Éléments de compréhension des statistiques

Jeffery P.

Doctorant au Laboratoire des Sciences du Numérique de Nantes (LS2N)

2019

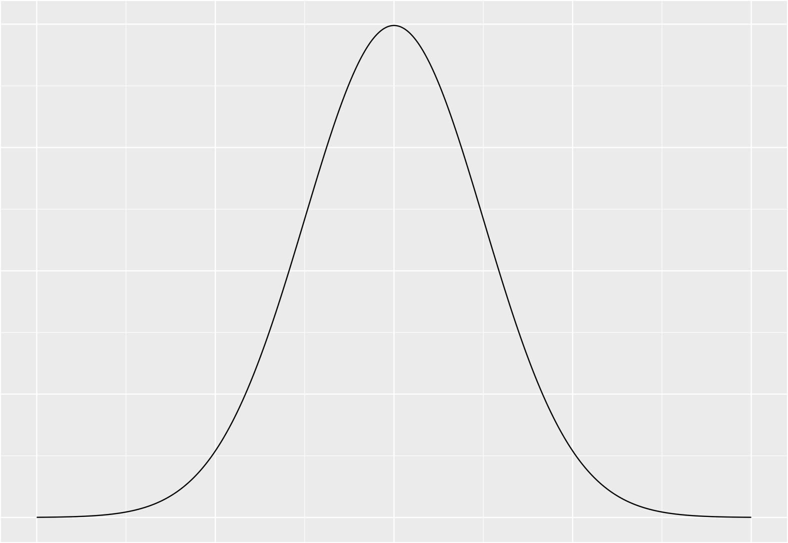
Crédits

- Ce support est inspiré du cours de M. Jean-Philippe Babin, responsable pédagogique de Licence à l'Université de Nantes - Laboratoire de Psychologie des Pays de la Loire.
- Également, plusieurs formulations ont pu être améliorées grâce aux commentaires avisés de M. Paul Marti et M. Damien Schnebelen, tous deux doctorants au LS2N - pôle SIEL, team PACCE.

La loi normale

Introduction

S'il y a bien une loi populaire en statistique, il s'agit de la loi normale. . .la célèbre courbe en cloche !

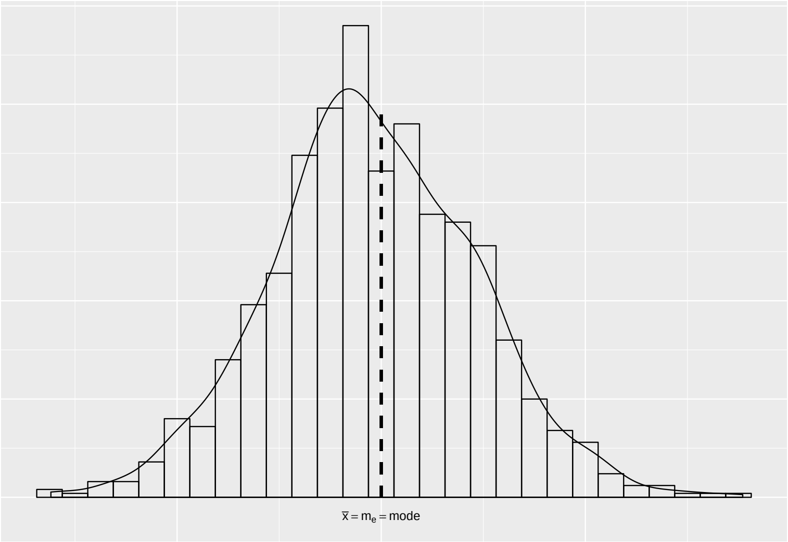


Caractéristiques de la loi normale

- La loi normale est une loi **symétrique, centrée autour de sa moyenne**
- La symétrie de la distribution implique que **la médiane est égale à la moyenne**
- C'est une loi unimodale, **son mode est égale à la moyenne**

Reconnaître une loi normale

En pratique, on peut supposer une distrbution normale d'après l'histogramme, ou le diagramme en barre



Quelques mots sur la loi normale

- La loi normale est définie par deux paramètres :

1. Sa moyenne souvent notée μ (où *mean*)
2. Son écart-type notée σ (où *sd* pour “Standard Deviation”)

Sa densité de probabilité est la suivante :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

avec une variable X qui suit une loi normale de moyenne μ et d’écart-type σ

On note $X \sim \mathrm{N}(\mu, \sigma^2)$

Illustration de la loi normale

Changer la moyenne d'une loi normale revient à translater la distribution vers la droite ou la gauche

