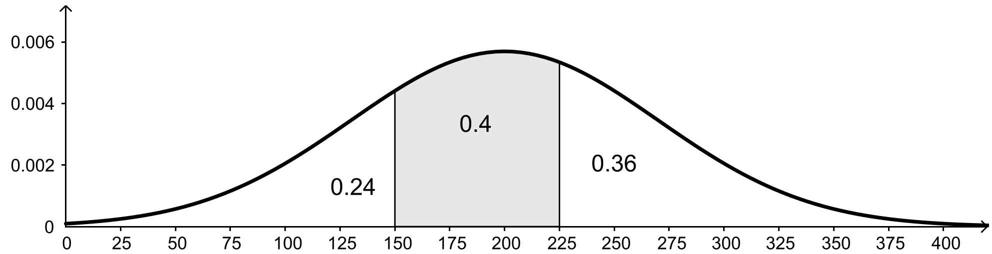
Loi normale et échantillonnage

1. Loi normale

❖ **Exemple introductif**

Une entreprise produit des ampoules. Soit X la variable qui à chaque ampoule associe sa durée de vie en heures. Cette durée n'est pas nécessairement un nombre entier d'heures et théoriquement on ne connaît pas la durée maximale de vie d'une telle ampoule. La variable aléatoire X est donc continue à valeurs dans $[0; +\infty[$.

La fonction f représentée ci-dessous a une aire sous la courbe égale à 1. Elle va permettre de calculer des probabilités.

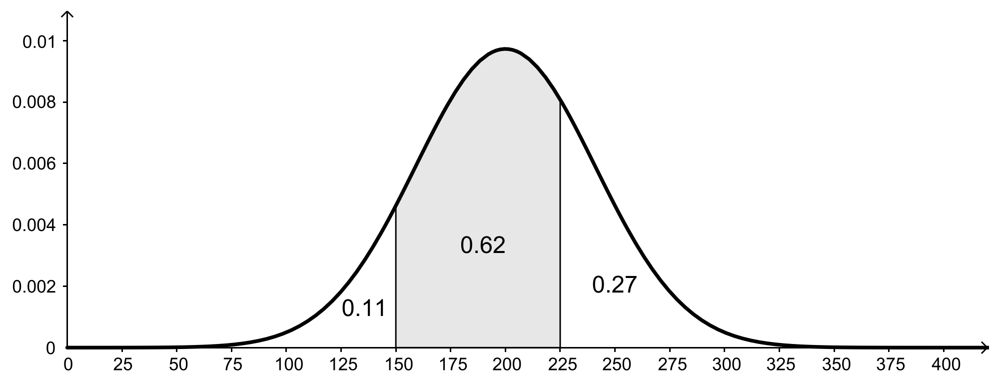


Considérons l'événement « la durée de vie de l'ampoule est comprise entre 150 et 200 heures ». Cet événement se note $(150 \leq X \leq 225)$. Sa probabilité se calcule en mesurant l'aire comprise entre la courbe de f, l'axe des abscisses et les droites d'équations $x = 150$ et $x = 225$.
Ici on a donc $P(150 \leq X \leq 225) = 0{,}4$.

De même, la probabilité que l'ampoule ait une durée de vie supérieure à 225 heures est $P(X \geq 225) = 0{,}36$.

La courbe étant symétrique par rapport à la droite d'équation $x = 200$, la durée de vie moyenne d'une ampoule est 200 heures.
On dit que **l'espérance** de X est 200 heures. On la note μ (lire « mu »).

Considérons à présent une nouvelle fonction f ci-dessous. Sa courbe est encore symétrique par rapport à la droite d'équation $x = 200$, mais elle davantage « resserrée » que la précédente.



Concrètement, les ampoules ont encore une durée de vie moyenne de 200 heures, mais il est plus probable qu'elle meurt autour de 200 heures dans le premier exemple.

La différence entre ces deux cas est mesuré par un nombre positif, appelé écart-type, noté σ (lire « sigma »). Plus σ est proche de 0, plus la courbe est resserrée autour de la moyenne.

Les deux variables aléatoires que l'on vient d'étudier sont des exemples de variable aléatoires suivant des lois normales.

