Statistiques

En statistiques, je fais des sondages, des essais sur un échantillon ou la totalité d'une population. Je traduis alors ces données et les décris par les nombres.
Je ne cherche pas à calculer les résultats que j'aurais le plus de « chance » d'avoir en théorie, sinon je fais des probabilités. Si l'échantillon, le nombre de sondages ou les essais tendent vers l'infini, ces résultats statistiques doivent correspondre aux valeurs théoriques déterminées par les lois de probabilité.

Vocabulaire - Définitions

- **x :** **valeur** obtenue au cours d'une étude statistique
  Ex. : somme d'argent, note d'élève, score, etc. …

- **(x_1 ; x_2 ; x_3 ; … ; x_{k-1} ; x_k) :** **série statistique de taille k** regroupant toutes les valeurs x_i obtenues (elles sont généralement toutes différentes deux à deux, donc **k ≤ N**).

  - **n_i : effectif associé à la valeur x_i** , c'est-à-dire le nombre de fois où cette valeur x_i est « apparue » dans la série statistique.

  - **N : effectif total** (ou population totale), somme de tous les effectifs n_i → $\mathbf{N = \sum_{i=1}^{k} n_i}$

  <u>Remarque</u> : Si la série statistique est (x_1 ; x_2 ; x_3 ; … ; x_{N-1} ; x_N), alors k = N, les valeurs x_j n'étant pas forcément toutes différentes, donc on aura : $n_j = 1 \quad \forall j$. → $N = \sum_{j=1}^{N} n_j = \sum_{j=1}^{N} 1$

- **Événement :** il désigne le fait qu'une valeur x_i apparaisse ou qu'une action se produise

Ex. : « Voici le tableau de notes d'une classe après un devoir.

x = Notes (/20)	**7,5**	**9**	**11**	**11,5**	**13**	**15**	**16**	**19**
n = nombre d'élèves	**2**	**4**	**6**	**5**	**4**	**2**	**3**	**2**

Soit A l'événement : 'j'ai entre 10/20 et 15/20 (notes incluses)'.
Calcule l'effectif total de cette classe N et l'effectif n_A correspondant à l'événement A.
Précise quelle valeur x_i a le plus grand effectif associé. »

*Nous travaillons avec la série **statistique de taille 8 : (7,5 ; 9 ; 11 ; 11,5 ; 13 ; 15 ; 16 ; 19)**.*

→ Effectif total de la classe : $N = \sum_{i=1}^{k} n_i = 2 + 4 + 6 + 5 + 4 + 2 + 3 + 2 = 28$

***Il y a N = 28 élèves dans cette classe**.*
*→ **n_A = 6 + 5 + 4 + 2 = 17 élèves ont eu entre 10/20 et 15/20 (notes incluses)**.*
*→ la valeur x_3 = **11/20 est la note la plus fréquemment obtenue par les élèves** (n_3 = 6).*

- **Fréquence :** $\mathbf{f_i = \frac{n_i}{N}}$ → indique la **proportion de la valeur x_i** dans la série statistique.

  NB : **Si N → +∞ , alors : f_i → p_i** (probabilité associée à x_i)

- **Moyenne :** $\bar{x} = \frac{1}{N}.\sum_{i=1}^{k} n_i.x_i = \sum_{i=1}^{k} \frac{n_i}{N}.x_i = \sum_{i=1}^{k} f_i.x_i$ → donne le « centre » de la répartition des valeurs x_i relevées.

  NB : $\overline{x^2} = \frac{1}{N}.\sum_{i=1}^{k} n_i.x_i^2 = \sum_{i=1}^{k} \frac{n_i}{N}.x_i^2 = \sum_{i=1}^{k} f_i.x_i^2$ → (moyenne des carrés des valeur x_i)

- **Variance :** $\mathbf{V = \frac{1}{N}.\sum_{i=1}^{k} n_i.(x_i-\bar{x})^2} = \sum_{i=1}^{k} f_i.(x_i-\bar{x})^2$

  $\mathbf{V = \overline{(x_i-\bar{x})^2}}$

  → V est le nombre égal à la moyenne des valeurs $(x_j-\bar{x})^2$, c'est-à-dire la **moyenne des carrés des écarts des valeurs x_i à la moyenne $\bar{x}$** . Il est très abstrait mais permet de mesurer la « dispersion » des valeurs x_i autour de la valeur moyenne. **Il permet surtout de calculer ensuite l'écart-type σ**.

