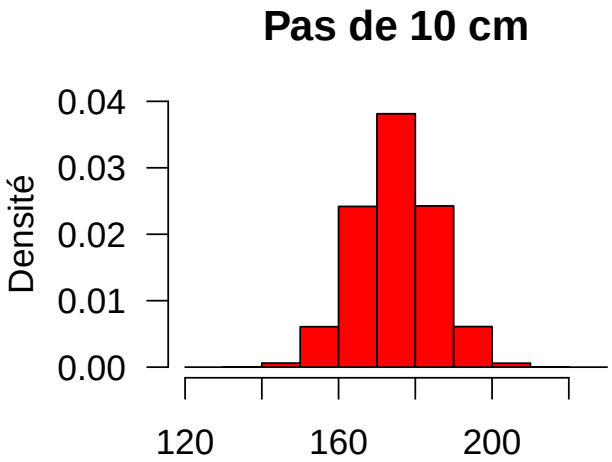
La loi de Poisson (de Siméon Denis Poisson, 1781-1840) est la loi de probabilité décrivant le *nombre d'évenements* aléatoires et indépendants arrivant dans le même intervalle de temps ou d'espace. Elle se note $\mathcal{P}(\lambda)$ avec λ l'espérance et la variance de la loi.

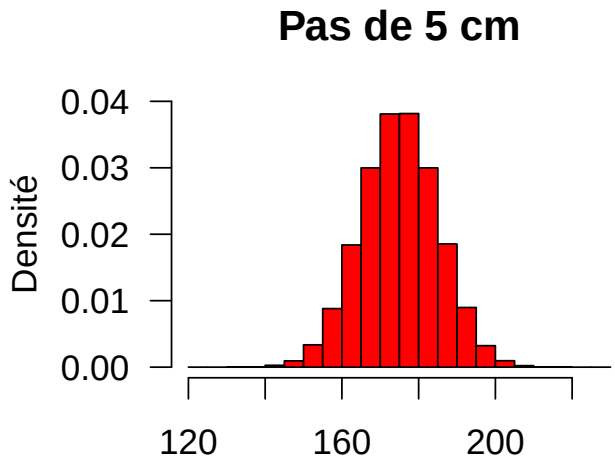


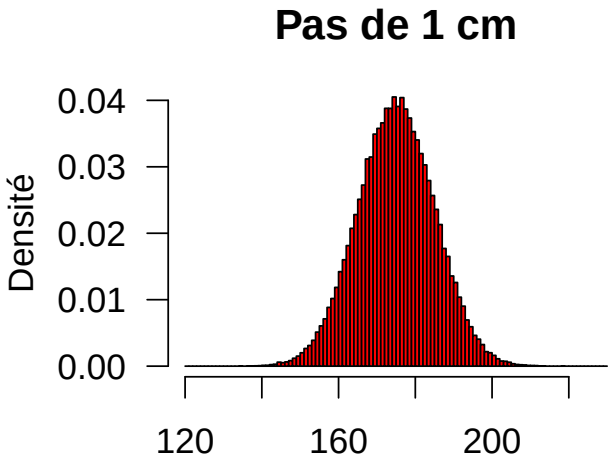
Probabilité absolue

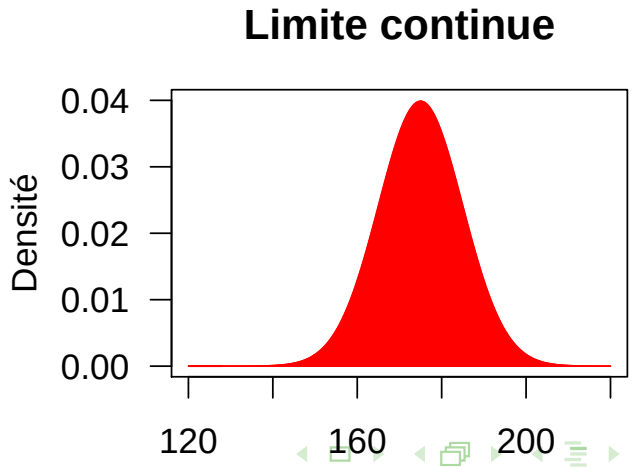


Densité de probabilité









Loi normale

La loi normale est la loi de probabilité des variables aléatoires continues dépendantes d'un grand nombre de causes indépendantes et additives. Elle se note $\mathcal{N}(\mu, \sigma)$ avec μ l'espérance de la loi et σ l'écart-type.

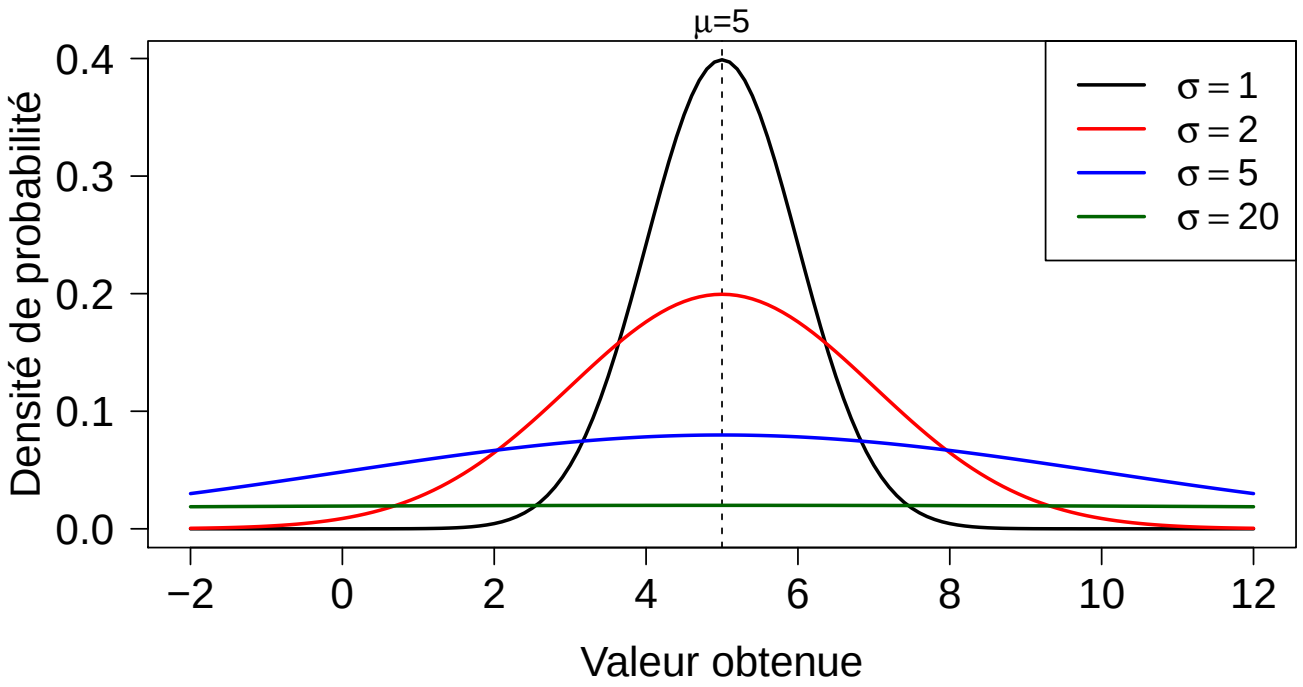


Table des matières

Variable discrète

Le balanin est un parasite de la châtaigne.

Nb. de parasites x_i	0	1	2	3	4	5	6 et plus
Nombre de fruits n_i ayant x_i parasites	1043	172	78	15	10	7	4
Fréquence f_i	0.785	0.129	0.059	0.011	0.007	0.005	0.004
Fréquence cumulée $\sum_{j=1}^{i} f_j$	0.785	0.914	0.973	0.984	0.991	0.996	1

Variable continue

On observe la concentration en glucose dans plusieurs mangues.