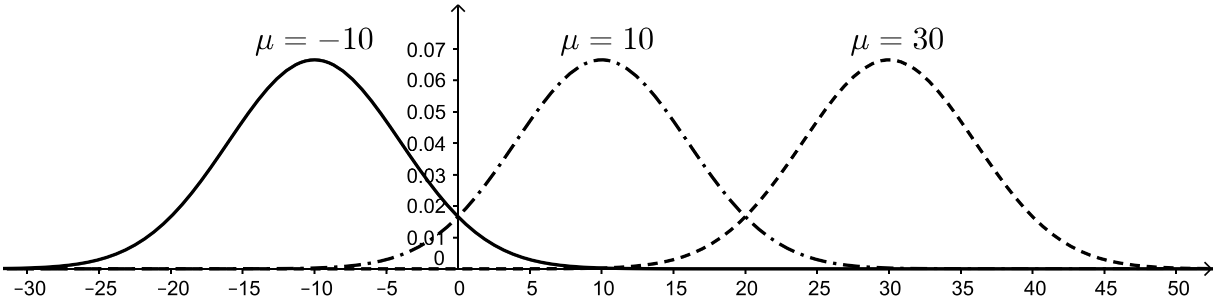
❖ Courbe de la loi normale

Comme on vient de le voir, deux paramètres caractérisent une loi normale :
- son espérance μ ;
- son écart-type σ.

On parle de loi normale de paramètres μ et σ, ce que l'on note $\mathcal{N}(\mu; \sigma^2)$.
Ainsi par exemple, la loi $\mathcal{N}(7; 9)$ a pour paramètres $\mu = 7$ et $\sigma = \sqrt{9} = 3$.

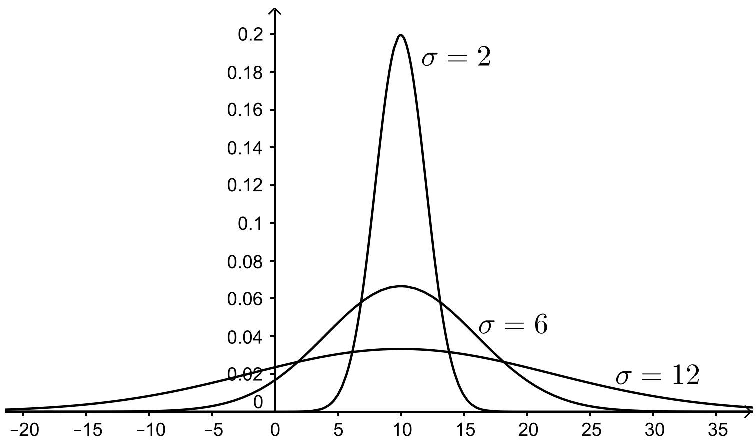
- **Rôle de l'espérance μ**

Les 3 lois normales suivantes ont le même écart-type. On constate que μ permet de déplacer horizontalement la courbe, mais ne modifie pas sa forme.
La courbe de la loi normale est symétrique par rapport à la droite d'équation $x = \mu$.



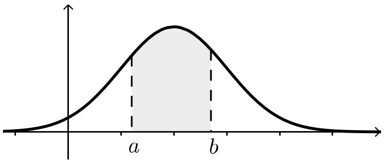
- **Rôle de l'écart-type σ**

Les 3 lois normales suivantes ont la même espérance. On constate que σ modifie la hauteur de la « cloche » sans modifier la symétrie.
Plus σ est petit, plus la courbe est « resserrée » autour de son axe de symétrie.

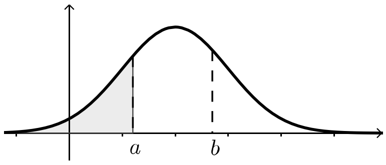


❖ Calcul de probabilité avec la loi normale

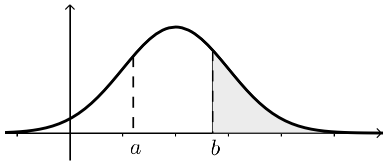
Soit X suivant une loi normale. La probabilité $P(a \leq X \leq b)$ que la variable aléatoire X appartienne à $[a;b]$ est égale à l'aire du domaine compris entre la courbe de la loi normale, l'axe des abscisses et les droites d'équation $x = a$ et $x = b$.
On définit de même les probabilités $P(X \leq a)$ ou $P(X \geq b)$.



