

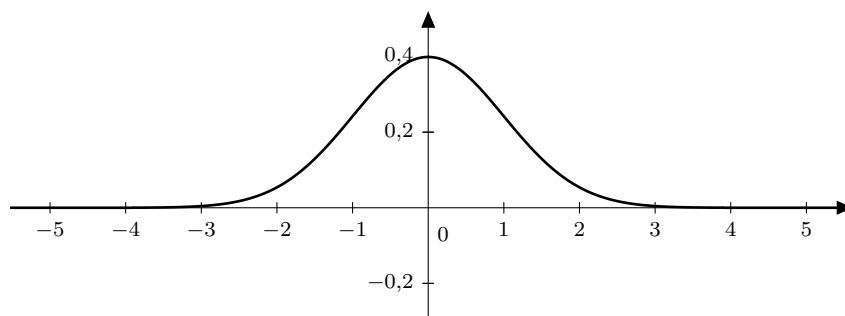
◆ Chapitre 16 : Lois normales et applications

I. — Loi normale centrée réduite

Théorème 1. — (admis)

La fonction f définie sur $\mathbb{R}$ par $f(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ est une densité de probabilité sur $\mathbb{R}$.

La courbe de la fonction f est la courbe suivante appelée courbe gaussienne ou courbe en « cloche ». La fonction f étant paire, la courbe de f est symétrique par rapport à l'axe des ordonnées.



Définition 2

On dit qu'une variable aléatoire X suit une loi normale centrée réduite, notée $\mathcal{N}(0, 1)$, si la densité de X est la fonction f .

Autrement dit, X suit une loi $\mathcal{N}(0, 1)$ si, pour tous réels a et b tels que $a \leq b$,

1. $\mathbf{P}(X \leq a) = \mathbf{P}(X < a) = \lim_{x \rightarrow -\infty} \int_x^a \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$;
2. $\mathbf{P}(a \leq X \leq b) = \mathbf{P}(a < X \leq b) = \mathbf{P}(a \leq X < b) = \mathbf{P}(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$
3. $\mathbf{P}(X \geq b) = \mathbf{P}(X > b) = \lim_{x \rightarrow +\infty} \int_b^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$.

Rappels. Comme pour toute loi continue, si X suit une loi $\mathcal{N}(0, 1)$ alors, pour tout réel a , $\mathbf{P}(X = a) = 0$.

Propriété 3

Si X suit une loi $\mathcal{N}(0, 1)$ alors $\mathbf{P}(X \leq 0) = \mathbf{P}(X \geq 0) = \frac{1}{2}$.

Démonstration. Comme la fonction densité f est paire, sa courbe est symétrique par rapport à l'axe des ordonnées. Ainsi, l'aire sous la courbe sur l'intervalle $]-\infty; 0]$ est la même que l'aire sous la courbe sur l'intervalle $[0; +\infty[$. Autrement dit, $\mathbf{P}(X \leq 0) = \mathbf{P}(X \geq 0)$. De plus, $\mathbf{P}(X \leq 0) + \mathbf{P}(X \geq 0) = 1$ donc $\mathbf{P}(X \leq 0) = \mathbf{P}(X \geq 0) = \frac{1}{2}$. □

Calcul pratique. On ne connaît pas de primitive de f sur $\mathbb{R}$. Pour calculer des probabilités en lien avec une loi $\mathcal{N}(0, 1)$, le seul moyen à notre disposition est de déterminer des valeurs approchées à l'aide de la calculatrice. Celle-ci permet de calculer $\mathbf{P}(a \leq X \leq b)$.

	T.I.	Casio	NumWorks
Touches	2nde var 2	OPTN F5 F3 F1 F2	Menu Probabilités puis Normale
Affichage	<code>normalFRép(a,b,0,1)</code>	<code>NormCD(a,b,1,0)</code>	$\mathbf{P}(\text{a} \leq X \leq \text{b}) = \text{}$



Les calculatrices T.I. et Casio ne calculent que des probabilités du type $\mathbf{P}(a \leq X \leq b)$ (et donc aussi $\mathbf{P}(a < X \leq b)$, $\mathbf{P}(a \leq X < b)$ et $\mathbf{P}(a < X < b)$). Si on doit calculer une probabilité du type $\mathbf{P}(X \leq a)$ ou $\mathbf{P}(X \geq b)$, il faut s'y ramener en utilisant les propriétés de symétrie de la courbe de f .

Exemple 4. Soit X une variable aléatoire suivant une loi $\mathcal{N}(0, 1)$. Calculer une valeur approchée à 10^{-3} de $\mathbf{P}(-2 \leq X < 1)$, $\mathbf{P}(X < 0,5)$ et $\mathbf{P}(X \geq 1)$, $\mathbf{P}(X < -0,8)$ et $\mathbf{P}(X \geq -0,3)$.