Concentration ($g.L^{-1}$) X	Nb de mangues n_j	Fréquence $\frac{n_i}{N}$	Fréquence cumulée $\sum_{j=1}^{i} f_j$
$[135, 150[$	7	0.113	0.113
$[150, 165[$	10	0.161	0.274
$[165, 180[$	23	0.371	0.645
$[180, 195[$	14	0.226	0.871
$[195, 210[$	5	0.080	0.951
$[210, 225[$	3	0.049	1

Moyenne observée sur des données groupées

On veut la moyenne du taux de glucose dans le mélange final de nos 4 types de mangues :

Concentration ($g.L^{-1}$) X	Moyenne $x_j^\star$	Nb de mangues n_j
$[135, 165[$	150	17
$[165, 180[$	172.5	23
$[180, 195[$	187.5	14
$[195, 225[$	210	8

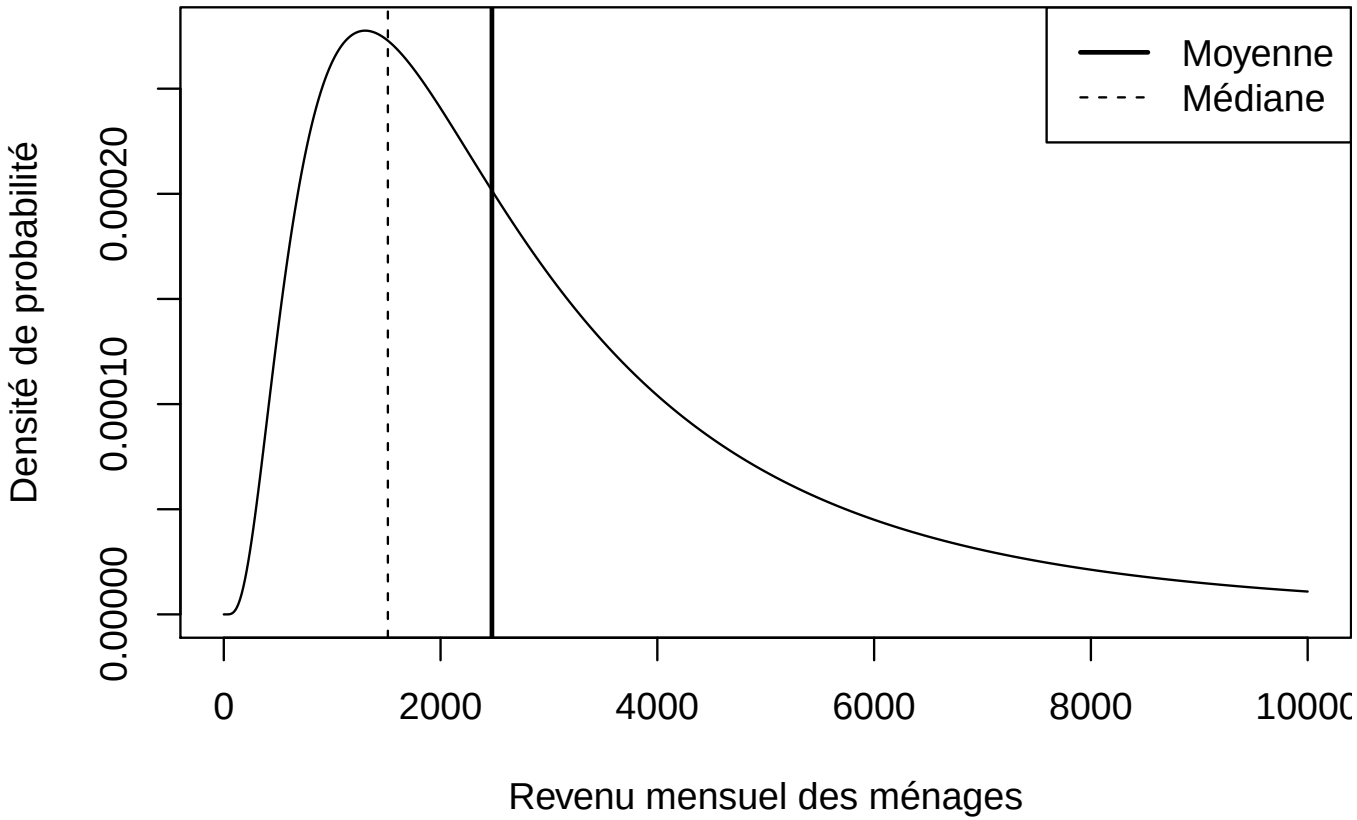
$$\bar{x} = \frac{1}{62}(150 \times 17 + 172.5 \times 23 + \ldots) = \frac{10822.5}{62} = 174.56 \text{ g.L}^{-1}$$

Différence entre médiane et moyenne

Revenu mensuel moyen des ménages en France : 2474 euros
Revenu mensuel médian des ménages en France : 1514 euros

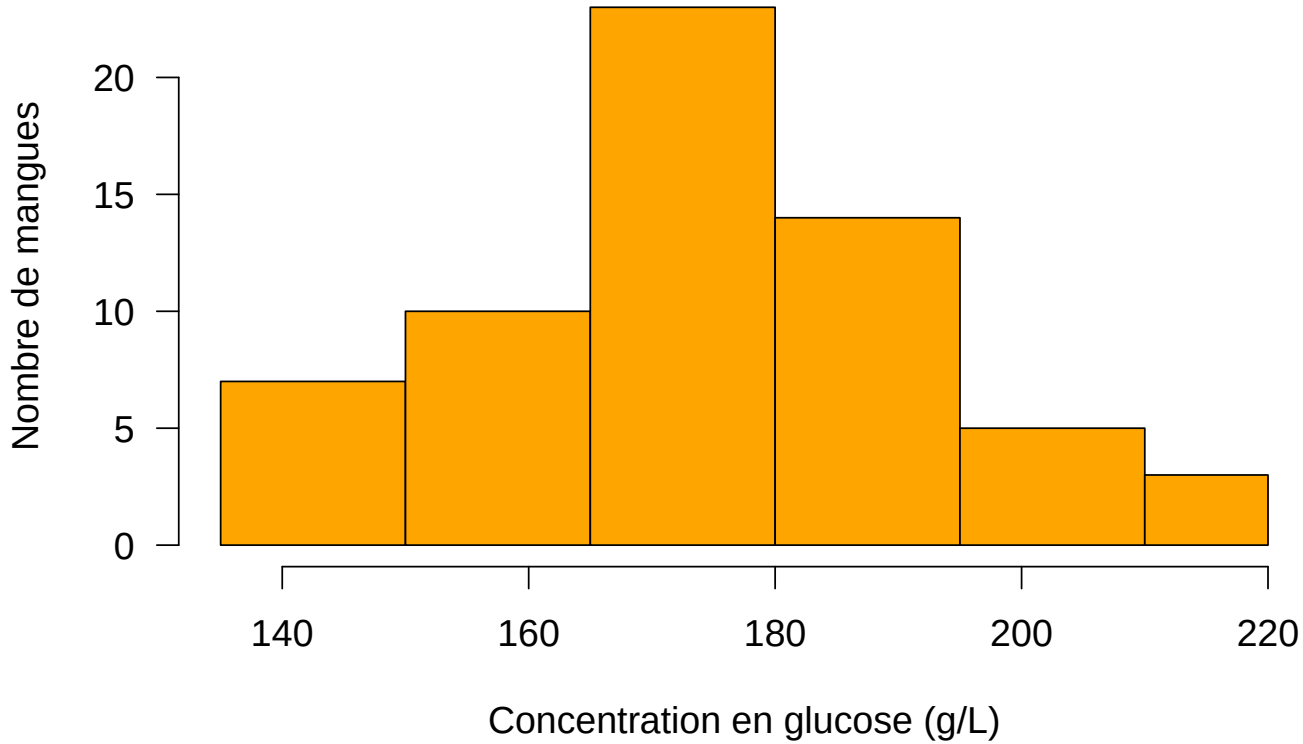
Différence entre médiane et moyenne

Revenu mensuel moyen des ménages en France : 2474 euros
Revenu mensuel médian des ménages en France : 1514 euros



Les mangues sont à la mode

On observe la concentration en glucose dans plusieurs mangues.