  Remarque : Il existe une autre formule de la variance qui peut être plus pratique à calculer sur tableur ou calculatrice :

$$V = \frac{1}{N}.\sum_{i=1}^{k} n_i.(x_i-\bar{x})^2 = \frac{1}{N}.\sum_{i=1}^{k} n_i.(x_i^2 - 2.x_i.\bar{x} + \bar{x}^2)$$
$$= \frac{1}{N}.\sum_{i=1}^{k} n_i.x_i^2 - 2.\bar{x}.\frac{1}{N}.\sum_{i=1}^{k} n_i.x_i + \bar{x}^2.\frac{1}{N}\sum_{i=1}^{k} n_i$$
$$= \overline{x^2} - 2.\bar{x}.\bar{x} + \bar{x}^2.1$$
$$\mathbf{V = \overline{x^2} - \bar{x}^2}$$

  Remarque : Si la série statistique relève toutes les valeurs x_j , même si celles-ci ne sont pas toutes différentes deux à deux, alors k = N et $n_j = 1$ ∀ j . On peut alors aussi écrire que :

$$V = \frac{1}{N}.\sum_{j=1}^{N} (x_j-\bar{x})^2 = \frac{1}{N}.\sum_{j=1}^{k} x_j^2-\bar{x}^2 = \overline{x^2}-\bar{x}^2$$

- **Écart-type :** $\mathbf{\sigma = \sqrt{V}}$

  → σ mesure la **dispersion des valeurs x_i autour de la valeur $\bar{x}$ moyenne**, mais il a la <u>même unité de</u> <u>mesure que x_i</u> .

Ex. : Si l'on revient sur les notes de la classe de 28 élèves précédente, nous trouvons :

x = Notes (/20)	**7,5**	**9**	**11**	**11,5**	**13**	**15**	**16**	**19**
n = nombre d'élèves	**2**	**4**	**6**	**5**	**4**	**2**	**3**	**2**
$(x_i - \bar{x})^2$	**22,39**	**10,45**	**1,52**	**0,54**	**0,59**	**7,66**	**14,20**	**45,80**

→ *Moyenne de la classe (**<u>vivent les tableurs ou calculatrices</u>** !) :*

$$\bar{x} = \frac{1}{28}.\sum_{i=1}^{8} n_i.x_i = \frac{1}{28}.(2\times7{,}5+4\times9+6\times11+5\times11{,}5+4\times13+2\times15+3\times16+2\times19) = \frac{342{,}5}{28}$$

$\bar{x} \approx$ ***12,23***

→ *Variance des notes de la classe :*

$$V \approx \frac{1}{28}.\sum_{i=1}^{8} n_i.(x_i - 12{,}23)^2$$

$$V \approx \frac{1}{28}.(2\times 22{,}39+4\times 10{,}45+6\times 1{,}52+5\times 0{,}54+4\times 0{,}59+2\times 7{,}66+3\times 14{,}2+2\times 45{,}8)$$

$$\boldsymbol{V \approx \frac{250{,}24}{28} \approx 8{,}94}$$ *J'en déduis l'écart-type σ :*

→ *Écart-type des notes de la classe :*

$\boldsymbol{\sigma = \sqrt{8{,}94} \approx 2{,}99}$ ***Cela signifie qu'une grande majorité des élèves a une note qui ne s'écarte pas plus de 3 points de la note moyenne 12,23 / 20 .***

Probabilités

En probabilités, je cherche à calculer les résultats que j'ai le plus de « chance » d'avoir en théorie. Si l'échantillon, le nombre de sondages ou les essais tendent vers l'infini, les résultats statistiques doivent rejoindre les valeurs théoriques déterminées par les lois de probabilité.

Vocabulaire - Définitions

- **Ω :** **Univers**, ensemble des événements possibles.
- **Événement A_i :** chaque événement A_i peut être associé à :
  - un **nombre de cas possibles n_i**
  - une **valeur** (ou un gain) **x_i**
- **Événements indépendants :** événements qui n'ont **aucun lien** (la réalisation de l'un n'a aucun effet sur la réalisation de l'autre)
- **Tirages successifs :** les événements ne sont **pas simultanés**, ils se suivent
- **Tirages successifs avec remise :** à chaque tirage, **le nombre d'événements possibles ne diminue pas** (événements indépendants)
- **Tirages successifs sans remise :** à chaque tirage, **le nombre d'événements possibles diminue**

Ex. : Si je place dans une urne 2 boules noires et 3 boules rouges, je peux effectuer 3 tirages successifs d'une boule avec remise.
Soit R l'événement : « une boule rouge est tirée »
Soit N l'événement : « une boule noire est tirée »

L'univers Ω sera alors :
Ω = { (R ; R ; R) ; (R ; R ; N) ; (R ; N ; R) ; (N ; R ; R) ; (R ; N ; N) ; (N ; R ; N) ; (N ; N ; R) }, soient 7 tirages possibles.