- Pour la première, la calculatrice donne directement $\mathbf{P}(-2 \leq X < 1) \approx 0,819$.
- Pour la seconde, on écrit $\mathbf{P}(X < 0,5) = \mathbf{P}(X \leq 0) + \mathbf{P}(0 < X < 0,5)$ donc, par la propriété 3, $\mathbf{P}(X < 0,5) = \frac{1}{2} + \mathbf{P}(0 < X < 0,5)$. Ensuite, à l'aide de la calculatrice, $\mathbf{P}(0 < X < 0,5) \approx 0,191$ donc $\mathbf{P}(X < 0,5) \approx 0,691$.
- Pour la troisième, on écrit $\mathbf{P}(X \geq 1) = \mathbf{P}(X \geq 0) - \mathbf{P}(0 \leq X < 1)$ donc, par la propriété 3, $\mathbf{P}(X \geq 1) = \frac{1}{2} - \mathbf{P}(0 \leq X < 1)$. Ensuite, à l'aide de la calculatrice, $\mathbf{P}(0 \leq X < 1) \approx 0,341$ donc $\mathbf{P}(X \geq 1) \approx 0,159$.
- Pour la quatrième, on écrit $\mathbf{P}(X < -0,8) = \mathbf{P}(X \leq 0) - \mathbf{P}(-0,8 \leq X \leq 0) = \frac{1}{2} - \mathbf{P}(-0,8 \leq X \leq 0)$. Ensuite, à l'aide de la calculatrice, $\mathbf{P}(-0,8 \leq X \leq 0) \approx 0,288$ donc $\mathbf{P}(X < -0,8) \approx 0,212$.
- Pour la cinquième, on écrit $\mathbf{P}(X \geq -0,3) = \mathbf{P}(-0,3 \leq X < 0) + \mathbf{P}(X \geq 0) = \mathbf{P}(-0,3 \leq X < 0) + \frac{1}{2}$. Ensuite, à l'aide de la calculatrice, $\mathbf{P}(-0,3 \leq X < 0) \approx 0,118$ donc $\mathbf{P}(X \geq -0,3) \approx 0,618$.

Théorème 5

Soit X une variable aléatoire suivant une loi $\mathcal{N}(0, 1)$. Pour tout réel $\alpha \in]0; 1[$, il existe un unique réel u_α tel que $\mathbf{P}(-u_\alpha \leq X \leq u_\alpha) = 1 - \alpha$.

Démonstration. On considère $G : u \mapsto \mathbf{P}(-u \leq X \leq u)$ définie sur $[0; +\infty[$. Par symétrie de la courbe de f , pour tout réel $u \geq 0$, $G(u) = 2\mathbf{P}(0 \leq X \leq u)$ i.e. $G : u \mapsto 2 \int_0^u f(t) dt$.

Notons $F : u \mapsto \int_0^u f(t) dt$ définie sur $\mathbb{R}$. Par le théorème fondamental du chapitre 13, F est la primitive de f qui s'annule en 0. Ainsi, F est dérivable donc continue sur $\mathbb{R}$ et $F(0) = 0$. De plus, pour tout réel u , $F'(u) = f(u) > 0$ donc F est strictement croissante sur $\mathbb{R}$ et $\lim_{u \rightarrow +\infty} F(u) = \frac{1}{2}$.

Soit $\alpha \in]0; 1[$. Alors, $0 < \alpha < 1$ donc $0 < 1 - \alpha < 1$ i.e. $0 < 1 - \alpha < 1$. Dès lors, $\frac{1-\alpha}{2} \in]0; \frac{1}{2}[$ donc, par le corollaire du théorème des valeurs intermédiaires, il existe un unique réel $u_\alpha > 0$ tel que $F(u_\alpha) = \frac{1-\alpha}{2}$. Or, $F(u) = \frac{1-\alpha}{2}$ équivaut à $2F(u) = 1 - \alpha$ i.e. par définition, à $G(u) = 1 - \alpha$. On conclut que u_α est l'unique réel tel que $G(u_\alpha) = 1 - \alpha$ i.e. $\mathbf{P}(-u_\alpha \leq X \leq u_\alpha) = 1 - \alpha$. $\square$

Calcul pratique. Étant donné $p \in]0; 1[$, la calculatrice permet de trouver une valeur approchée de l'unique réel c tel que $\mathbf{P}(X \leq c) = p$.

	T.I.	Casio	NumWorks
Touches	2nde var 3	OPTN F5 F3 F1 F3	Menu Probabilités puis Normale
Affichage	$\text{FracNormale}(p, 0, 1)$	$\text{InvNormCD}(p, 1, 0)$	$\mathbf{P}(\square \leq X) = \square$

En utilisant la symétrie de la courbe de f , on peut en déduire u_α .

Propriété 6

On a les valeurs remarquables $u_{0,05} \approx 1,96$ et $u_{0,01} \approx 2,58$. Autrement dit, si X suit une loi $\mathcal{N}(0, 1)$ alors

$$\mathbf{P}(-1,96 \leq X \leq 1,96) \approx 0,95 \quad \mathbf{P}(-2,58 \leq X \leq 2,58) \approx 0,99.$$

Démonstration. En reprenant les notations de la démonstration précédente, $\mathbf{P}(-u_{0,05} \leq X \leq u_{0,05}) = 1 - 0,05 = 0,95$ donc $G(u) = 0,95$ i.e. $2F(u) = 0,95$ soit encore $F(u) = 0,475$. Ainsi,

$$\mathbf{P}(X \leq u_{0,05}) = \mathbf{P}(X \leq 0) + \mathbf{P}(0 < X \leq u_{0,05}) = 0,5 + 0,475 = 0,975$$

On utilise donc la calculatrice avec $p = 0,975$ et on trouve $u_{0,05} \approx 1,96$.

Par le même raisonnement, on a $\mathbf{P}(X \leq u_{0,01}) = 0,5 + \frac{1-0,01}{2} = 0,995$ et on en déduit que $u_{0,01} \approx 2,58$. $\square$

Remarque 7. Graphiquement, cela signifie qu'environ 95% de l'aire sous la courbe de f se trouve entre les abscisses $-1,96$ et $1,96$ et 99% de l'aire sous la courbe de f se trouve entre les abscisses $-2,58$ et $2,58$.