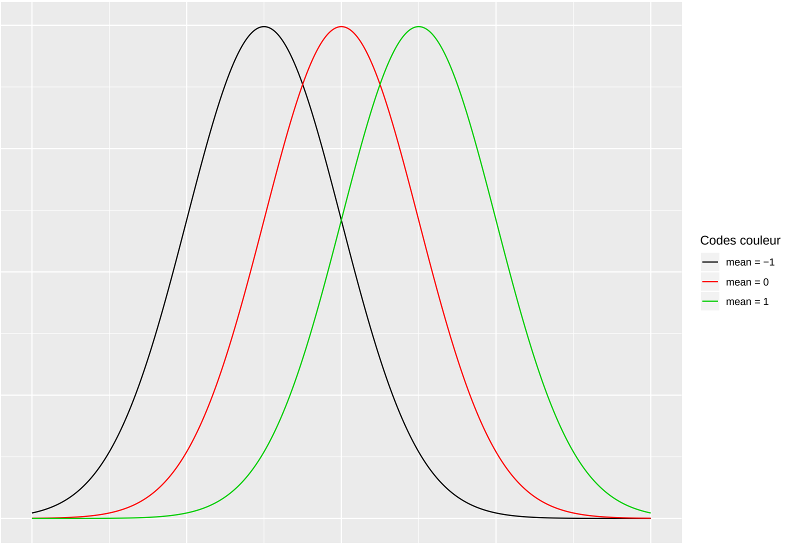
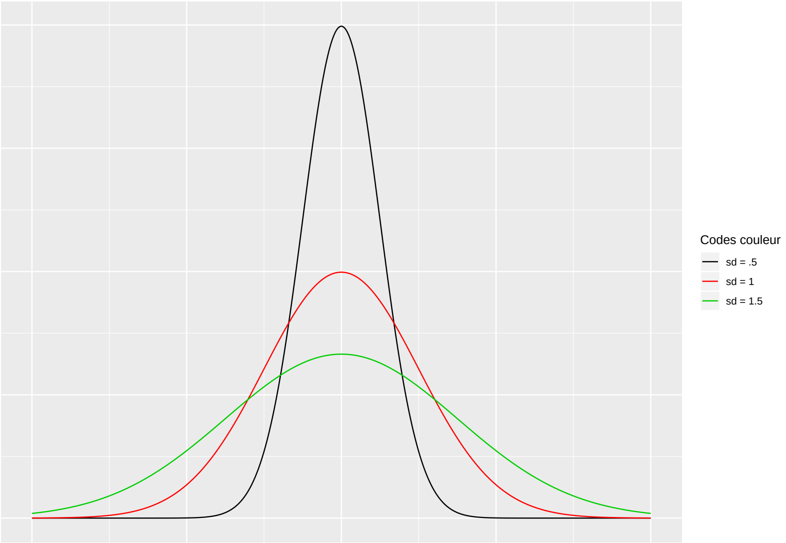


Illustration de la loi normale

Changer l'écart-type d'une loi normale revient à aplatir ou resserrer sa distribution autour de sa moyenne



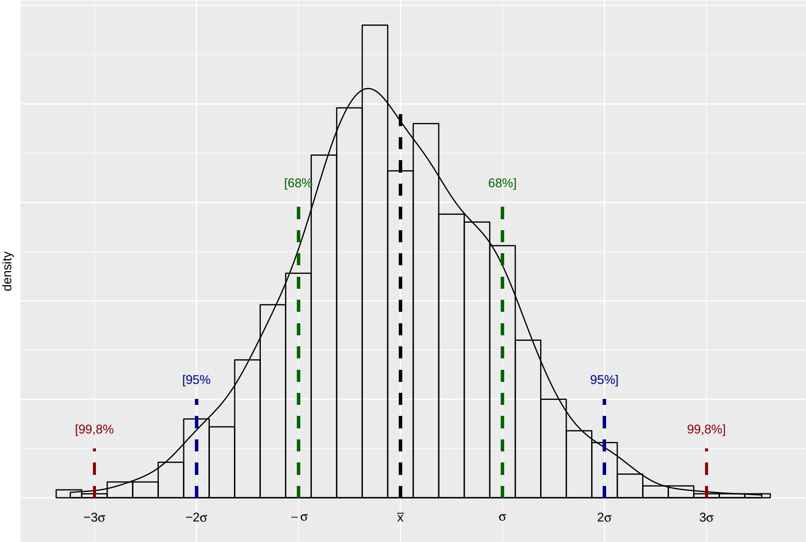
Propriété intéressante

Si l'on dispose d'observations d'une distribution N(μ, σ^2) alors :