- **Probabilité :** $\mathbf{p(A) = \frac{n_A}{n_\Omega}}$ → « chances », **probabilité que l'événement A se réalise**.

NB : $\mathbf{p(\Omega) = \sum_{i=1}^{k} p_i = 1}$ → La somme des probabilités qu'au moins un des événements de l'univers Ω se réalise vaut toujours 1 (soient 100 % de chances de réalisation).

Ex. : dans le cas précédent, $p(2\,rouges\,tirées) = \frac{3}{7}$ → ***3 chances sur 7 de tirer 2 boules rouges***.

- **Espérance :** $\mathbf{E = \frac{1}{n_\Omega}.\sum_{i=1}^{k} n_i.x_i = \sum_{i=1}^{k} \frac{n_i}{n_\Omega}.x_i = \sum_{i=1}^{k} p_i.x_i}$ → Avec $k \leq n_\Omega$, chaque événement A_i étant associé à un nombre de cas possibles n_i et une valeur (ou gain) x_i , **E mesure la valeur (le gain) que l'on peut espérer gagner (moyenne théorique)**.

NB : **en statistiques**, si $N \to +\infty$, alors : $f_i \to p_i$ (probabilité associée à x_i). Si l'on compare les deux formules, on constate donc que :
si N → +∞ , alors : $\bar{x}$ → **E** (les statistiques rejoignent les probabilités sur de grands échantillons). **L'espérance E est donc (en probabilités) la valeur théorique vers laquelle tend la moyenne** $\bar{x}$ **(en statistiques)**.

Ex. : Dans le cas précédent, si je perds 3 € par boule rouge et gagne 5 € par boule noire, alors j'obtiens le tableau suivant :

i	*1*	*2*	*3*
Événements A_i	3 fois R	2 fois R ; 1 fois N	1 fois R ; 2 fois N
Gains x_i (en €)	**- 9 €**	**-1 €**	**7 €**
Nombre de cas favorables n_i	**1**	**3**	**3**
Probabilité p_i	$\frac{1}{7}$	$\frac{3}{7}$	$\frac{3}{7}$

Je peux donc calculer l'espérance : $\mathbf{E = \sum_{i=1}^{3} p_i.x_i = \frac{-9}{7} + \frac{-3}{7} + \frac{21}{7} = \frac{9}{7} \approx 1{,}29\ €}$

(ce jeu n'est donc pas une arnaque !).

Statistiques et probabilités avancées

(Du discret au continu…)

✔ Quand la valeur x n'est plus discrète, on parle de **variable aléatoire <u>continue</u> :**

$$\mathbf{X = k \in I} \text{ (X est un réel)}$$

Celle-ci est **associée à un événement** par la notation suivante :

$$\{\mathbf{X = k \in [a\ ;\ b]}\}, \text{ avec } [a\ ;\ b] \subset I.$$

✔ La probabilité $p(A_i)$ est remplacée par une probabilité **$P(X \in [a\ ;\ b])$** ou **$P(a \leq X \leq b)$** ou **$P([a\ ;\ b])$**.

✔ Pour calculer une probabilité $P(X \in [a\ ;\ b])$, il faut connaître sa **<u>fonction de densité</u>** (ou **<u>densité de probabilité</u>**) **f** :

- f est **définie et continue sur I = $(x_{min}\ ;\ x_{MAX})$**
- f est **positive sur I**
- l'**intégrale** de f sur I doit être **égale à 1** *(la somme des probabilités vaut toujours 1…)*
- Je peux alors calculer la probabilité $P(X \in [a\ ;\ b])$ à partir de sa densité de probabilité f :

$$\mathbf{P(X \in [a\ ;\ b]) = \int_a^b f(t).dt}$$

<u>NB</u> : ► $\mathbf{[a\ ;\ b] \subset I}$
► **t** est la variable pouvant prendre **toutes les valeurs de X sur [a ; b]**
► **f(t).dt** est la probabilité associée à un **intervalle de largeur dt infinitésimale** autour de t.