Variance et écart-type observés, données groupées

La variance sur des données groupées se calcule ainsi :

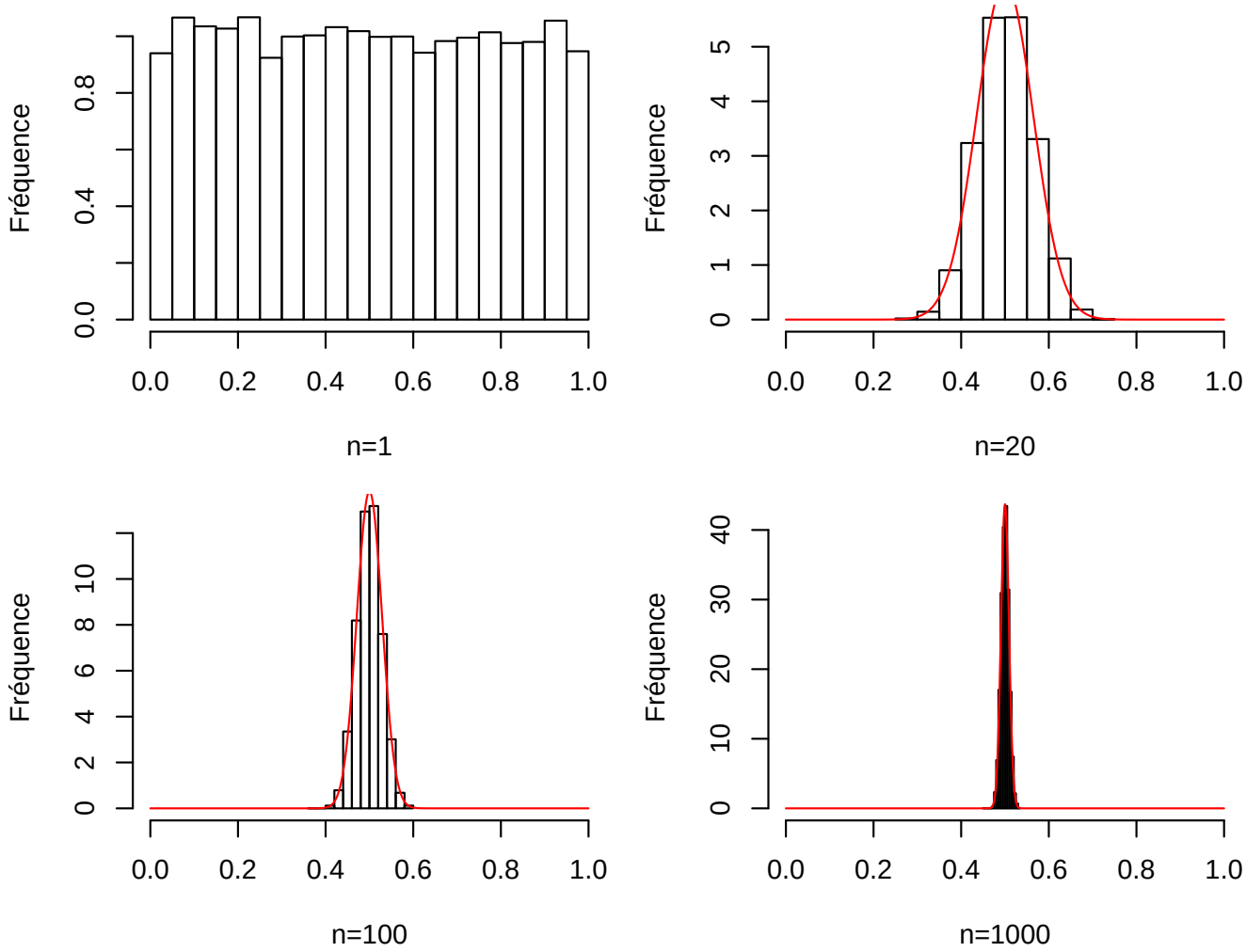
Concentration ($g.L^{-1}$)	Moyenne	Nb de mangues
X	$x_j^\star$	n_j
$[135, 165[$	150	17
$[165, 180[$	172.5	23
$[180, 195[$	187.5	14
$[195, 225[$	210	8

$$\bar{x} = 174.56 \text{ g.L}^{-1}$$

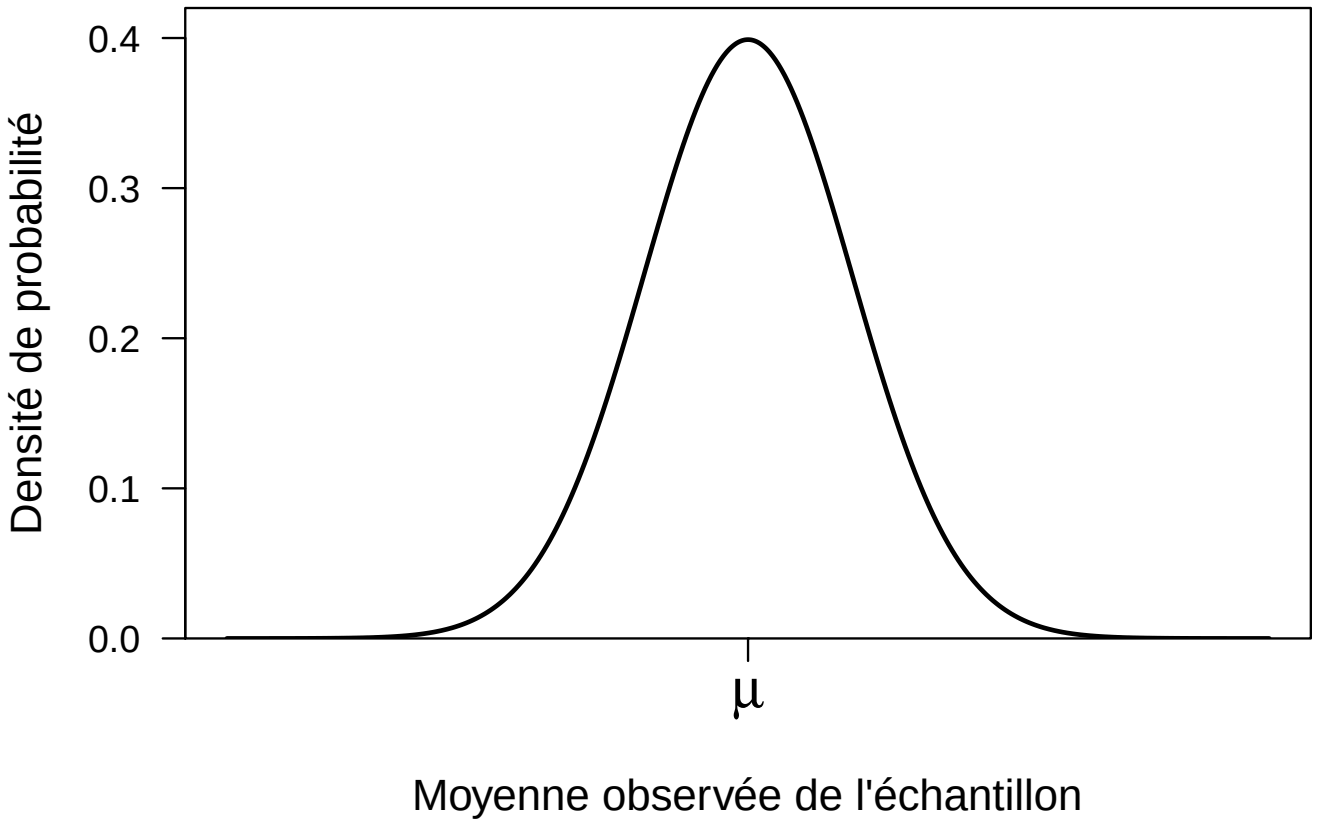
$$s^2 = \frac{1}{62}\left(17 \times 150^2 + 23 \times 172.5^2 + \dots\right) - 174.56^2$$
$$= 365.60$$