- 68% des observations sont comprises dans l'intervalle $[\mu - \sigma; \mu + \sigma]$
- 95% des observations sont comprises dans l'intervalle $[\mu - 2\sigma; \mu + 2\sigma]$
- 99,8% des observations sont comprises dans l'intervalle $[\mu - 3\sigma; \mu + 3\sigma]$

Estimation graphique de μ et σ

- En pratique on estime μ en prenant le centre de la distribution, et on déduit σ en estimant l'intervalle vert



Calcul de probabilité : généralité

La probabilité pour qu'une variable aléatoire X soit inférieure à une quelconque valeur x s'écrit $P(X \leq x)$

- **Un probabilité est toujours positive et inférieure à 1**

Remarque

- Pour une variable aléatoire réelle on a :
  $P(-\infty < X < +\infty) = 1$
- Pour une variable continue, on a $P(X = x) = 0$, d'où :

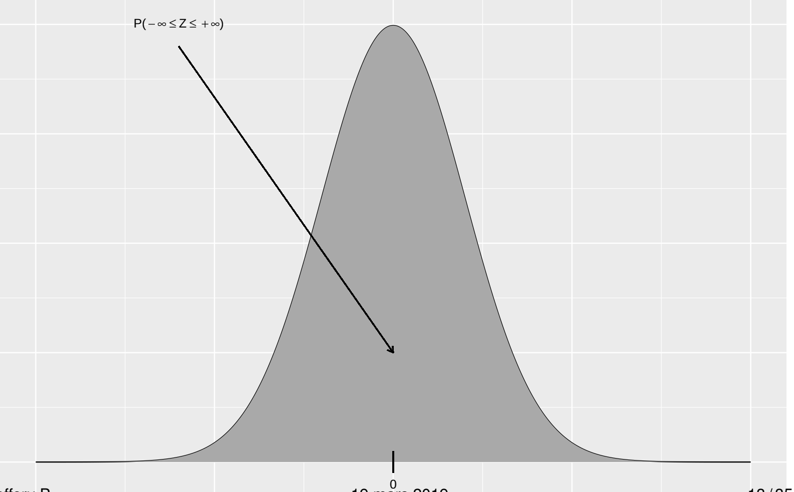
$$P(X \leq x) = P(X < x) \text{ où encore } P(X \geq x) = P(X > x)$$

Ce résultat s'explique avec un peu de théorie mathématique que l'on ne détaillera pas ici, mais il peut être utile d'avoir ces propriétés en tête pour les exercices. . .