Définition 8

Soit X une variable aléatoire suivant une loi ayant une densité continue g sur $\mathbb{R}$. Si les deux fonctions $x \mapsto \int_x^0 t \times g(t) dt$ et $x \mapsto \int_0^x t \times g(t) dt$ ont une limite finie respectivement en $-\infty$ et en $+\infty$ alors on définit l'espérance de X par $E(X) = \lim_{x \rightarrow -\infty} \int_x^0 t \times g(t) dt + \lim_{x \rightarrow +\infty} \int_0^x t \times g(t) dt$.

Si, de plus, la variable $(X - E(X))^2$ admet une espérance, on définit la variance de X par $V(X) = E((X - E(X))^2)$ et l'écart-type de X par $\sigma(X) = \sqrt{V(X)}$.

Propriété 9

Si X suit une loi $\mathcal{N}(0, 1)$ alors $E(X) = 0$ et $\sigma(X) = 1$.

Démonstration. Calcul de l'espérance. Soit un réel $x > 0$. Alors,

$$\int_0^x tf(t) dt = \frac{1}{\sqrt{2\pi}} \int_0^x te^{-\frac{t^2}{2}} dt = -\frac{1}{\sqrt{2\pi}} \int_0^x \underbrace{te^{-\frac{t^2}{2}}}_{u'e^u} dt = -\frac{1}{\sqrt{2\pi}} \left[e^{-\frac{t^2}{2}} \right]_0^x = -\frac{1}{\sqrt{2\pi}} \left[e^{-\frac{x^2}{2}} - 1 \right]$$

Lorsqu'on fait tendre x vers $+\infty$, $-\frac{x^2}{2}$ tend vers $-\infty$. Or, la fonction $\exp$ tend vers 0 en $-\infty$ donc, par composition, $\lim_{x \rightarrow +\infty} e^{-\frac{x^2}{2}} = 0$. Ainsi, $\lim_{x \rightarrow +\infty} \int_0^x tf(t) dt = \frac{1}{\sqrt{2\pi}}$.

Soit maintenant $x < 0$. Alors, par le même calcul,

$$\int_x^0 tf(t) dt = - \int_0^x tf(t) dt = \frac{1}{\sqrt{2\pi}} \left[e^{-\frac{x^2}{2}} - 1 \right]$$

donc, comme $\lim_{x \rightarrow -\infty} -\frac{x^2}{2} = -\infty$, on conclut de même que $\lim_{x \rightarrow -\infty} \int_x^0 tf(t) dt = -\frac{1}{\sqrt{2\pi}}$.

On conclut que X admet une espérance et que $E(X) = \frac{1}{\sqrt{2\pi}} - \frac{1}{\sqrt{2\pi}} = 0$.

Calcul de l'écart-type. Comme $E(X) = 0$, on a $V(X) = E(X^2)$. Soit un réel $x > 0$. Alors,

$$\int_0^x t^2 f(t) dt = \frac{1}{\sqrt{2\pi}} \int_0^x t^2 e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_0^x (-t) \times (-te^{-\frac{t^2}{2}}) dt$$

Considérons les fonctions $u : t \mapsto -t$ et $v : t \mapsto e^{-\frac{t^2}{2}}$. Alors, u et v sont dérivables sur $\mathbb{R}$ et leurs dérivées sont $u' : t \mapsto -1$ et $v' : t \mapsto -te^{-\frac{t^2}{2}}$. Comme ces dérivées sont continues, par la formule d'intégration par parties,

$$\begin{aligned} \int_0^x t^2 f(t) dt &= \frac{1}{\sqrt{2\pi}} \int_0^x u(t)v'(t) dt = \frac{1}{\sqrt{2\pi}} \left[-te^{-\frac{t^2}{2}} \right]_0^x - \frac{1}{\sqrt{2\pi}} \int_0^x (-1)e^{-\frac{t^2}{2}} dt \\ &= \frac{1}{\sqrt{2\pi}} \left(-xe^{-\frac{x^2}{2}} \right) + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt. \end{aligned}$$

Remarquons que $-xe^{-\frac{x^2}{2}} = \frac{2}{x} \left(-\frac{x^2}{2} e^{-\frac{x^2}{2}} \right)$. Or, $\lim_{x \rightarrow +\infty} -\frac{x^2}{2} = -\infty$ et, par théorème, $\lim_{X \rightarrow -\infty} Xe^X = 0$ donc, par composition, $\lim_{x \rightarrow +\infty} -\frac{x^2}{2} e^{-\frac{x^2}{2}} = 0$. De plus, $\lim_{x \rightarrow +\infty} \frac{2}{x} = 0$ donc, par produit, $\lim_{x \rightarrow +\infty} -xe^{-\frac{x^2}{2}} = 0$. De plus, $\frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt = \mathbf{P}(0 \leq X \leq x)$ donc $\lim_{x \rightarrow +\infty} \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt = \mathbf{P}(0 \leq X) = \frac{1}{2}$ par la propriété 3.

Soit maintenant $x < 0$. Alors, par le même calcul,

$$\int_x^0 t^2 f(t) dt = - \int_0^x t^2 f(t) dt = -\frac{1}{\sqrt{2\pi}} \left(-xe^{-\frac{x^2}{2}} \right) - \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt = -\frac{1}{\sqrt{2\pi}} \left(-xe^{-\frac{x^2}{2}} \right) + \frac{1}{\sqrt{2\pi}} \int_x^0 e^{-\frac{t^2}{2}} dt.$$

Comme précédemment, on a $\lim_{x \rightarrow +\infty} -xe^{-\frac{x^2}{2}} = 0$ et $\lim_{x \rightarrow +\infty} \frac{1}{\sqrt{2\pi}} \int_x^0 e^{-\frac{t^2}{2}} dt = \lim_{x \rightarrow -\infty} \mathbf{P}(x \leq X \leq 0) = \mathbf{P}(X \leq 0) = \frac{1}{2}$ par la propriété 3. On conclut que X admet une variance qui vaut $V(X) = \frac{1}{2} + \frac{1}{2} = 1$ et donc $\sigma(X) = \sqrt{1} = 1$. $\square$

Remarque 10. Cette proposition justifie le nom de la loi $\mathcal{N}(0, 1)$. Elle est « centrée » car sa moyenne est nulle et elle est « réduite » car son écart-type vaut 1.