► **une <u>intégrale</u> (fonctions <u>continues</u>) étant l'équivalent d'une <u>somme</u> (de termes <u>discrets</u>)**, on retrouve donc ici une correspondance entre les formules :

Valeurs discrètes ***(en nombre fini)***	**Valeurs continues** ***(en nombre infini)***
Valeurs $\{x_m\ ;\ \dots\ ;\ x_{m+q}\} \in [a\ ;\ b]$	Valeurs de $\mathbf{X} = \mathbf{t} \in [a\ ;\ b]$
Probabilités $\{p_m = p(x_m)\ ;\ \dots\ ;\ p_{m+q} = p(x_{m+q})\}$	Infinité de probabilités de **f(a).dt** à **f(b).dt**
Événement $A_i = \{x_i \in [a\ ;\ b]\}$	Événement $\{X \in [a\ ;\ b]\}$
Probabilité $p(A_i) = \sum_{i=m}^{m+q} p_i$	Probabilité $P([a\ ;\ b]) = \int_a^b f(t).dt$

<u>NB</u> : Que les inégalités ci-dessus soient **<u>strictes ou larges, cela ne change rien</u> !**

Preuve : $P(a \leq X) = P(X < a)$ *car :* $P([a\ ;\ a]) = \int_a^a f(t).dt = F(a) - F(a) = 0$

✔ Si $I = [x_{min}\ ;\ x_{MAX}]$, alors : $\mathbf{E(X) = \int_{x_{min}}^{x_{MAX}} t.f(t).dt}$

<u>NB</u> : De même que ci-dessus, on retrouve l'équivalent avec la formule discrète : $\mathbf{E = \sum_{i=1}^{k} p_i.x_i}$

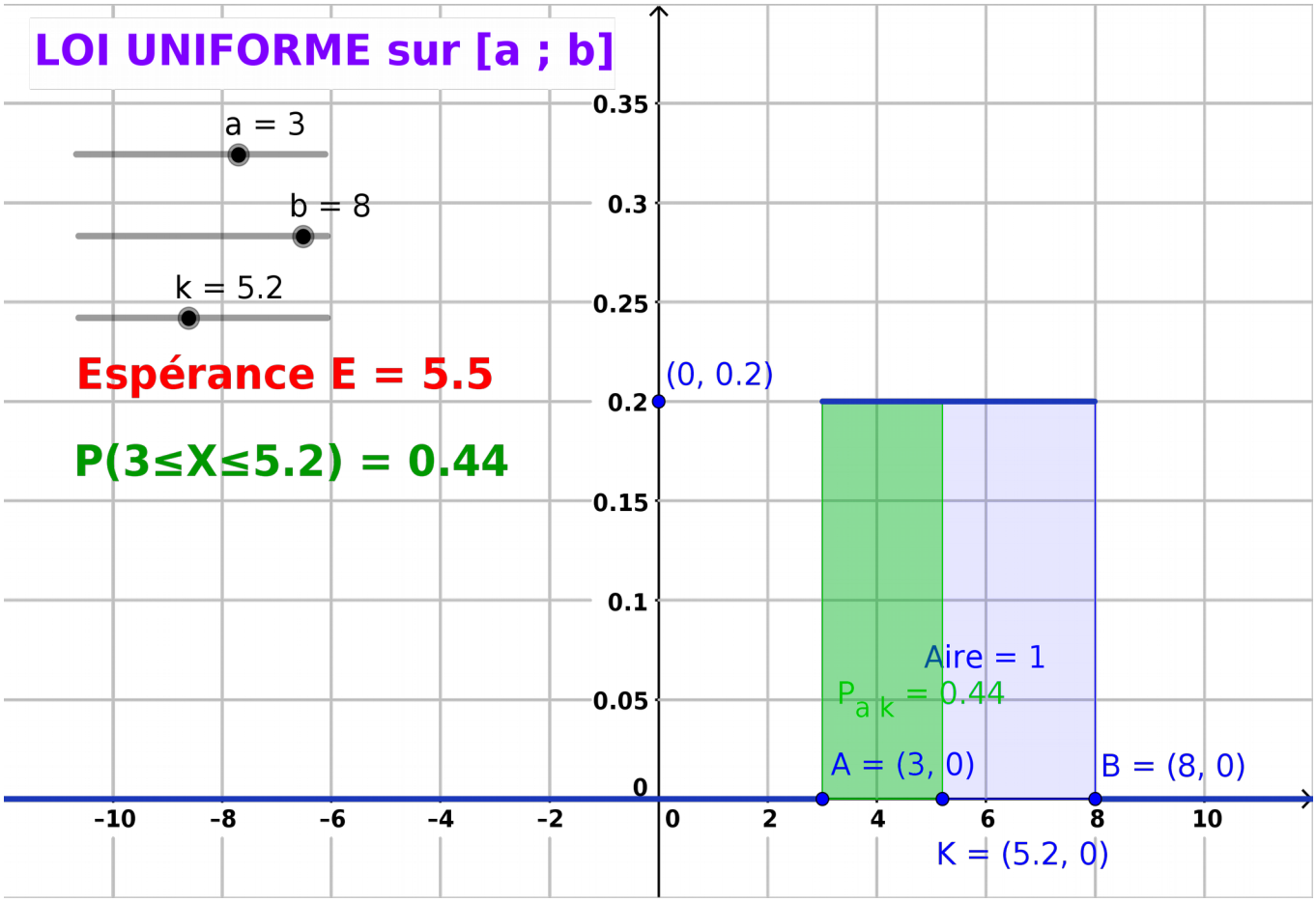
✗ **Loi uniforme** (notée : **U([a ; b])**) :