Calcul de probabilité : lien avec l'aire sous la courbe de densité

Cas 0 : soit $Z \sim N(0, 1)$, l'aire sous la courbe est égale à 1

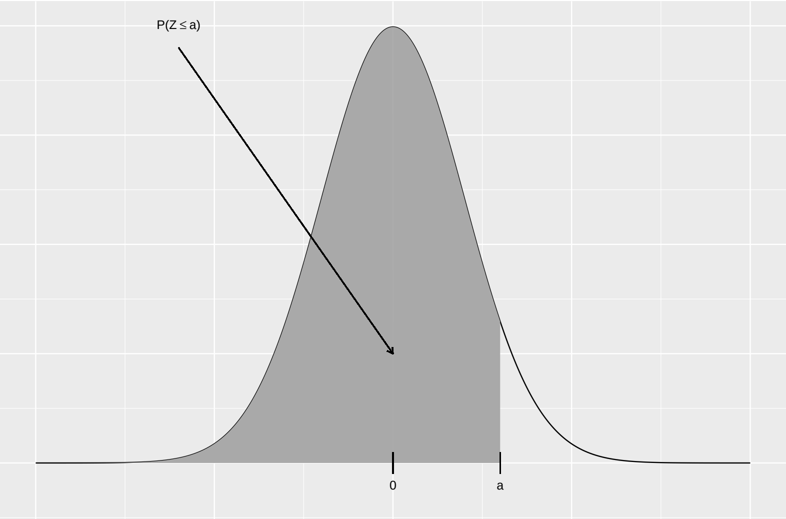
$$P(-\infty < Z < +\infty) = 1$$



Calcul de probabilité : valeurs de la table

Cas I : soient $Z \sim N(0,1)$ et a un nombre réel **positif**

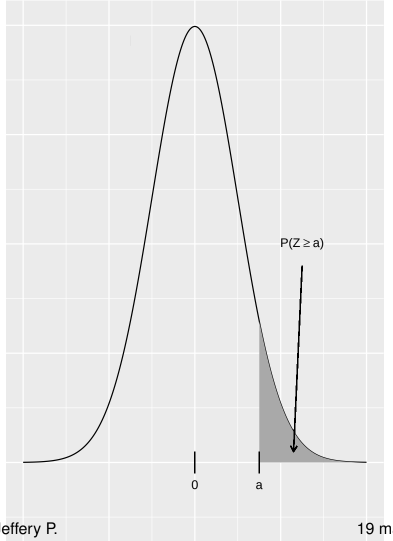
$$P(Z \leq a) \longrightarrow \text{le résultat se trouve dans la table !}$$

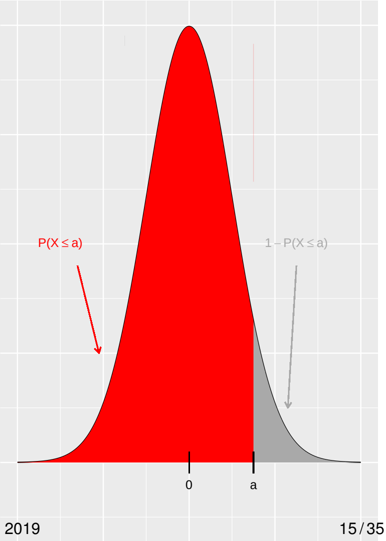


Calcul de probabilité avec la loi normale

Cas II : soient $Z \sim N(0,1)$ et a un nombre réel **positif**

$$P(Z \geq a) = 1 - P(Z \leq a) \longrightarrow \text{on se ramène au } \textbf{cas I}$$

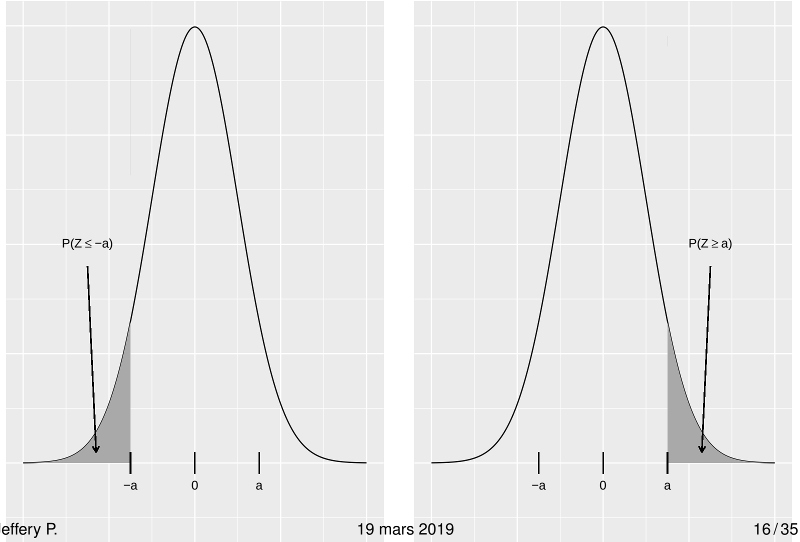




Calcul de probabilité avec la loi normale

Cas III : soient $Z \sim N(0,1)$ et a un nombre réel **positif**

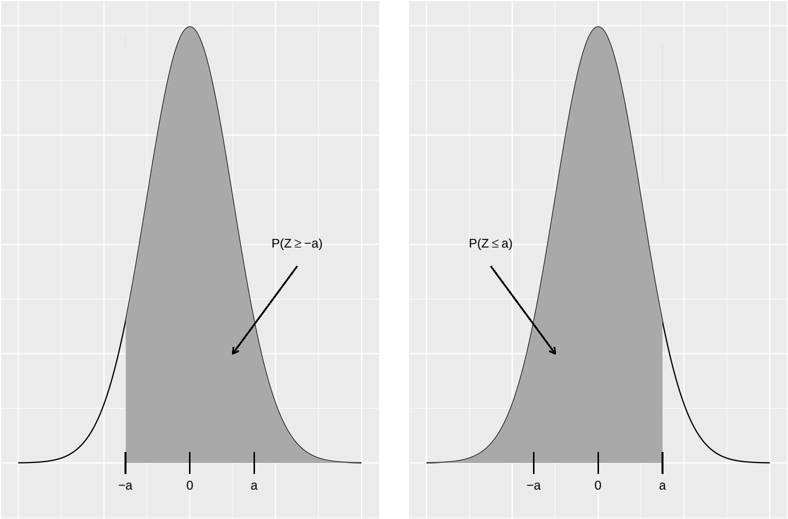
$$P(Z \leq -a) = P(Z \geq a) \longrightarrow \text{on se ramène au } \textbf{cas II}$$