$P(a \leq X \leq b)$



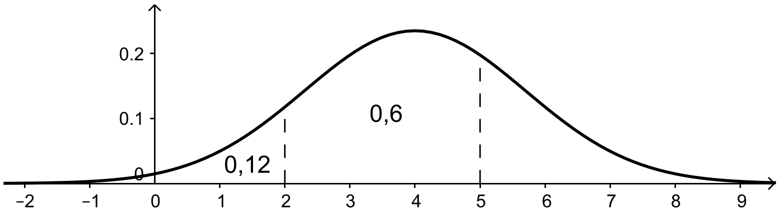
$P(X \leq a)$



$P(X \geq b)$

Exemple A

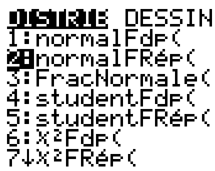
Soit X une variable aléatoire suivante une loi normale de paramètre $\mu = 4$ et $\sigma = 1{,}7$.
On donne ci-dessous la courbe représentative de la densité de X ainsi que les valeurs de deux aires.



Calculer

- **a.** $P(2 \leq X \leq 5) =$
- **b.** $P(X \leq 0{,}2) =$
- **c.** $P(X \leq 5) =$
- **d.** $P(X \geq 5) =$
- **e.** $P(X \geq 2) =$
- **f.** $P(X = 3) =$

❖ Utilisation de la calculatrice



Soit X une variable aléatoire suivant une loi normale d'écart-type μ et de variance σ.
Le menu Distrib, obtenue en faisant $\boxed{\text{2nde}}$ $\boxed{\text{var}}$, permet d'accéder à la commande `normalFRép`.

	Syntaxe TI
Calcul de $P(a \leq X \leq b)$	`normalFrép(a,b,μ,σ)`
Calcul de $P(X \leq a)$	`normalFrép(-10^99,b,μ,σ)`
Calcul de $P(X \geq b)$	`normalFrép(b,10^99,μ,σ)`

Exemple A

La calculatrice permet d'effectuer les calculs de probabilité ci-dessus.

- **a.** $P(2 \leq X \leq 5)$ s'obtient avec `NormalFrép(2,5,4,1.7)`.
- **b.** $P(X \leq 2)$ s'obtient avec `NormalFrép(-10^99,2,4,1.7)`.
- **d.** $P(X \geq 5)$ s'obtient avec `NormalFrép(5,10^99,4,1.7)`.

En Casio, la commande est `normCD(a,b,σ,μ)` qui s'obtient par « OPTN », « STAT », puis « DIST ».

- **Un intervalle à connaître**

Soit X une variable aléatoire suivant une loi normale d'écart-type μ et de variance σ. Alors $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0{,}95$. L'intervalle $[\mu - 2\sigma ; \mu + 2\sigma]$ est appelé <u>intervalle de fluctuation de X au seuil de 95 %</u>.

2. Échantillonnage et estimation

- **Principe de l'échantillonnage et de l'estimation**

On illustre la situation ainsi : on dispose d'une grande urne où se trouvent un très grand nombre de boules rouges et bleues.

Cas 1 : la proportion p des boules rouges est connue	**Cas 2 : la proportion p des boules rouges est inconnue**
On tire au hasard avec remise n boules de l'urne. On obtient donc une fréquence f de boules rouges sur cet échantillon. On s'attend à ce que la fréquence f observée soit « proche » de p, fréquence théorique. On est ici dans le cadre d'un <u>échantillonnage</u>.	On veut donc estimer la proportion p de boules rouges dans l'urne. Comme il y en a beaucoup, on ne peut pas toutes les compter. On tire donc au hasard avec remise n boules de l'urne. On obtient donc une fréquence f de boules rouges sur cet échantillon. On s'attend à ce que la fréquence p théorique soit « proche » de f, fréquence observée. On est ici dans le cadre d'une <u>estimation</u>.