$$s = \sqrt{365.60} = 19.12 \text{ g.L}^{-1}$$

Loi de la moyenne de n v.a., n grand

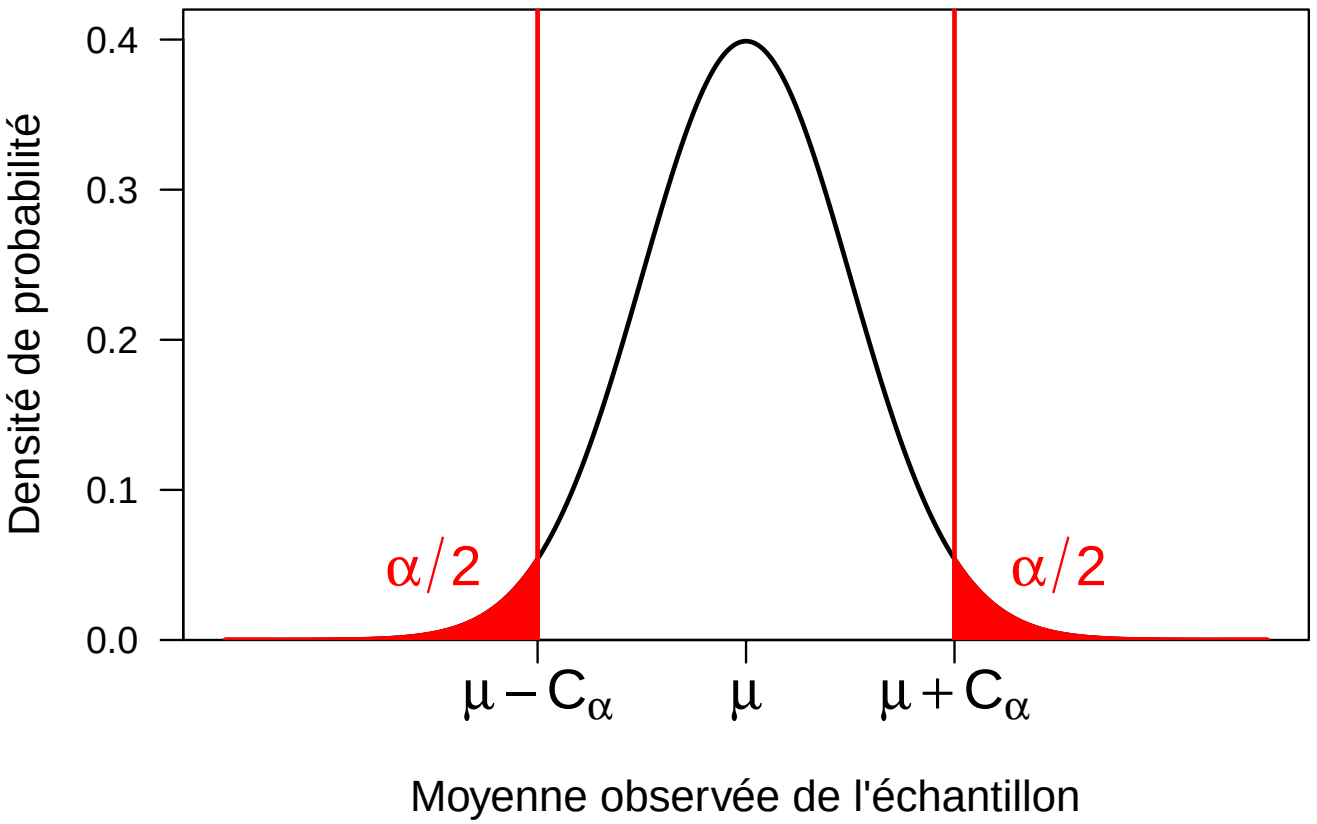


Distribution d'échantillonnage d'une moyenne observée



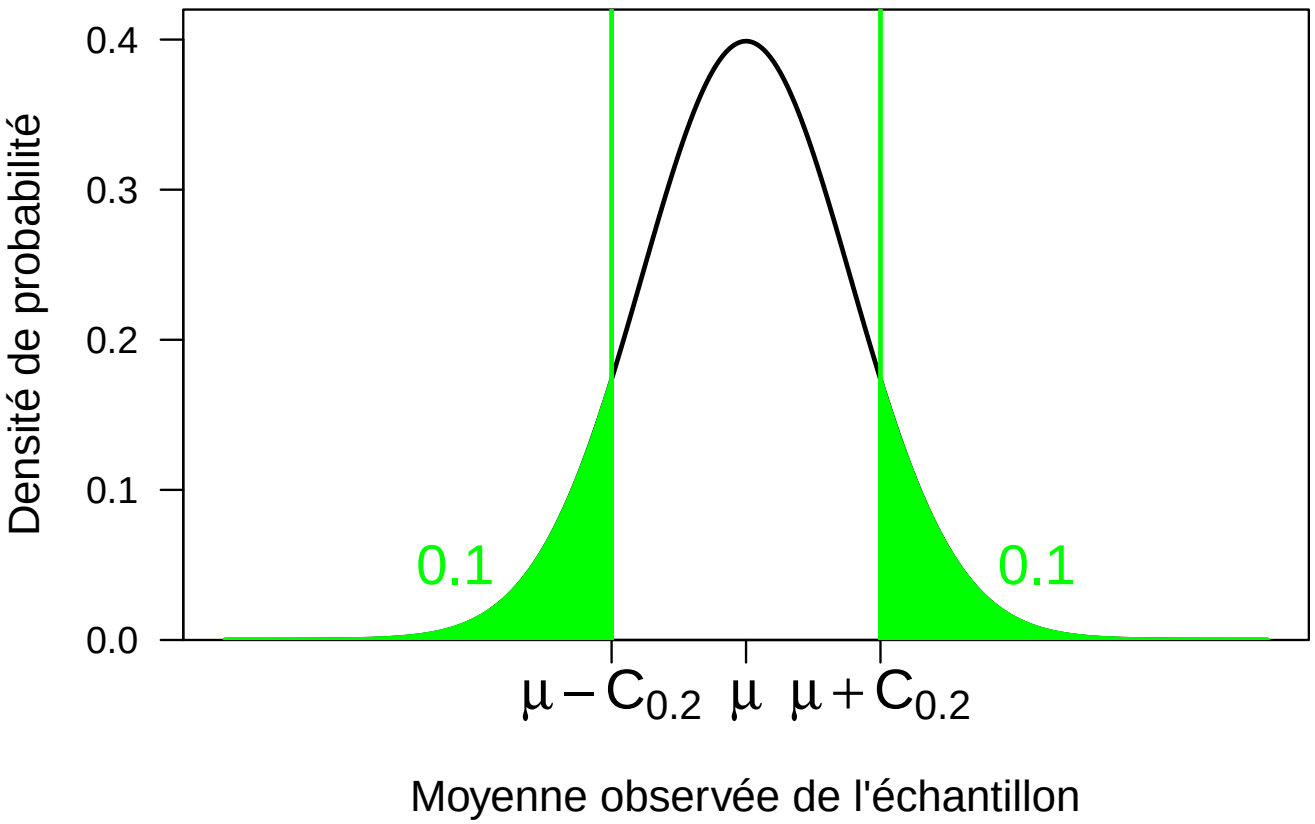
Quantiles de la loi normale