I = [a ; b]

$f(x) = \frac{1}{b-a}$ (fonction constante)

$P(a \leqslant X \leqslant x) = \frac{x-a}{b-a}$ avec $x \in [a\ ;\ b]$

$E(X) = \frac{a+b}{2}$



✗ **Durée de vie et lois exponentielles de paramètre λ** :

I = [0 ; +∞[

$f(x) = \lambda . e^{-\lambda . x}$ avec $\lambda \in \mathbb{R}$

$P(X \leqslant t) = 1 - e^{-\lambda . t}$ pour tout $t \in [0\ ;\ +\infty[$ donc :

$P(X \geqslant t) = e^{-\lambda . t}$ pour tout $t \in [0\ ;\ +\infty[$

$P([a\ ;\ b]) = e^{-\lambda . a} - e^{-\lambda . b}$

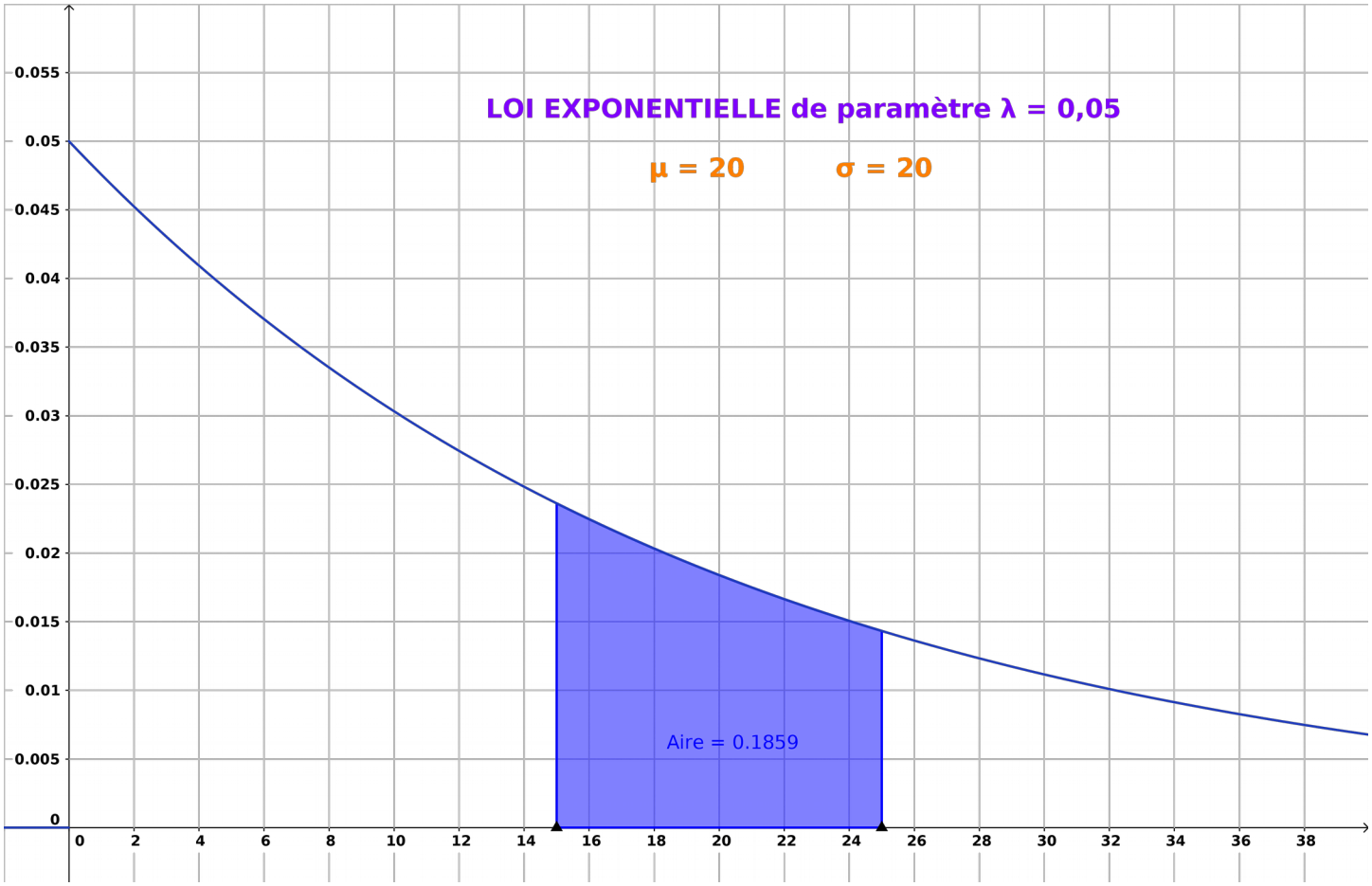
➢ **Durée de vie sans vieillissement :**