Calcul de probabilité avec la loi normale

Cas IV : soient $Z \sim N(0,1)$ et a un nombre réel **positif**

$$P(Z \geq -a) = P(Z \leq a) \longrightarrow \text{on se ramène au } \textbf{cas I}$$



Résumé

[Calcul de probabilité *(Extrait du cours de M. Gérin, Paris Ouest 2012-2013)*][normale]

Formule de calcul avec une intervalle quelconque

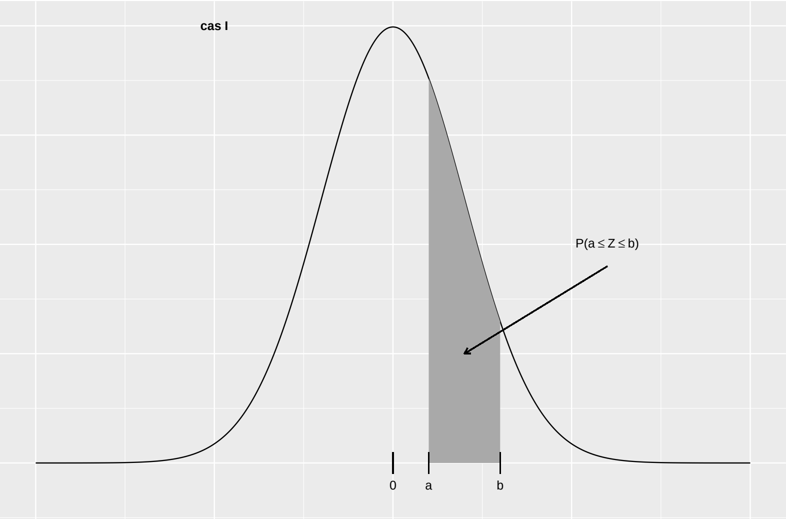
Soient $Z \sim N(0,1)$ et u, v deux nombres réels avec $u \leq v$

$$P(u \leq Z \leq v) = P(Z \leq v) - P(Z \leq u)$$

Calcul de probabilité avec la loi normale

Cas V.1 : soient $Z \sim N(0, 1)$ et $0 \leq a \leq b$

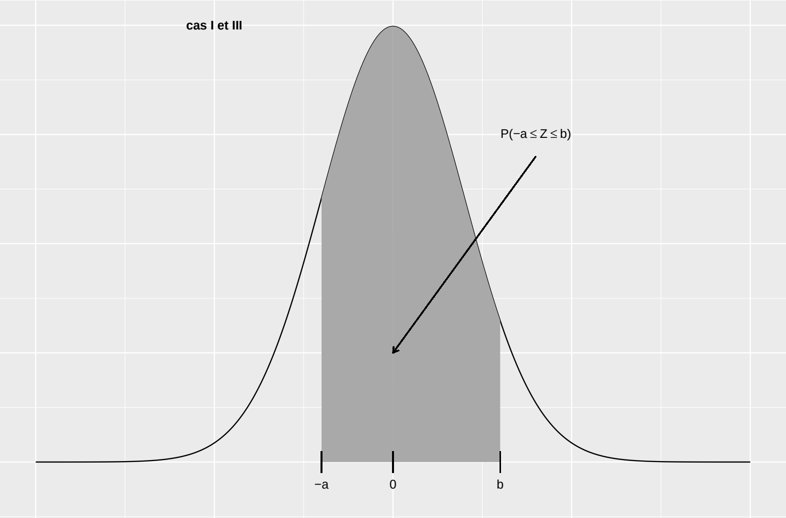
$$P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$$