$$P(\mu - C_\alpha < \bar{x} < \mu + C_\alpha) = 1 - \alpha$$



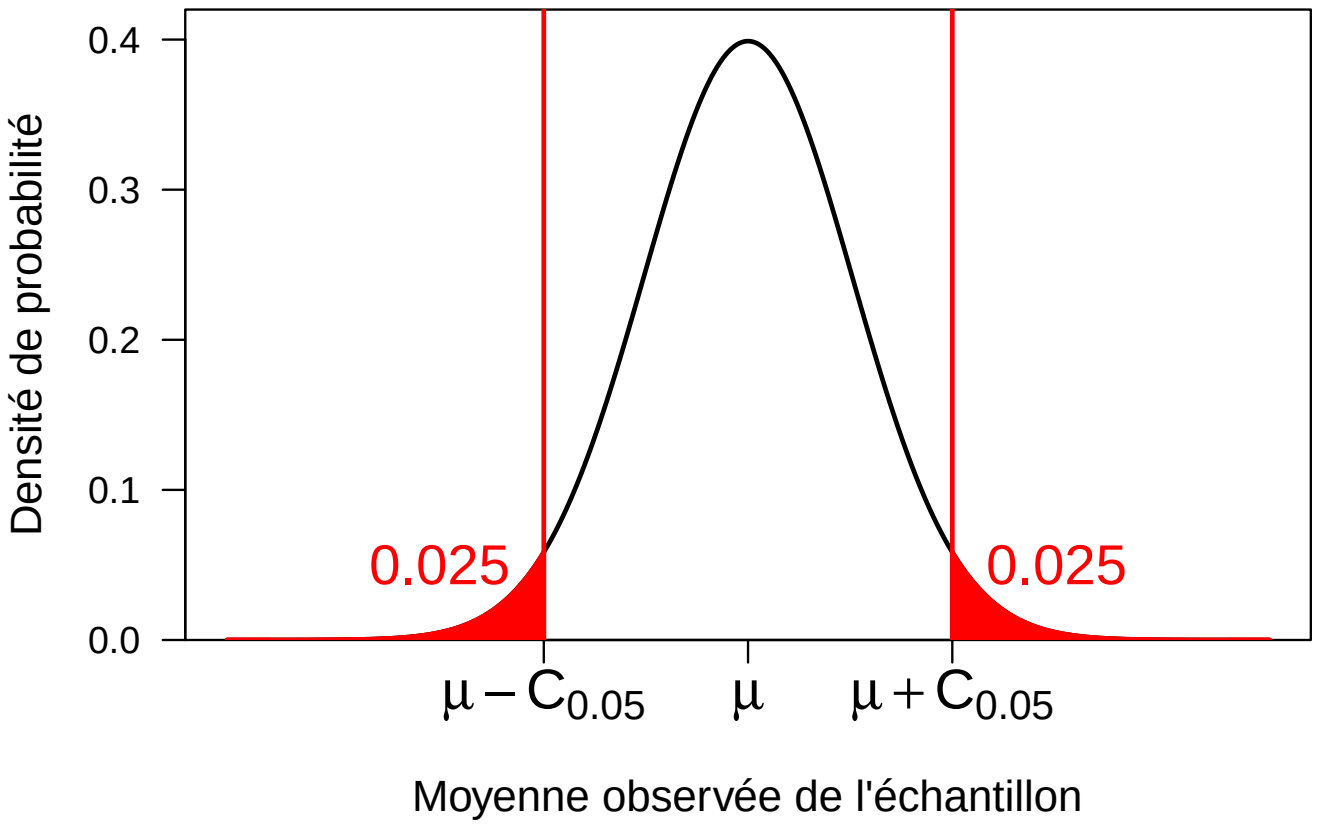
Quantiles de la loi normale, $\alpha = 0.20$

$$P(\mu - C_{0.20} < \bar{x} < \mu + C_{0.20}) = 0.80$$



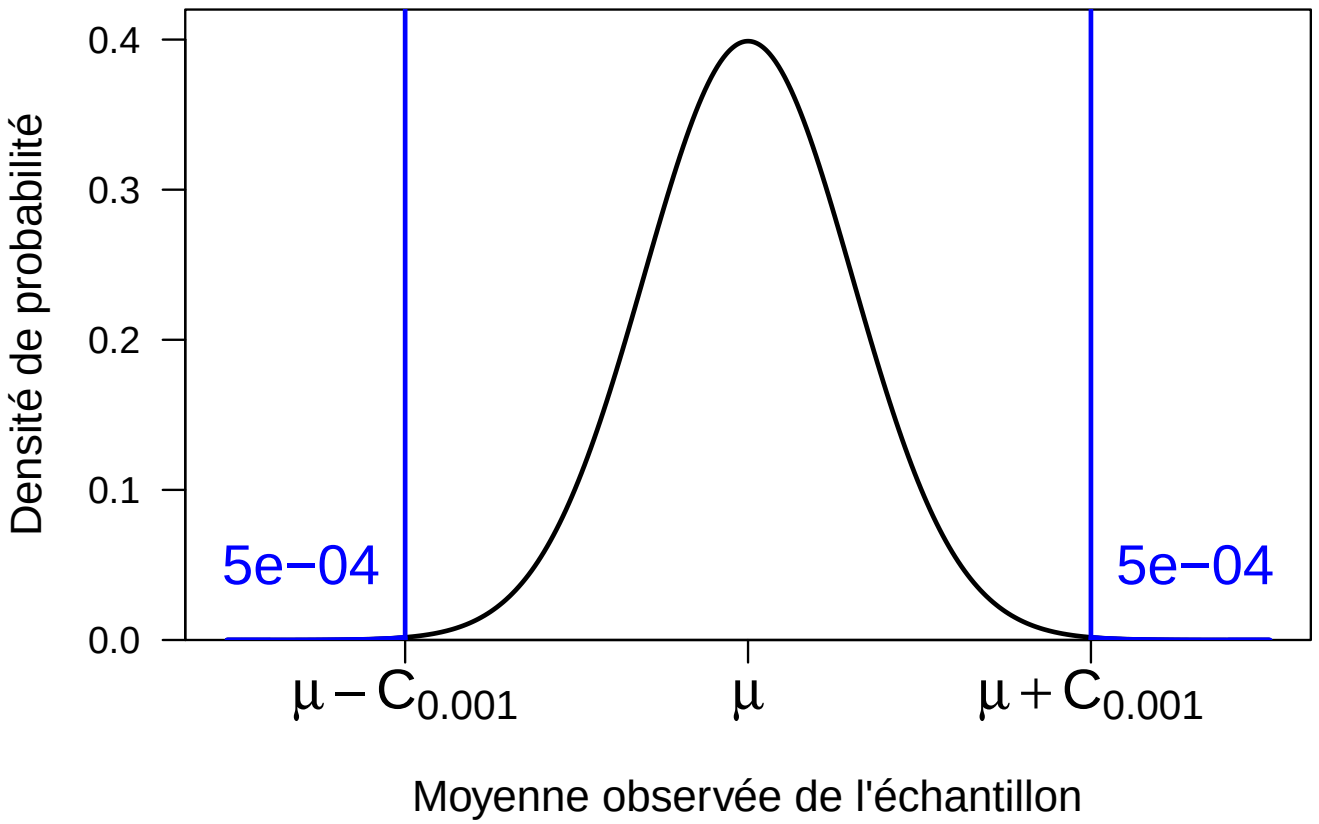
Quantiles de la loi normale, $\alpha = 0.05$