II. — Loi normale $\mathcal{N}(\mu, \sigma^2)$

Définition 11

Soit μ et σ deux réels tels que $\sigma > 0$. On dit qu'une variable aléatoire X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ si la variable aléatoire Y définie par $Y = \frac{X - \mu}{\sigma}$ suit une loi $\mathcal{N}(0, 1)$.

Exemple 12. Soit X une variable aléatoire suivant un loi $\mathcal{N}(2, 9)$. Calculer $\mathbf{P}(-2 \leq X \leq 1)$.

Remarquons que

$$-2 \leq X \leq 1 \Leftrightarrow -2 - 2 \leq X - 2 \leq 1 - 2 \Leftrightarrow -4 \leq X - 2 \leq -1 \Leftrightarrow -\frac{4}{3} \leq \frac{X - 2}{3} \leq -\frac{1}{3}.$$

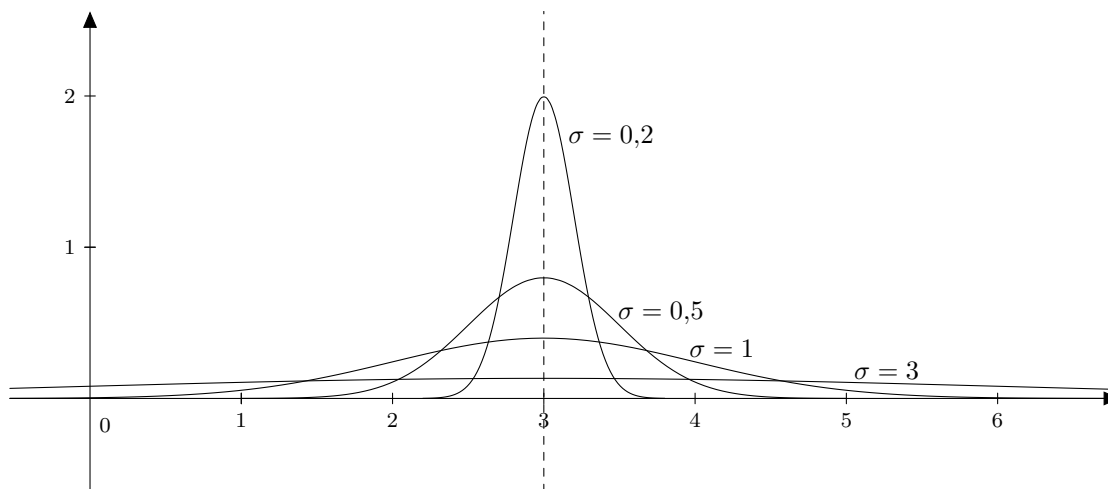
Posons $Y = \frac{X-2}{3}$. Ainsi, $(-2 \leq X \leq 1) = (-\frac{4}{3} \leq Y \leq -\frac{1}{3})$. Or, par définition, Y suit une loi $\mathcal{N}(0, 1)$ donc on peut calculer $\mathbf{P}(-\frac{4}{3} \leq Y \leq -\frac{1}{3})$ à l'aide de la calculatrice. On trouve $\mathbf{P}(-\frac{4}{3} \leq Y \leq -\frac{1}{3}) \approx 0,278$ et on conclut donc que $\mathbf{P}(-2 \leq X \leq 1) \approx 0,278$.

Propriété 13. — Admise

Si X suit une loi $\mathcal{N}(\mu, \sigma^2)$ alors

1. X est une variable aléatoire à densité continue sur $\mathbb{R}$;
2. la courbe de la densité de X est une courbe « en cloche » symétrique par rapport à la droite d'équation $x = \mu$;
3. $E(X) = \mu$ et $\sigma(X) = \sigma$.

Remarque 14. L'écart-type d'une variable aléatoire traduisant la dispersion des valeurs autour de la moyenne, plus σ est grand, plus la courbe est « étalée » et plus σ est petit, plus la courbe est « resserrée » autour de μ .



Remarque 15. Pour la même raison de symétrie de la courbe, si X suit une loi $\mathcal{N}(\mu, \sigma^2)$ alors $P(X \leq \mu) = P(X \geq \mu) = \frac{1}{2}$.

Calcul pratique. Comme pour la loi $\mathcal{N}(0, 1)$, la calculatrice permet de calculer $\mathbf{P}(a \leq X \leq b)$ et de déterminer le réel c tel que $\mathbf{P}(X \leq c) = p$ si X suit une loi $\mathcal{N}(\mu, \sigma^2)$.