➔ « **X ≥ t** » est l'événement **« la durée de vie dépasse t années »**

➔ On dit que la durée de vie est « sans vieillissement » lorsque la probabilité qu'il fonctionne encore h heures supplémentaires, sachant qu'il fonctionne à l'instant t, ne dépend pas de t.
La durée de vie sur une période h ne dépend pas de l'âge t à partir duquel on considère cet événement.

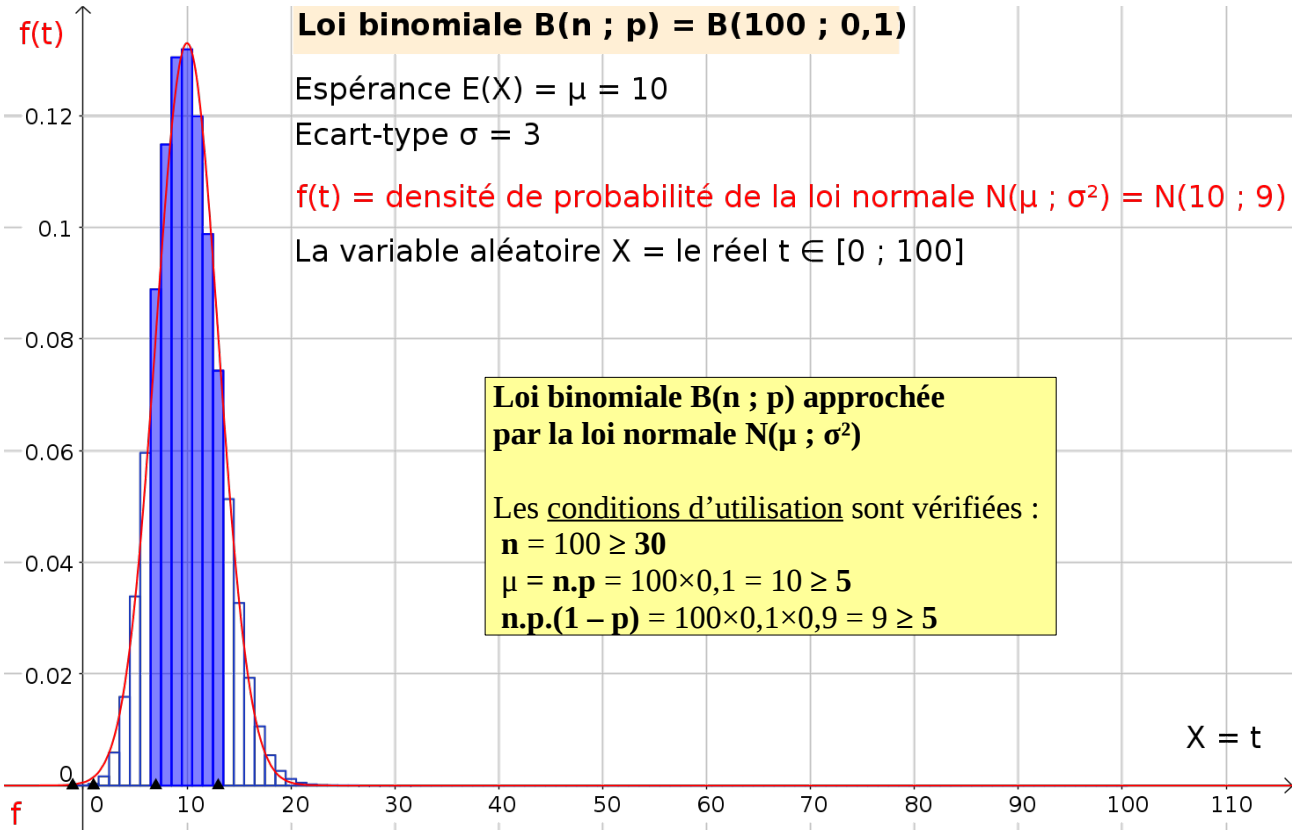
➔ Pour tout t et h positifs : $P_{X \geqslant t}(X \geqslant t+h) = P(X \geqslant h)$