$$P(\mu - C_{0.05} < \bar{x} < \mu + C_{0.05}) = 0.95$$

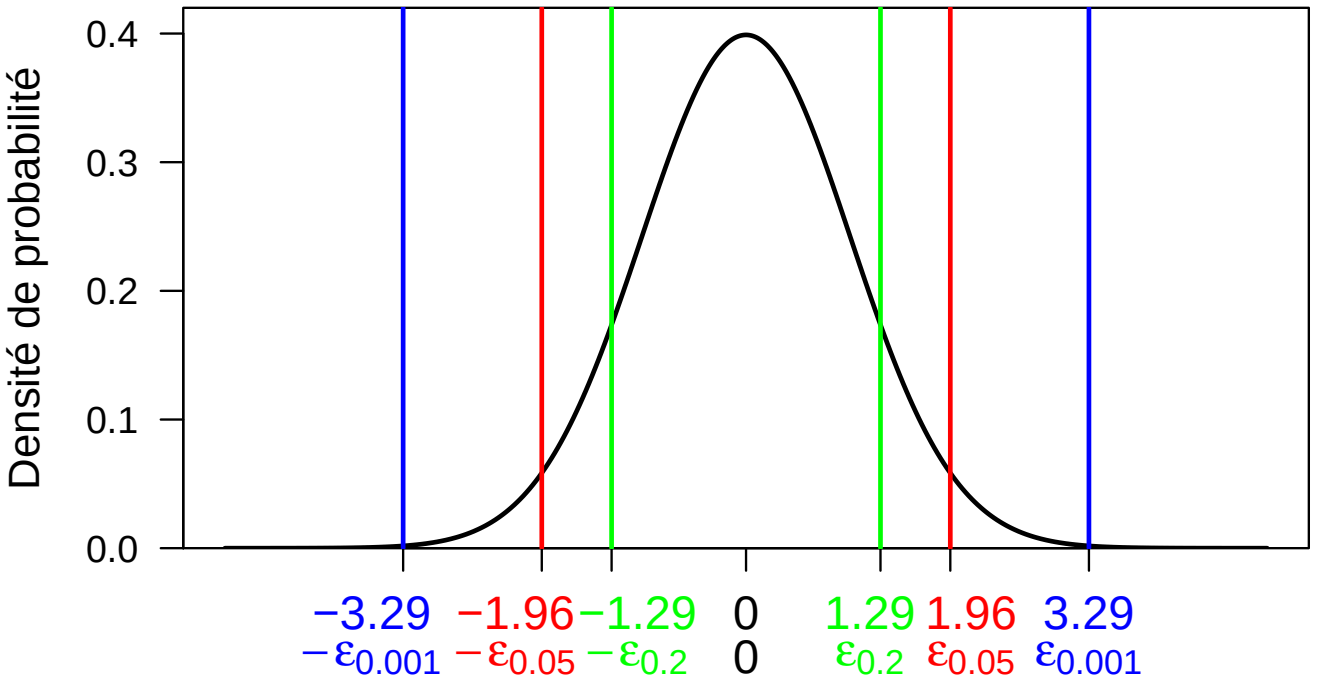


Quantiles de la loi normale, $\alpha = 0.001$

$$P(\mu - C_{0.001} < \bar{x} < \mu + C_{0.001}) = 0.999$$



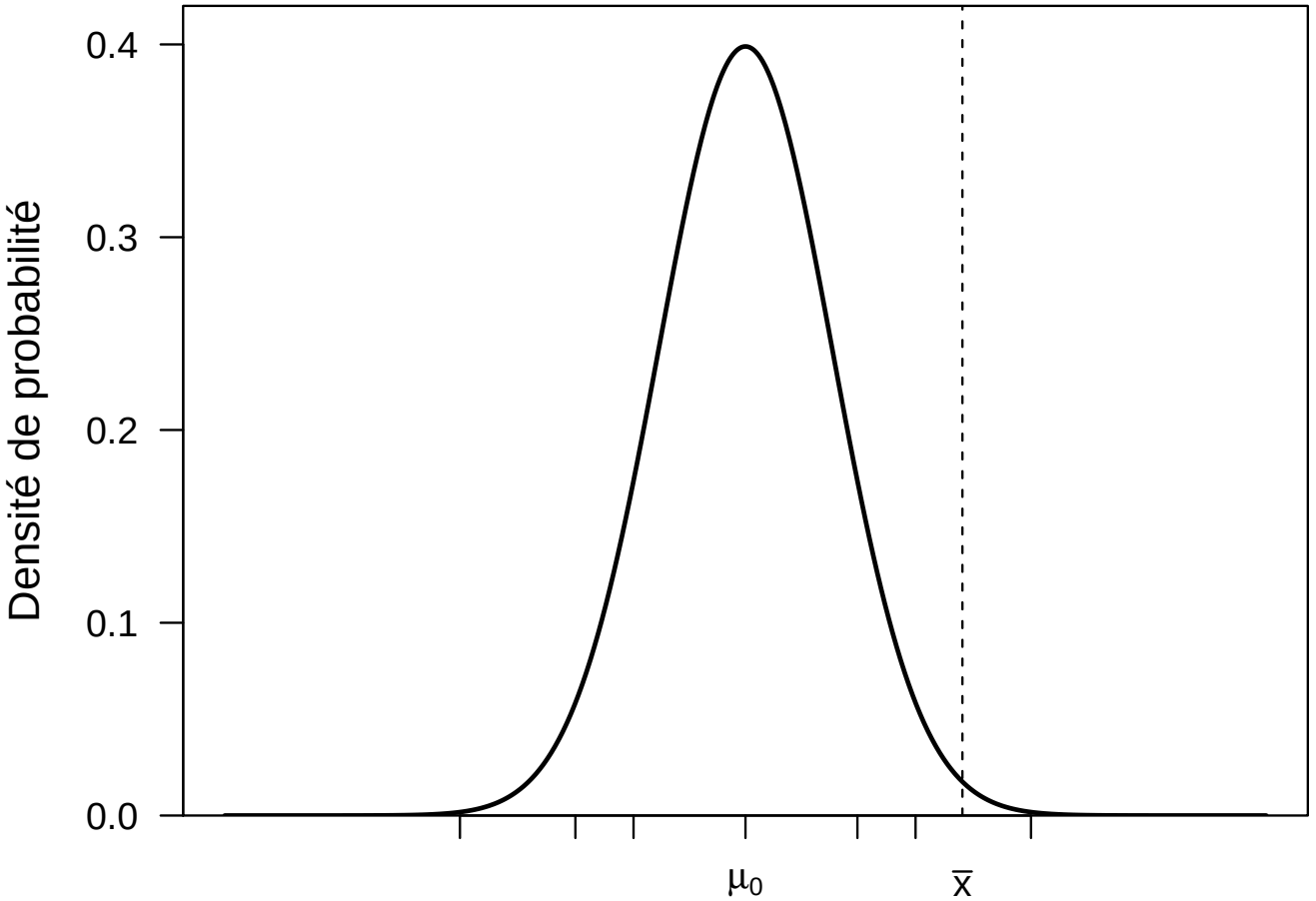
Quantiles de la loi normale centrée réduite



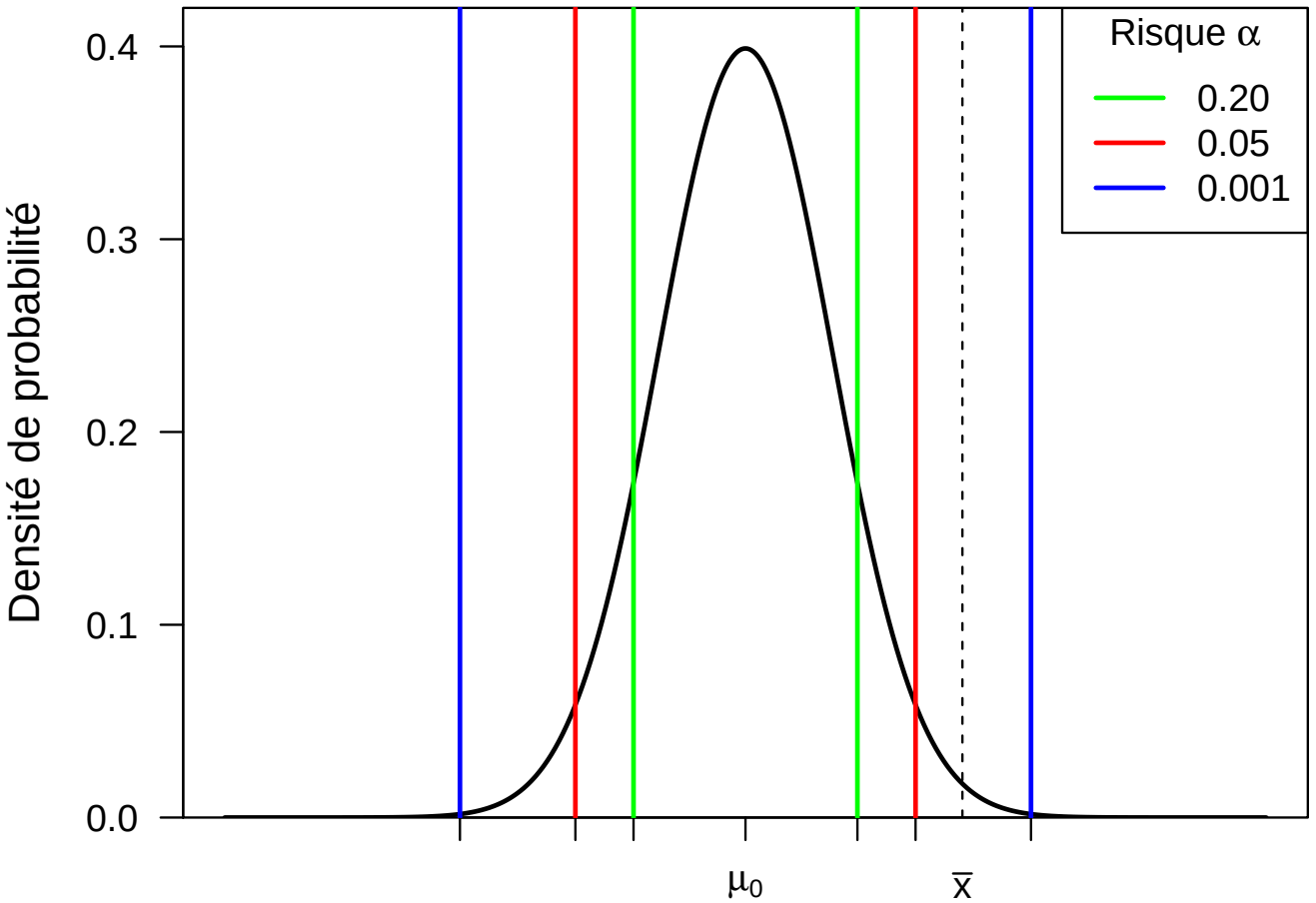
$$z = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