Calcul de probabilité avec la loi normale

Cas V.2 : soient $Z \sim N(0,1)$ et $0 \leq a \leq b$

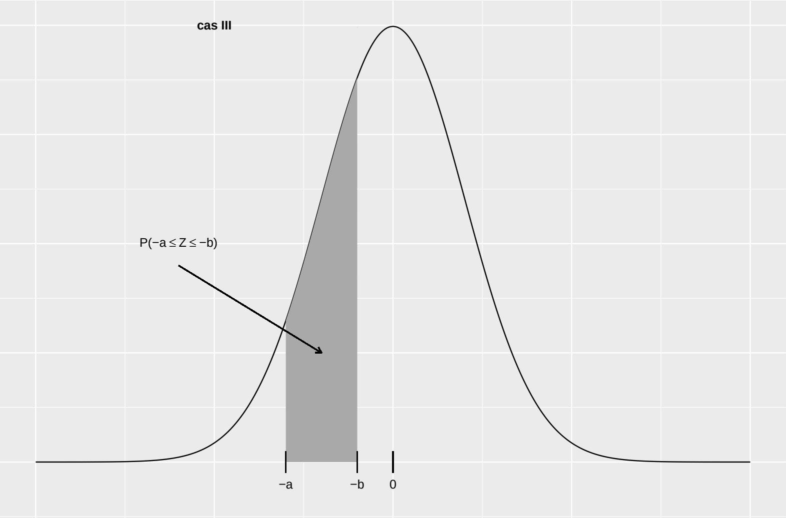
$$P(-a \leq Z \leq b) = P(Z \leq b) - P(Z \leq -a) = P(Z \leq b) + P(Z \leq a) - 1$$



Calcul de probabilité avec la loi normale

Cas V.3 : soient $Z \sim N(0,1)$ et $-a \leq -b \leq 0$

$$P(-a \leq Z \leq -b) = P(Z \leq -b) - P(Z \leq -a) = P(Z \leq a) - P(Z \leq b)$$



Transformation de la loi normale

Toute transformation *affine* d'une loi normale est encore une loi normale i.e., quelque soit les nombre réels a et $b \neq 0$ on a :

$$X \sim \mathrm{N}(\mu, \sigma^2) \Rightarrow aX + b \sim \mathrm{N}(\mu + b, a^2 \times \sigma^2)$$

Par conséquent :

- si $X \sim \mathrm{N}(\mu, \sigma^2)$ alors $X - \mu \sim \mathrm{N}(0, \sigma^2)$ (**centrage**)
- si $X \sim \mathrm{N}(\mu, \sigma^2)$ alors $\frac{X}{\sigma} \sim \mathrm{N}(\mu, 1)$ (**réduction**)

En pratique, on fait souvent les deux en même temps :

- si $X \sim \mathrm{N}(\mu, \sigma^2)$ alors $\frac{X-\mu}{\sigma} \sim \mathrm{N}(0, 1)$ (**normalisation=centrage+réduction**)

Calcul de probabilité : en pratique

Si la variable X ne suite pas une loi normale centrée réduite, on peut toujours se ramener aux cas précédents, par exemple :

$$P(X \leq x) = P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = P(Z \leq \frac{x-\mu}{\sigma})$$

Où $Z \sim \mathrm{N}(0, 1)$

$\rightarrow$ Rappel : il s'agit de la **normalisation**, cette petite transformation est très utile et beaucoup utilisée !

Cas pratique : calcul de probabilité

68% des observations sont comprises dans l'intervalle $[\bar{x} - \sigma; \bar{x} + \sigma]$... Est-ce bien vrai ?

$$\begin{aligned} P(\bar{x} - \sigma \leq X \leq \bar{x} + \sigma) &= P(-\sigma \leq X - \bar{x} \leq \sigma), \quad \text{on centre} \\ &= P(-1 \leq \tfrac{X-\bar{x}}{\sigma} \leq 1), \quad \text{on réduit} \\ &= P(-1 \leq Z \leq 1) \end{aligned}$$

Où $Z \sim N(0, 1)$

Il faut maintenant trouver la valeur de cette probabilité sur les tables...

Cas pratique

$$
\begin{aligned}
P(\bar{x} - \sigma \leq X \leq \bar{x} + \sigma) &= P(-1 \leq Z \leq 1) \\
&= P(Z \leq 1) - P(Z \leq -1) \\
&= P(Z \leq 1) - (1 - P(Z \leq 1)) \\
&= 2P(Z \leq 1) - 1
\end{aligned}
$$

Où $Z \sim N(0, 1)$

Remarque

Plus généralement, on a la formule suivante quelque soit k nombre entier positif :

$$P(\bar{x} - k \times \sigma \leq X \leq \bar{x} + k \times \sigma) = P(-k \leq Z \leq k)$$