- **Intervalles de fluctuation et de confiance**

Échantillonnage	**Estimation**
On connait p, fréquence théorique d'un caractère sur une population. On a un échantillon de taille n. **L'intervalle $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$ est appelé intervalle de fluctuation au seuil 95 %** de la fréquence de ce caractère aléatoire de taille n issu de la population. **Conditions de validité :** $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$. **Interprétation.** La fréquence observée f sur un échantillon de taille n appartient à l'intervalle de fluctuation au seuil 95 % dans 95 % des cas.	On connait f, fréquence observée d'un caractère sur un échantillon de taille n d'une population. **L'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$ est appelé intervalle de confiance au seuil 95 %** de la proportion p de ce caractère aléatoire de la population. **Interprétation.** Au moins 95 % des intervalles de confiance au seuil 95 % contiennent la fréquence théorique p.

Exemple

Dans une urne contenant 4 boules rouges et 6 boules blanches, on effectue 10000 tirages au hasard avec remise. On obtient donc un échantillon de taille $n = 10000$.
La proportion p de boules rouges est égale à $p = 0{,}4$. Comme $n \geq 30$, $np = 4000 \geq 5$ et $n(1-p) = 6000 \geq 5$, l'intervalle de fluctuation au seuil de 95 % de la fréquence de boules rouges observées est

$$\left[0{,}4 - \frac{1}{\sqrt{10000}}; 0{,}4 + \frac{1}{\sqrt{10000}}\right] = [0{,}39; 0{,}41].$$

Cela signifie qu'avec une probabilité d'au moins 95 %, la fréquence de boules rouges observées dans un échantillon de taille 10000 sera comprise entre 39 % et 41 %.

Exemple

On dispose d'une pièce de monnaie que l'on sait truquée. On veut estimer la probabilité p de l'issue « pile » pour cette pièce.
On procède à 10000 tirages et l'on observe 4230 « pile ».
La fréquence observée sur cet échantillon de taille $n = 10000$ de l'issue « pile » est $f = 0{,}423$.
On calcule l'intervalle de confiance au seuil 95 % :

$$\left[0{,}423 - \frac{1}{\sqrt{10000}}; 0{,}423 + \frac{1}{\sqrt{10000}}\right] = [0{,}413; 0{,}433].$$

Il y a donc 95 % de chance que la probabilité p de l'issue « pile » appartienne à cet intervalle.

- **Prise de décision**

Dans ce paragraphe, la proportion du caractère étudié est supposée être égale à p.
La prise de décision consiste, à partir d'un échantillon de taille n, à valider ou non cette hypothèse faite sur la proportion p. Pour ce faire,

- on calcule la fréquence observée f du caractère étudié dans cet échantillon ;
- on détermine l'intervalle de fluctuation au seuil de 95 % de la fréquence f ;
- enfin on applique la règle de décision suivante
  - si la fréquence observée f appartient à l'intervalle de fluctuation asymptotique au seuil de 95 %, on ne rejette pas l'hypothèse faite sur p ;
  - si la fréquence observée f n'appartient pas à l'intervalle de fluctuation asymptotique au seuil de 95 %, on rejette l'hypothèse faite sur la proportion p avec un risque de 5 % de se tromper.

Remarque.
Le risque de 5 % signifie que la probabilité que l'on rejette à tort l'hypothèse faite sur la proportion p alors qu'elle est vraie est approximativement égale à 0,05. C'est une probabilité conditionnelle.
Dans le cas où l'on ne rejette pas l'hypothèse faite sur la proportion p, le risque d'erreur n'est pas quantifié.

Exemple

Une machine fabrique des composants électroniques qui présentent un défaut de fabrication avec une probabilité 0,4.
On constitue un échantillon de taille 100 de composants issus de la fabrication. On en détecte 49 présentant un défaut de fabrication. La machine nécessite-t-elle d'être réparée ?

On fait l'hypothèse que la machine est bien réglée, donc que $p = 0{,}4$.
Un intervalle de fluctuation au seuil de 95 % de la fréquence de pièces présentant un défaut sur des échantillons de taille 100 est $\left[0{,}4 - \frac{1}{\sqrt{100}} ; 0{,}4 + \frac{1}{\sqrt{100}}\right] = [0{,}3; 0; 5]$.
La fréquence observée f de pièces présentant un défaut sur un échantillon de taille 100 est $f = \frac{49}{100} = 0{,}49$. Comme $f \in [0{,}304; 0; 496]$, on ne rejette pas l'hypothèse $p = 0{,}4$.
On peut donc dire que la machine ne nécessite pas d'être réparée, au seuil de confiance de 95 %.