Table des matières

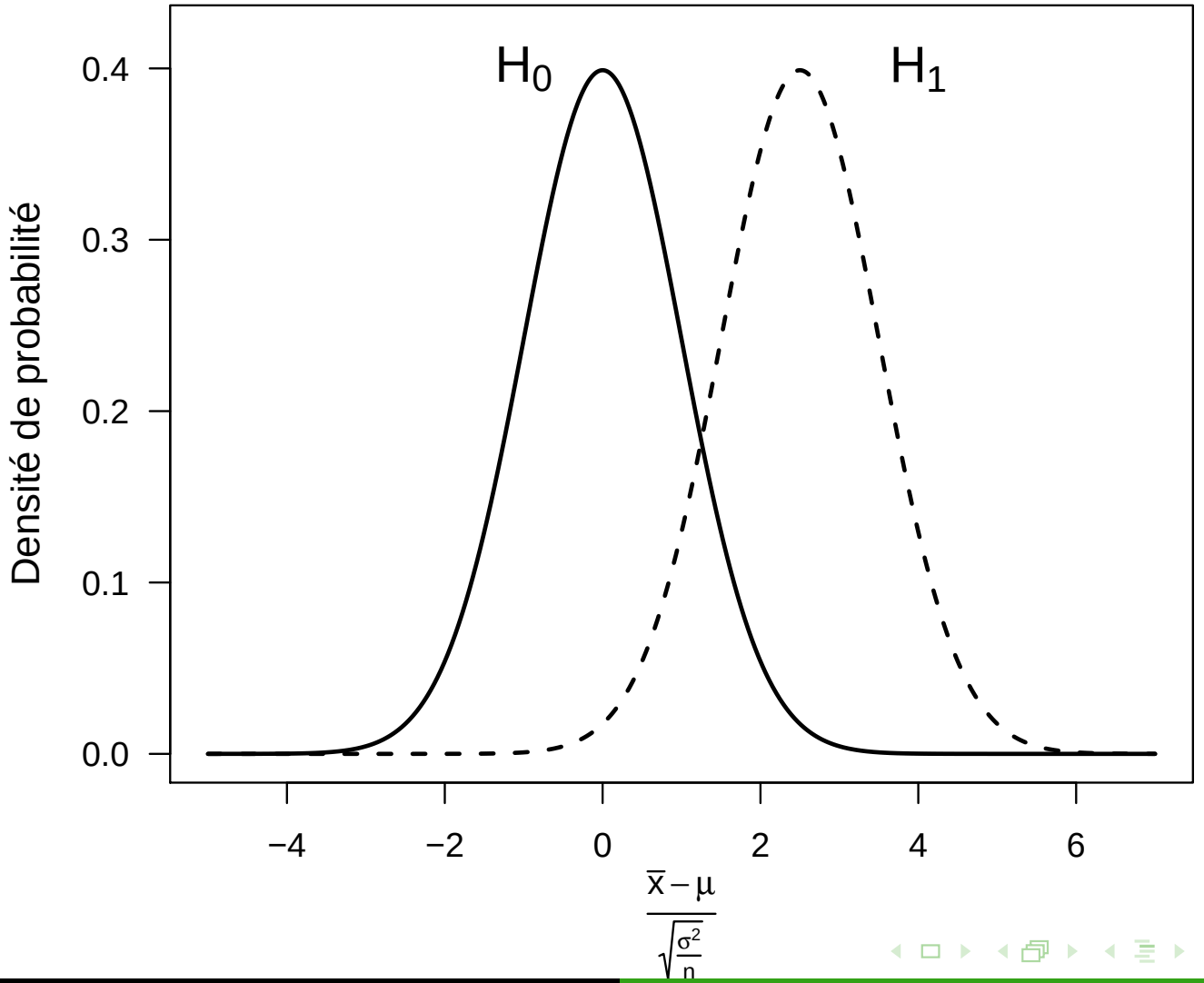
Distribution d'échantillonnage et moyenne observée



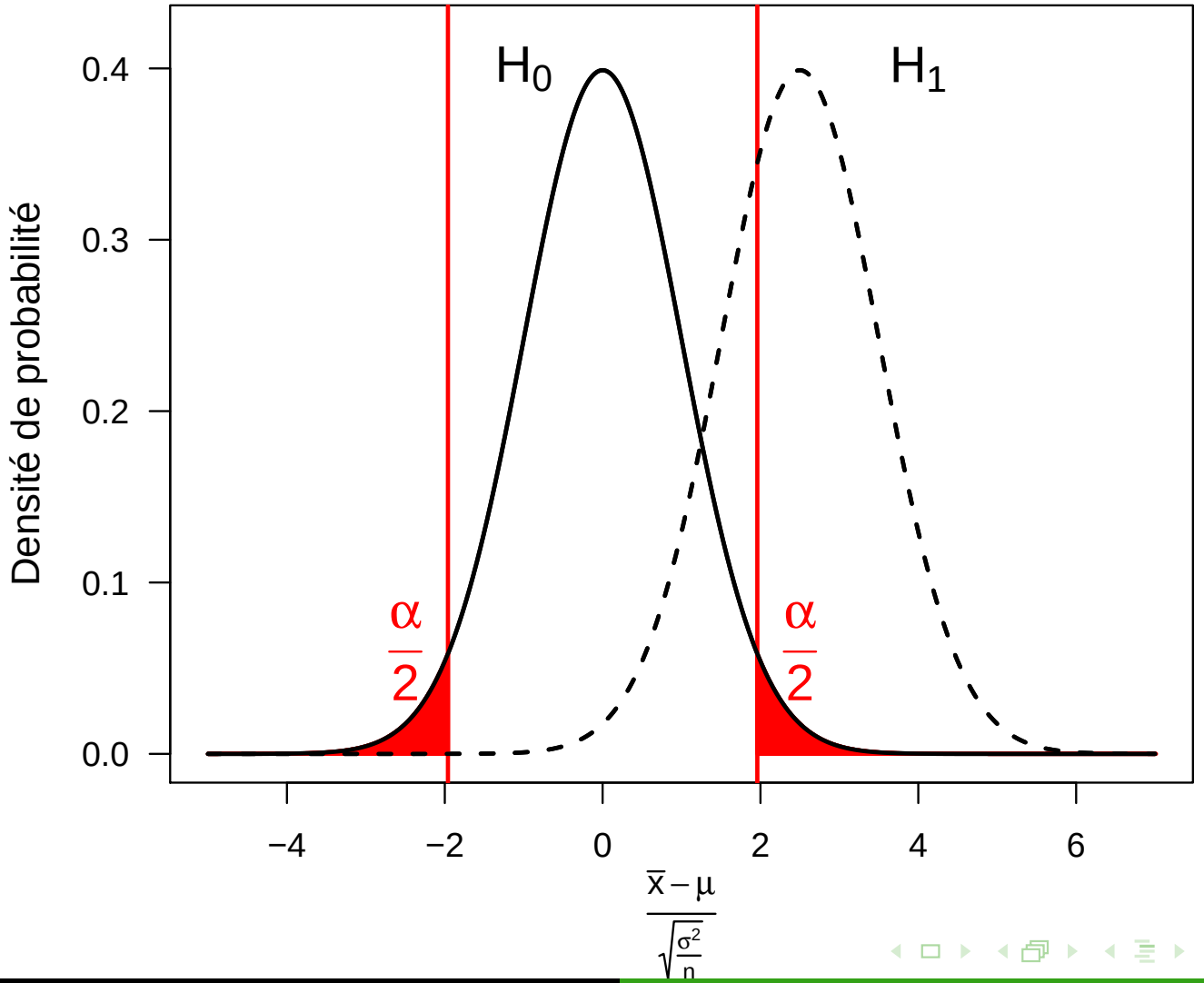
Distribution d'échantillonnage et moyenne observée