		T.I.	Casio	NumWorks
Calculer	Touches	2nde var 2	OPTN F5 F3 F1 F2	Menu Probabilités puis Normale
$P(a \leq X \leq b)$	Affichage	<code>normalFRép(a,b,μ,σ)</code>	<code>NormCD(a,b,σ,μ)</code>	$P(\text{a} \leq X \leq \text{b}) = \text{ }$
Trouver c tel que	Touches	2nde var 3	OPTN F5 F3 F1 F3	Menu Probabilités puis Normale
$P(X \leq c) = p$	Affichage	<code>FracNormale(p,μ,σ)</code>	<code>InvNormCD(p,σ,μ)</code>	$P(X \leq \text{ }) = \text{p}$



Il faut entrer dans la calculatrice l'écart-type σ et non pas la variance σ^2 .

Exemple 16. Soit X une variable aléatoire de loi $\mathcal{N}(10, 25)$.

- Calculer $P(7 \leq X \leq 11)$, $P(X \leq 12)$, $P(X > 8)$, $P(X < 9)$ et $P(X \geq 11)$.
 - Déterminer c tel que $P(X \leq c) = 0,7$ et d tel que $P(X \geq d) = 0,8$.
- On a directement $P(7 \leq X \leq 11) \approx 0,305$. Ensuite,
 - $P(X \leq 12) = P(X \leq 10) + P(10 < X \leq 12) \approx \frac{1}{2} + 0,155 = 0,655$;
 - $P(X > 8) = P(8 < X < 10) + P(X \geq 10) \approx \frac{1}{2} + 0,155 = 0,655$;
 - $P(X < 9) = P(X \leq 10) - P(9 \leq X \leq 10) \approx \frac{1}{2} - 0,079 = 0,421$;
 - $P(X \geq 11) = P(X \geq 10) - P(10 \leq X < 11) \approx \frac{1}{2} - 0,079 = 0,421$.
 - On a directement $c \approx 12,622$. Pour déterminer d , on peut remarquer que $P(X \leq d) = 1 - P(X \geq d) = 0,2$ et donc $d \approx 5,792$.

Propriété 17

Soit X une variable aléatoire suivant une loi $\mathcal{N}(\mu, \sigma^2)$. Alors, pour tout réel $k > 0$, les probabilités des événements $(\mu - k\sigma \leq X \leq \mu + k\sigma)$ sont indépendantes de μ et de σ .

En particulier,

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0,683 \quad P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0,954 \quad P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0,997.$$

Démonstration. Soit k un réel strictement positif. Alors,

$$\mu - k\sigma \leq X \leq \mu + k\sigma \Leftrightarrow -k\sigma \leq X - \mu \leq k\sigma \Leftrightarrow -k \leq \frac{X - \mu}{\sigma} \leq k.$$

Posons $Y = \frac{X - \mu}{\sigma}$. Alors, par définition, Y suit une loi normale centrée réduite et on vient de montrer que $P(\mu - k\sigma \leq X \leq \mu + k\sigma) = P(-k \leq Y \leq k)$.

Ensuite, à l'aide de la calculatrice, on trouve $P(-1 \leq Y \leq 1) \approx 0,683$, $P(-2 \leq Y \leq 2) \approx 0,954$ et $P(-3 \leq Y \leq 3) \approx 0,997$ ce qui permet de conclure. $\square$

Exemple 18. Par exemple, si X suit une loi $\mathcal{N}(5, 36)$ alors $P(-1 \leq X \leq 11) \approx 0,683$ car $-1 = 5 - 6 = \mu - \sigma$ et $11 = 5 + 6 = \mu + \sigma$.

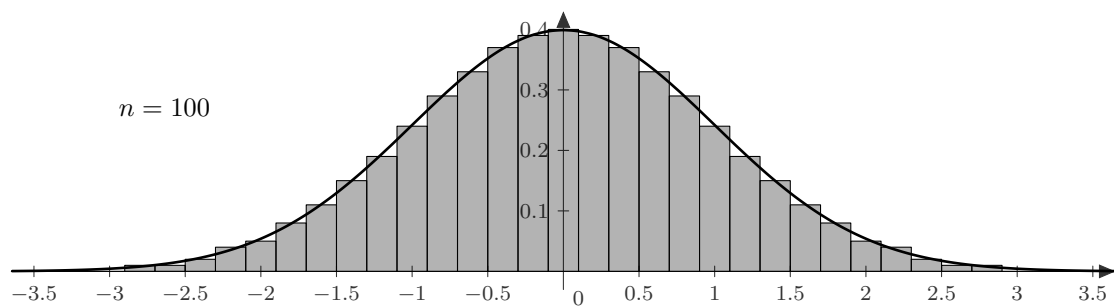
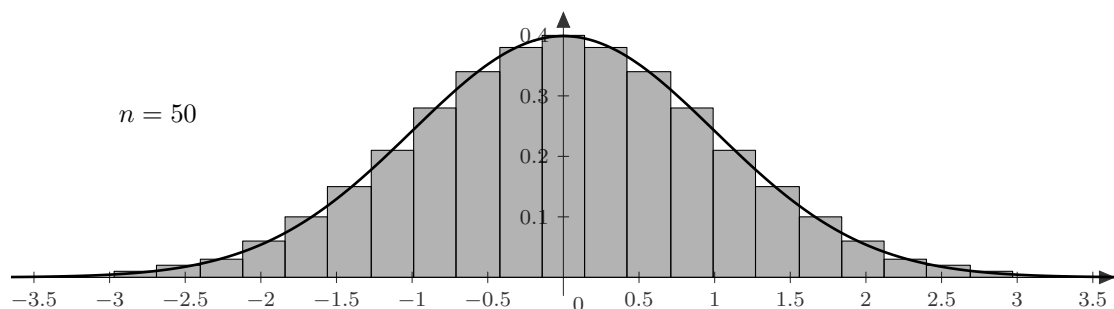
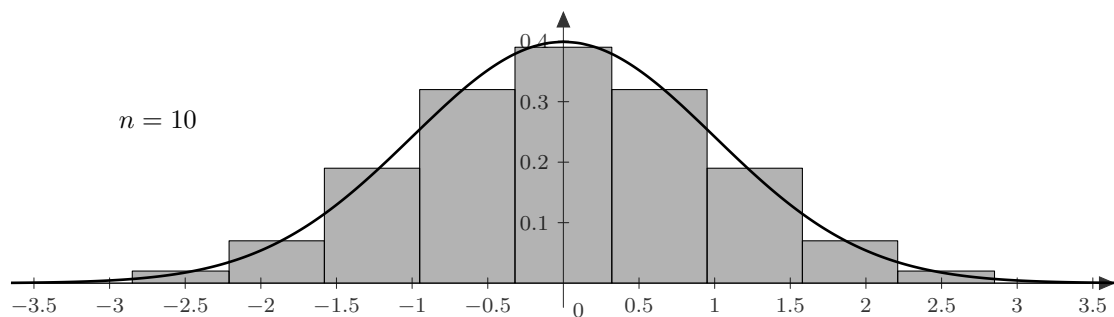
III. — Théorème de De Moivre-Laplace et applications