Calcul des quantiles

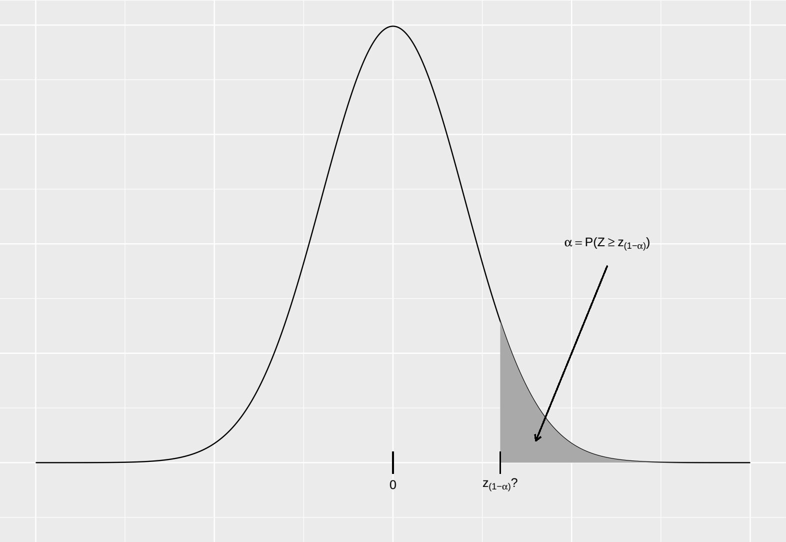
Dans certains cas, on ne cherche pas à calculer une probabilité pour un b donné (e.g., $P(Z \leq b)$) mais on cherche b telle que la probabilité soit égale à une certaine valeur :

On veut trouver b tel que $P(X \geq b) = \alpha$ où $P(|X| \geq b) = \alpha$

- Dans ce cas, on s'assure que la probabibilité concerne une variable suivant une **loi normale centrée réduite**, puis on regarde dans une des tables

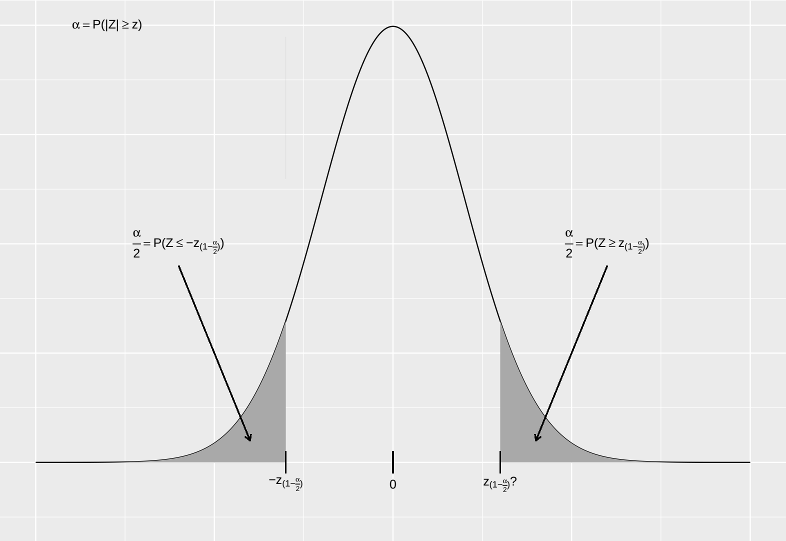
Calcul des quantiles : table 1

On recherche une valeur z telle que $P(Z > z)$ avec $Z \sim N(0, 1)$:



Calcul des quantiles : table 2

On recherche une une valeur z telle que $P(|Z| > z)$ avec $Z \sim N(0, 1)$:



Calcul des quantiles

Attention, cette problématique est faussement facile...dans la majorité des cas elle nécessite une bonne maîtrise du calcul des probabilités de la loi normale !

Cas fréquents

- On ne demande pas de trouver z tel que $P(Z \geq z) = \alpha$ mais plutôt $P(Z \leq z) = 1 - \alpha$.
- On ne demande pas de trouver z tel que $P(|Z| \geq z) = \alpha$ mais plutôt $P(|Z| \leq z) = 1 - \alpha$.
- La probabilité porte sur une variable $X \sim N(\mu, \sigma^2)$. Il faut bien penser à centrer et réduire l'évènement (ce qui se trouve dans la probabilité) e.g., $P(X > x) = P(Z > \frac{x-\mu}{\sigma})$ où et $Z \sim N(0, 1)$

Calcul des quantiles

Exemple général (théorique mais utile !)

Soit $X \sim N(\mu, \sigma)$, on suppose $\alpha \in]0; 0.5[$ connu.