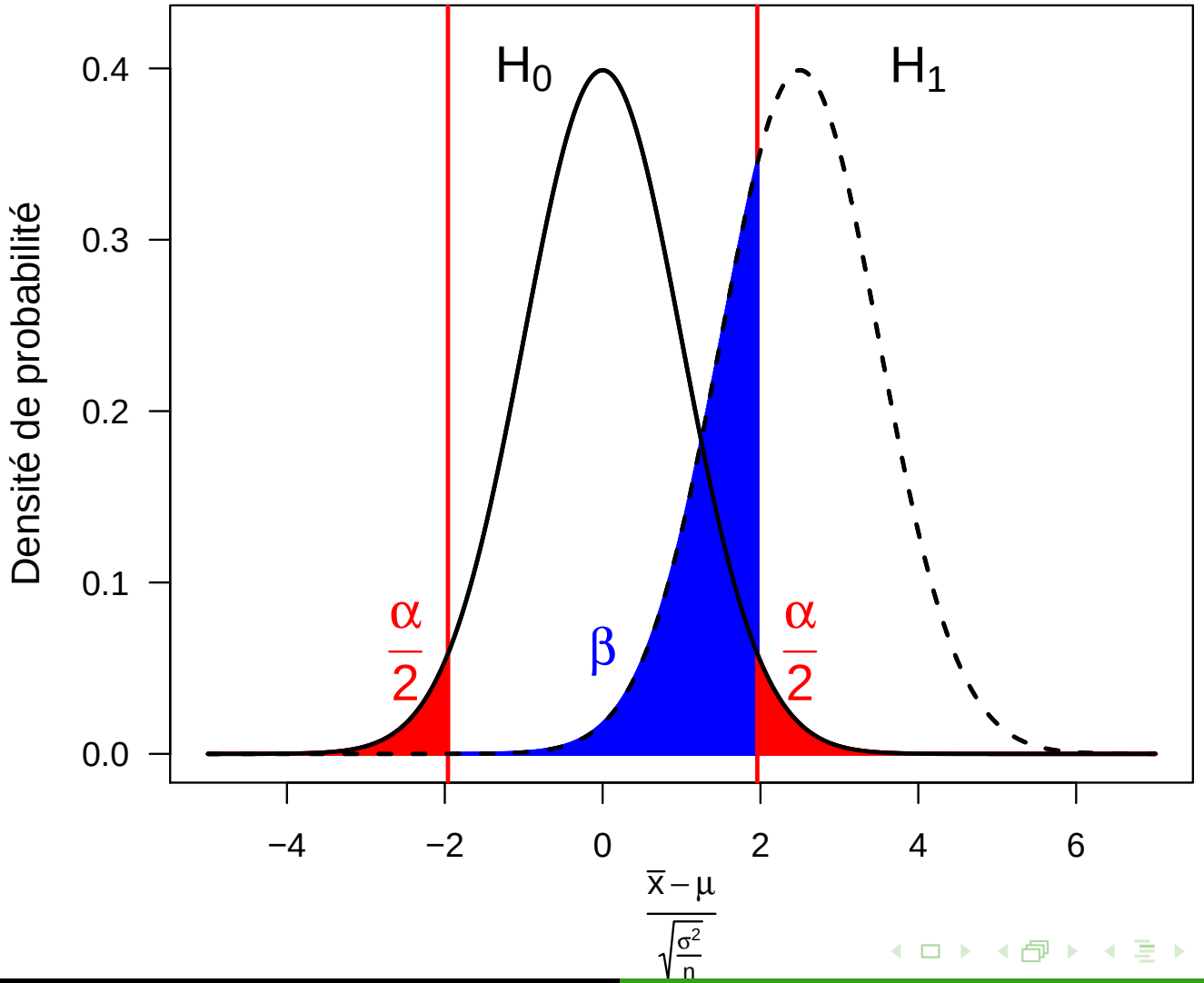
Risque de deuxième espèce



Risque de deuxième espèce

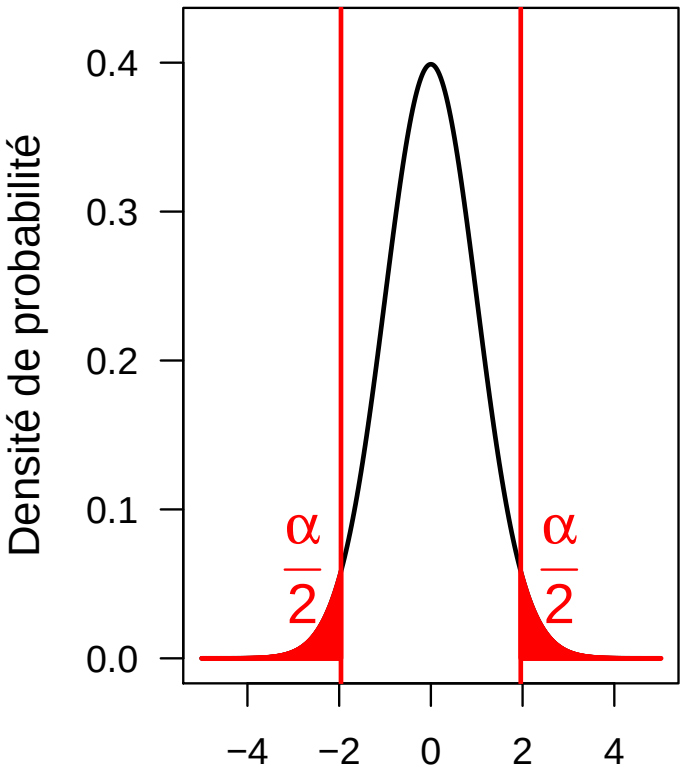


Risque de deuxième espèce



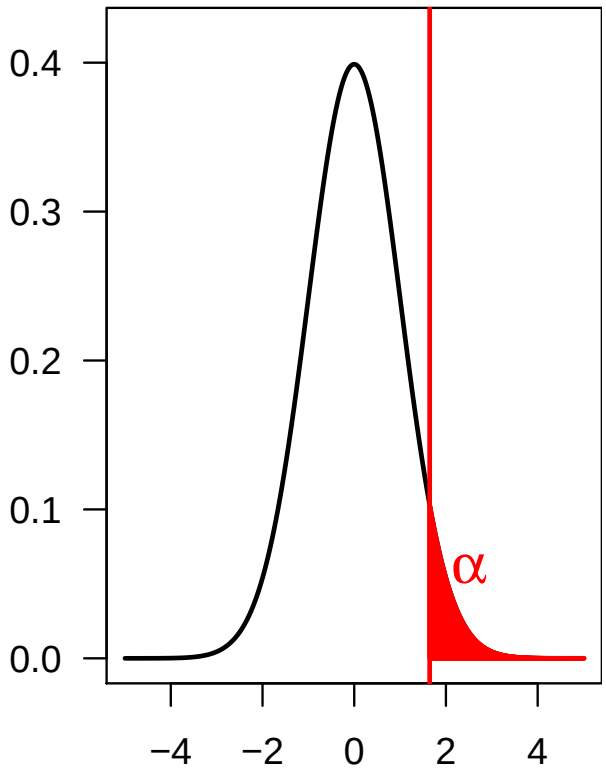
Test unilatéral, $\alpha = 5\%$

$H_1 : \mu \neq \mu_0$



$\varepsilon_{0.05} = 1.96$

$H_1 : \mu > \mu_0$



$\varepsilon_{0.1} = 1.645$