1) Théorème de De Moivre-Laplace

On considère une variable aléatoire X_n qui suit une loi binomiale de paramètres n et $p \in]0; 1[$. L'espérance de X_n est donc $\mu = np$ et l'écart-type de X_n est donc $\sigma = \sqrt{np(1-p)}$. On considère la variable centrée réduite

$Z_n = \frac{X_n - \mu}{\sigma}$ et on observe le comportement de la loi de probabilité de Z_n lorsque n augmente.

Dans les exemples ci-dessous, on a choisi $p = 0,5$.



On constate que l'histogramme représentant la loi de probabilité de Z_n se rapproche de la courbe gaussienne centrée réduite. Ceci est la traduction graphique du théorème suivant.

Théorème 19. — Théorème de De Moivre-Laplace (admis)

Pour tous réels a et b tels que $a < b$,

$$\lim_{n \rightarrow +\infty} \mathbf{P}(a \leq Z_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Remarque 20. Ce théorème peut se comprendre ainsi : si X_n suit une loi binomiale de paramètres n et p alors, lorsque n devient grand, la loi de la variable $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$ peut être approchée par la loi $\mathcal{N}(0, 1)$.

2) Intervalle de fluctuation asymptotique et application

Le but de ce paragraphe est le suivant : on considère une expérience aléatoire et on s'intéresse à un événement particulier A dont la probabilité p est connue (ou supposée connue). On réalise un grand nombre n de fois l'expérience et on note la fréquence f_n de réalisation de A i.e. le nombre de fois où A est réalisé divisé par n . Lorsque n devient grand, f_n converge vers p . On aimerait estimer la probabilité que f_n appartienne à un intervalle de la forme $[p - r; p + r]$ (avec $r > 0$).

a) Intervalle de fluctuation asymptotique

Si Y est une variable aléatoire suivant une loi $\mathcal{N}(0, 1)$ et si $\alpha \in]0; 1[$, on note, comme dans le théorème 5, u_α l'unique réel tel que $\mathbf{P}(-u_\alpha \leq Y \leq u_\alpha) = 1 - \alpha$.

Théorème 21

Si une variable aléatoire X_n suit une loi binomiale de paramètres n et $p \in]0; 1[$ alors, pour tout $\alpha \in]0; 1[$,

$$\lim_{n \rightarrow +\infty} \mathbf{P} \left(\frac{X_n}{n} \in I_n \right) = 1 - \alpha$$

où I_n désigne l'intervalle $\left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$.

Démonstration. Notons que

$$\begin{aligned} \frac{X_n}{n} \in I_n &\Leftrightarrow p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \\ &\Leftrightarrow -u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{X_n}{n} - p \leq u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \\ &\Leftrightarrow -u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{X_n - np}{\sqrt{n^2}} \leq u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \\ &\Leftrightarrow -u_\alpha \sqrt{p(1-p)} \leq \frac{X_n - np}{\sqrt{n}} \leq u_\alpha \sqrt{p(1-p)} \\ &\Leftrightarrow -u_\alpha \leq \frac{X_n - np}{\sqrt{n} \times \sqrt{p(1-p)}} \leq u_\alpha \\ &\Leftrightarrow -u_\alpha \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha \end{aligned}$$

Ainsi, en posant $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$, on a $\left\{ \frac{X_n}{n} \in I_n \right\} = \{-u_\alpha \leq Z_n \leq u_\alpha\}$. Or, si on note Y une variable aléatoire suivant une loi $\mathcal{N}(0, 1)$ alors, d'après le théorème de de Moivre-Laplace,

$$\lim_{n \rightarrow +\infty} \mathbf{P}(-u_\alpha \leq Z_n \leq u_\alpha) = \mathbf{P}(-u_\alpha \leq Y \leq u_\alpha) = 1 - \alpha$$

par définition de u_α . On conclut donc que $\lim_{n \rightarrow +\infty} \mathbf{P} \left(\frac{X_n}{n} \in I_n \right) = 1 - \alpha$. □

Interprétation. On suppose qu'une urne contient une proportion p de boules blanches. On tire successivement et avec remise n boules de l'urne. La variable aléatoire X_n égale au nombre de boules blanches tirées suit une loi binomiale de paramètres n et p . La variable $F_n = \frac{X_n}{n}$ représente alors la fréquence de boules blanches parmi les n boules tirées. Lorsque n devient grand, cette fréquence se rapproche de p . Ce que dit le théorème, c'est que pour n suffisamment grand, la fréquence F_n appartient à l'intervalle I_n avec une probabilité proche de $1 - \alpha$. On admet que c'est une bonne approximation si on a les conditions suivantes :

$$n \geq 30 \quad np \geq 5 \quad n(1-p) \geq 5.$$

Définition 22

L'intervalle $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ est appelé intervalle de fluctuation asymptotique au seuil de confiance (ou au niveau de confiance) $1 - \alpha$ de la variable aléatoire F_n .