➔ $E(X) = \frac{1}{\lambda}$



✔ **Loi de BERNOULLI B_p** de paramètre p**,**
Loi BINOMIALE B(n ; p) *(soient n répétitions d'une épreuve de Bernoulli B_p)***,**
Loi NORMALE N(μ ; σ ²) *(espérance μ ; variance σ ²)* **:**

- NB : la loi normale **N(μ ; σ ²)** est aussi appelée **« loi de LAPLACE-GAUSS »** *(d'espérance μ et de variance σ ²)*



Loi :	**de Bernoulli** B_p ou B_α	**Binomiale** $B(n ; p)$ ou $B(n ; \alpha)$	*Passage au continu*	**Normale** $N(\mu ; \sigma^2)$ ou $N(\mu ; \sigma)$
Variable DISCRÈTE (ou « finie ») **: calculs exacts**			**Variable CONTINUE : calculs approchés**	
Contexte :	Définie sur un **univers à deux éventualités (ou <u>deux issues</u>) :** $x \in \{0 ; 1\}$ avec $p(1)=p \in]0 ; 1[$	Expérience suivant la **loi de Bernoulli B_p** et répétée **n fois** (expériences **indépendantes**)	***Contexte :***	**<u>Conditions de validité de l'approximation</u> :** $n \geqslant 30$ $n.p \geqslant 5$ $n.(1-p) \geqslant 5$ *(p n'est alors pas trop proche de 0 ou 1)* Si **X** suit une **loi binomiale B(n ; p)**, je peux alors utiliser les **fonctions continues** pour faire des **calculs approchés**.
Exemples :	***Pile ou face (<u>un seul lancé</u>) :*** *pile → 0* *face → 1* ***p = 0,5*** *= p(face)*	***Pile ou face (<u>n lancés</u>) :*** *pile → 0* *face → 1* ***p = 0,5*** *= p(face)* ***Combien de fois face (x=1) sur n lancés ?***	*Exemple :*	*Un **appareil en sortie d'usine** a **<u>2 issues</u>** :* ***bon état*** *→ x = 0* ***en panne*** *→ x= 1* *Production :* ***n = 1000.*** *Probabilité de tomber en panne :* ***p =0,08.***
Nombre d'expériences	**1 seule**	**n (indépendantes)**		$n \geqslant 30$
$n_\Omega =$	**2**	**n+1** → *tirage 1 :* $X \in \{0 ; 1\}$ *(2 cas pour face)* → *tirage 2 :* $X \in \{0 ; 1 ; 2\}$ *(3 cas pour face)* → *tirage n :* $X \in \{0 ; 1 ; ... ; n\}$ *soient **n+1 issues face***		*Grand*
Valeurs x :	$x \in \{0 ; 1\}$	$x \in \{0 ; 1\}$		$x \in \{0 ; 1\}$
Variable aléatoire $X_n = k$ *= nombre de fois ou (x = 1) se réalise*	$k \in [0 ; 1]$ → *(X = 1)* *ou* → *(X = 0)*	$k \in [0 ; n]$ → *Face peut sortir 1 fois (X_n = 1), ou 4 fois (X_n = 4), etc...*	***Variable aléatoire*** → $X = k$	**Réel** $k \in I$ <u>NB</u> : $[a;b] \subset I \subseteq \mathbb{R}$
$P(X_n = k) =$	$p^k.(1-p)^{1-k}$ → $P(X=0) = 1-p$ → $P(X=1) = p$	$\binom{n}{k} p^k.(1-p)^{n-k}$ *(Cf. binôme de Newton)* → $P(X_n = 0) = 1-p$ → $P(X_n = n) = p^n$	→ $P(a_X \leqslant X \leqslant b_X) = \int_{a_X}^{b_X} f(x).dx \approx$	$\int_{a_X}^{b_X} \frac{1}{\sigma\sqrt{2\pi}}.e^{\frac{-(x-\mu)^2}{2\sigma^2}}.dx$ *(admis)* *avec :* $f(x) = \frac{1}{\sigma\sqrt{2\pi}}.e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