On cherche b **tel que** $P(X > b) = \alpha$

Voici les étapes qui nous permettent de trouver la solution :

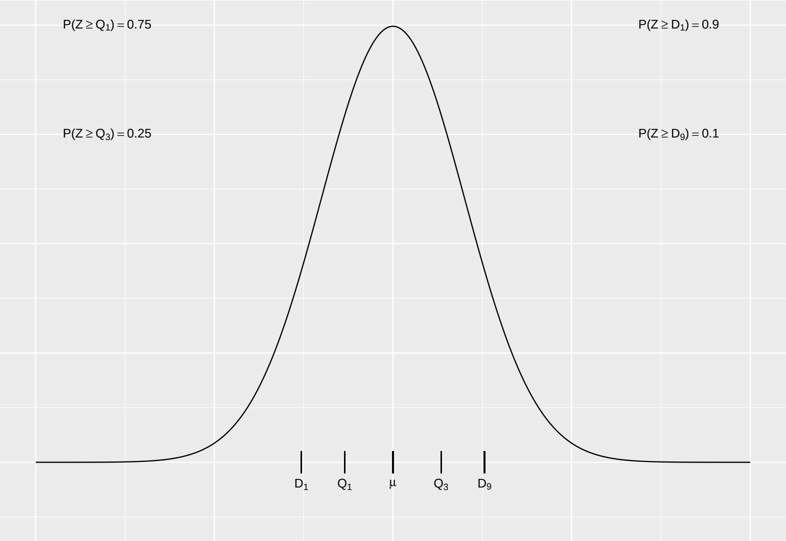
$$\begin{array}{rrcl} & P(X > b) & = & \alpha \\ \Leftrightarrow & P(Z > \frac{b-\mu}{\sigma}) & = & \alpha \\ \Leftrightarrow & P(Z \leq \frac{b-\mu}{\sigma}) & = & 1 - \alpha \\ \Rightarrow & \frac{b-\mu}{\sigma} & = & z_{(1-\alpha)} \\ & \text{Donc} \quad b & = & \mu + \sigma \times z_{(1-\alpha)} \end{array}$$

Avec :

- $Z \sim N(0, 1)$
- $z_{(1-\alpha)}$ est le **quantile** d'ordre $(1 - \alpha)$ de la loi normale centrée réduite

Calcul des quantiles

- Rappel sur la position des déciles et quartiles extrêmes



Calcul des quantiles

Soit $X \sim N(\mu, \sigma)$

Détermination des déciles et quartiles extrêmes

En utilisant la formule précédente, il vient :

$$\begin{array}{rrcl} & P(X > Q_3) & = & 0.25 \\ \Rightarrow & Q_3 & = & \mu + \sigma \times z_{(0.75)} \end{array}$$

$$\begin{array}{rrcl} & P(X > D_9) & = & 0.1 \\ \Rightarrow & D_9 & = & \mu + \sigma \times z_{(0.9)} \end{array}$$

Pour rappel :

- Q_3 : troisième quartile
- D_9 : neuvième décile

Calcul des quantiles

Soit $X \sim N(\mu, \sigma)$

Détermination des déciles et quartiles extrêmes

Pour D_1 et Q_1, le calcul est un peu **sioux.** Par exemple pour D_1, on a :

$$\begin{array}{rrcl} & P(X > D_1) & = & 0.9 \\ \Leftrightarrow & P(Z > \frac{D_1-\mu}{\sigma}) & = & 0.9 \quad \text{or } \frac{D_1-\mu}{\sigma} < 0 \\ \Leftrightarrow & P(Z \leq -\frac{D_1-\mu}{\sigma}) & = & 0.9 \\ \Rightarrow & -\frac{D_1-\mu}{\sigma} & = & z_{(0.9)} \\ & \text{Donc} \quad D_1 & = & \mu - \sigma \times z_{(0.9)} \end{array}$$

Calcul des quantiles

Soit $X \sim N(\mu, \sigma)$

Détermination des déciles et quartiles extrêmes

On peut utiliser les formules suivantes pour D_1 et Q_1 :

$$Q_1 = \mu - \sigma \times z_{(0.75)} = 2 \times \mu - Q_3$$

$$D_1 = \mu - \sigma \times z_{(0.9)} = 2 \times \mu - D_9$$

Avec :

- Q_1 : premier quartile i.e., $P(X \geq Q_1) = 0.75$
- Q_3 : troisième quartile i.e., $P(X \geq Q_3) = 0.25$
- D_1 : premier décile i.e., $P(X \geq D_1) = 0.9$
- D_9 : neuvième décile i.e., $P(X \geq D_9) = 0.1$