Exemple 23. Un intervalle de fluctuation asymptotique au seuil de confiance 95% de la variable aléatoire F_n est

$$\left[p - 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right].$$

Exercice 24. On lance 500 fois un dé équilibré. Déterminer un intervalle de fluctuation asymptotique au seuil de confiance 95% de la fréquence de 6.

Même question en remplaçant 95% par 99%.

Dans cet exemple, $n = 500 \geq 30$, $p = \frac{1}{6}$ donc $np = \frac{250}{3} \geq 5$ et $n(1 - p) = \frac{1250}{3} \geq 5$.

On en déduit qu'un intervalle de fluctuation asymptotique au seuil de confiance 95% pour la fréquence du nombre de 6 est

$$\left[\frac{1}{6} - 1,96 \frac{\sqrt{\frac{1}{6}(1 - \frac{1}{6})}}{\sqrt{500}}; \frac{1}{6} + 1,96 \frac{\sqrt{\frac{1}{6}(1 - \frac{1}{6})}}{\sqrt{500}} \right] \approx [0,134; 0,2].$$

Comme $u_{0,01} \approx 2,58$, un intervalle de fluctuation asymptotique au seuil de confiance 99% pour la fréquence du nombre de 6 est

$$\left[\frac{1}{6} - 2,58 \frac{\sqrt{\frac{1}{6}(1 - \frac{1}{6})}}{\sqrt{500}}; \frac{1}{6} + 2,58 \frac{\sqrt{\frac{1}{6}(1 - \frac{1}{6})}}{\sqrt{500}} \right] \approx [0,123; 0,21].$$

Remarque 25. Pour être sûr que l'intervalle arrondi contienne l'intervalle de fluctuation asymptotique, on arrondit la borne inférieure par défaut et la borne supérieure par excès (pour obtenir un intervalle légèrement plus grand que l'intervalle de fluctuation asymptotique).

b) Prise de décision

Exemple 26. Dans un pays, on estime que 12% des habitants sont porteurs d'un caractère génétique g . On a effectué des tests sur un échantillon de 500 personnes dans deux villes A et B de ce pays. Sur les 500 personnes testées de la ville A, 71 sont porteurs du caractère g et sur les 500 personnes testées dans la ville B, 76 personnes sont porteurs du caractère g .

Peut-on affirmer que le nombre de porteurs du caractère est anormalement élevé dans la ville A ? dans la ville B ?

Règle de prise de décision. On considère une population dans laquelle on étudie un caractère C . On souhaite tester l'hypothèse H : « la proportion du caractère C dans la population est p » à partir d'un échantillon de la population de taille n (avec $n \geq 30$, $np \geq 5$ et $n(1 - p) \geq 5$). Pour cela,

1. on fixe un seuil de risque α ;
2. on détermine I_n un intervalle de fluctuation asymptotique au seuil de confiance $1 - \alpha$ pour la fréquence et on suit la règle de décision suivante :
 - si $f_n \in I_n$, on accepte l'hypothèse H ;
 - si $f_n \notin I_n$, on rejette l'hypothèse H avec un risque d'erreur α .

Pour l'exemple 26, fixons un seuil de risque $\alpha = 0,05$. Ici, $n = 500 \geq 30$ et $p = \frac{12}{100}$ donc $np = 60 \geq 5$ et $n(1 - p) = 440 \geq 5$. Un intervalle de fluctuation asymptotique au seuil de confiance 95% est donc

$$I_{500} = \left[\frac{12}{100} - 1,96 \frac{\sqrt{\frac{12}{100}(1 - \frac{12}{100})}}{\sqrt{500}}; \frac{12}{100} + 1,96 \frac{\sqrt{\frac{12}{100}(1 - \frac{12}{100})}}{\sqrt{500}} \right] \approx [0,091; 0,149].$$

Or, dans la ville A, la fréquence observée est $f_A = \frac{71}{500} = 0,142 \in I_{500}$ donc on accepte l'hypothèse selon laquelle la répartition est normale (au vu de cet échantillon) et, dans la ville B, la fréquence observée est $f_B = \frac{76}{500} = 0,152 \notin I_{500}$ donc on rejette, avec un risque d'erreur de 5%, l'hypothèse selon laquelle la répartition est normale (au vu de cet échantillon).

On conclut donc que le nombre de porteurs du caractère g est normal dans la ville A et anormalement élevé dans la ville B.

Remarque 27. Le résultat de la prise de décision dépend évidemment du risque α choisi. Le choix de privilégier telle ou telle valeur de α repose sur des considérations extra-mathématiques.

3) Estimation par intervalle de confiance

Dans ce paragraphe, on se pose le problème « réciproque » du précédent. On suppose que, dans une population, on étudie un caractère C dont la proportion p est inconnue. On cherche à estimer p en calculant la fréquence sur un échantillon.

Théorème 28

Soit X_n une variable aléatoire suivant une loi binomiale de paramètres n et $p \in]0; 1[$. Alors, il existe un rang $N \in \mathbb{N}$ tel que, pour tout $n \geq N$, $\mathbf{P}\left(p - \frac{1}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + \frac{1}{\sqrt{n}}\right) > 0,95$.

Démonstration. Soit Y une variable aléatoire suivant une loi $\mathcal{N}(0, 1)$. Considérons le réel α tel que $u_\alpha = 2$. D'après la propriété 17, $\mathbf{P}(-2 \leq Y \leq 2) \approx 0,954$ donc $1 - \alpha \geq 0,953$. Dès lors, par le théorème 21,

$$\lim_{n \rightarrow +\infty} \mathbf{P}\left(\frac{X_n}{n} \in I_n\right) \geq 0,953$$

où I_n désigne l'intervalle $\left[p - 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$. Ainsi, il existe un entier N tel que, pour tout entier $n \geq N$, $\mathbf{P}\left(\frac{X_n}{n} \in I_n\right) > 0,95$. Or, la fonction trinôme du second degré $x \mapsto x(1-x)$ atteint son maximum en $-\frac{-1}{2} = \frac{1}{2}$ et ce maximum vaut $\frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{4}$ donc $p(1-p) \leq \frac{1}{4}$ et, par croissance de la racine carrée, $\sqrt{p(1-p)} \leq \frac{1}{2}$. Ainsi, pour tout $n \in \mathbb{N}^*$, $2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$ donc

$$p - \frac{1}{\sqrt{n}} \leq p - 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p + \frac{1}{\sqrt{n}}$$

ce qui revient à dire que I_n est inclus dans $J_n = \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$. Dès lors, pour tout entier $n \geq N$, $\mathbf{P}\left(\frac{X_n}{n} \in J_n\right) \geq \mathbf{P}\left(\frac{X_n}{n} \in I_n\right) > 0,95$ i.e. $\mathbf{P}\left(p - \frac{1}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + \frac{1}{\sqrt{n}}\right) > 0,95$. $\square$

Corollaire 29

On considère une population dans laquelle la proportion d'un caractère C est p . À tout échantillon de taille n de cette population, on associe la variable aléatoire fréquence F_n égale à la fréquence du caractère dans l'échantillon. Alors, pour n assez grand, la proportion p appartient à l'intervalle $\left[F_n - \frac{1}{\sqrt{n}}; F_n + \frac{1}{\sqrt{n}}\right]$ avec une probabilité supérieure ou égale à 0,95.

Démonstration. Il suffit de remarquer que, pour tout $n \in \mathbb{N}^*$,

$$\begin{aligned} p \in \left[F_n - \frac{1}{\sqrt{n}}; F_n + \frac{1}{\sqrt{n}}\right] &\Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}} \\ &\Leftrightarrow -\frac{1}{\sqrt{n}} \leq p - F_n \leq \frac{1}{\sqrt{n}} \\ &\Leftrightarrow \frac{1}{\sqrt{n}} \geq F_n - p \geq -\frac{1}{\sqrt{n}} \\ &\Leftrightarrow p + \frac{1}{\sqrt{n}} \geq F_n \geq p - \frac{1}{\sqrt{n}} \\ &\Leftrightarrow p - \frac{1}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + \frac{1}{\sqrt{n}} \end{aligned}$$

donc $\mathbf{P}(p \in [F_n - \frac{1}{\sqrt{n}}; F_n + \frac{1}{\sqrt{n}}]) = \mathbf{P}(\frac{X_n}{n} \in [p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}])$. Or, d'après le théorème 28, pour n assez grand, $\mathbf{P}(\frac{X_n}{n} \in [p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}]) \geq 0,95$ donc, pour n assez grand, p appartient à l'intervalle $[F_n - \frac{1}{\sqrt{n}}; F_n + \frac{1}{\sqrt{n}}]$ avec une probabilité supérieure ou égale à 0,95. $\square$

Définition 30

On considère une population dans laquelle la proportion d'un caractère C est p et on note f la fréquence observée du caractère C dans un certain échantillon de taille n . Alors, l'intervalle $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ est appelé intervalle de confiance de la proportion p au niveau de confiance 95%.

Dans la pratique, on utilise cet intervalle si les conditions suivantes sont remplies :

$$n \geq 30 \quad nf \geq 5 \quad n(1 - f) \geq 5.$$

Exemple 31. Dans une population, une maladie touche une proportion p de personnes. On a testé 500 personnes et on a dépisté 45 personnes atteintes de cette maladie.

1. Donner un intervalle de confiance de p au niveau de confiance 95%.
2. Combien de personnes au minimum suffit-il tester pour connaître, au niveau de confiance 95%, la proportion p avec une précision au moins égale 0,01 ?

1. Ici, $n = 500 \geq 30$, $f = \frac{45}{500}$ donc $nf = 45 \geq 5$ et $n(1 - f) = 455 \geq 5$.

Un intervalle de confiance de la probabilité p au niveau de confiance 95% est donc $\left[\frac{45}{500} - \frac{1}{\sqrt{500}}; \frac{45}{500} + \frac{1}{\sqrt{500}}\right] \approx [0,045; 0,135]$.

2. On cherche n tel que f soit, au niveau de confiance 95%, une valeur approchée de p à 0,01 près. Or, on sait que, au niveau de confiance 95%, p appartient à l'intervalle $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ i.e. f est une valeur approchée de p à $\frac{1}{\sqrt{n}}$ près. Il suffit donc que $\frac{1}{\sqrt{n}} \leq 0,01$ i.e. $\sqrt{n} \geq 100$ i.e. $n \geq 10000$.

Ainsi, pour que f fournisse, au seuil de confiance 95%, une valeur approchée de p à 0,01, il suffit de tester 10000 personnes.