	→ NB : **P(Ω)** = $\sum_{k=0}^{1} P(X=k)$ = **1**	→ NB : **P(Ω)** = $\sum_{k=0}^{n} P(X_n=k)$ = $(p + (1-p))^n$ = **1** *(binôme de Newton)*		
$E(X_n)$ =	**p** = *p.1 + (1 – p).0*	**n.p** *(admis)*	→ **μ** =	**n.p** *(admis)*
$V(X_n)$ =	**p.(1 – p)** *(Cf ci-dessous *)*	**n.p.(1 – p)** *(admis)*	→ σ^2 =	**n.p.(1 – p)** *(admis)*
$\sigma(X_n)$ =	$\sqrt{p.(1-p)}$	$\sqrt{n.p.(1-p)}$	→ **σ** =	$\sqrt{n.p.(1-p)}$

➢ *Démonstration (*) :* *selon la loi binomiale* B_p *, en choisissant que* $x_1 = 0$ *et* $x_2 = 1$ *:*

$$V = \sum_{i=1}^{n_\Omega} p_i.(x_i-\bar{x})^2 = (1-p)(0-p)^2 + p(1-p)^2 = [(-p)^2 + p(1-p)].(1-p) = p.(1-p)$$

➢ **Théorème de MOIVRE-LAPLACE :**

Si X est une variable aléatoire qui suit la loi binomiale **B(n ; p)**, et **n** → +∞ ;

soit la variable aléatoire : $Z = \frac{X-\mu}{\sigma} = \frac{X-n.p}{\sqrt{n.p(1-p)}}$.

→ Pour tous réels a_Z et b_Z tels que $a_Z \leqslant b_Z$, on a alors : $\lim_{n\to+\infty} P(a_Z \leqslant Z \leqslant b_Z) = \int_{a_Z}^{b_Z} \frac{1}{\sqrt{2\pi}}.e^{\frac{-t^2}{2}}.dt$

- **Conditions de validité de l'approximation** *(comme celles de la loi normale N(μ ; σ²))* **:**
  - ✔ $n \geqslant 30$
  - ✔ $n.p \geqslant 5$
  - ✔ $n.(1-p) \geqslant 5$

  (p n'est alors pas trop proche de 0 ou 1)

- Ce changement de variable aléatoire permet de calculer plus facilement $P(a_X \leqslant X \leqslant b_X)$ car, **quand n** → +∞ : $P(a_X \leqslant X \leqslant b_X) \approx \lim_{n\to+\infty} P(a_X \leqslant X \leqslant b_X) = \lim_{n\to+\infty} P(a_Z \leqslant Z \leqslant b_Z)$.

Le calcul est effectivement plus facile car <u>la variable aléatoire Z suit alors la loi normale centrée réduite N(0 ; 1)</u> *(cf. ci-dessous)* , de densité de probabilité : $f(t) = \frac{1}{\sqrt{2\pi}}.e^{\frac{-t^2}{2}}$.

On peut donc dire en résumé que **<u>si je vérifie les conditions ci-dessus</u>, alors je peux écrire que :**

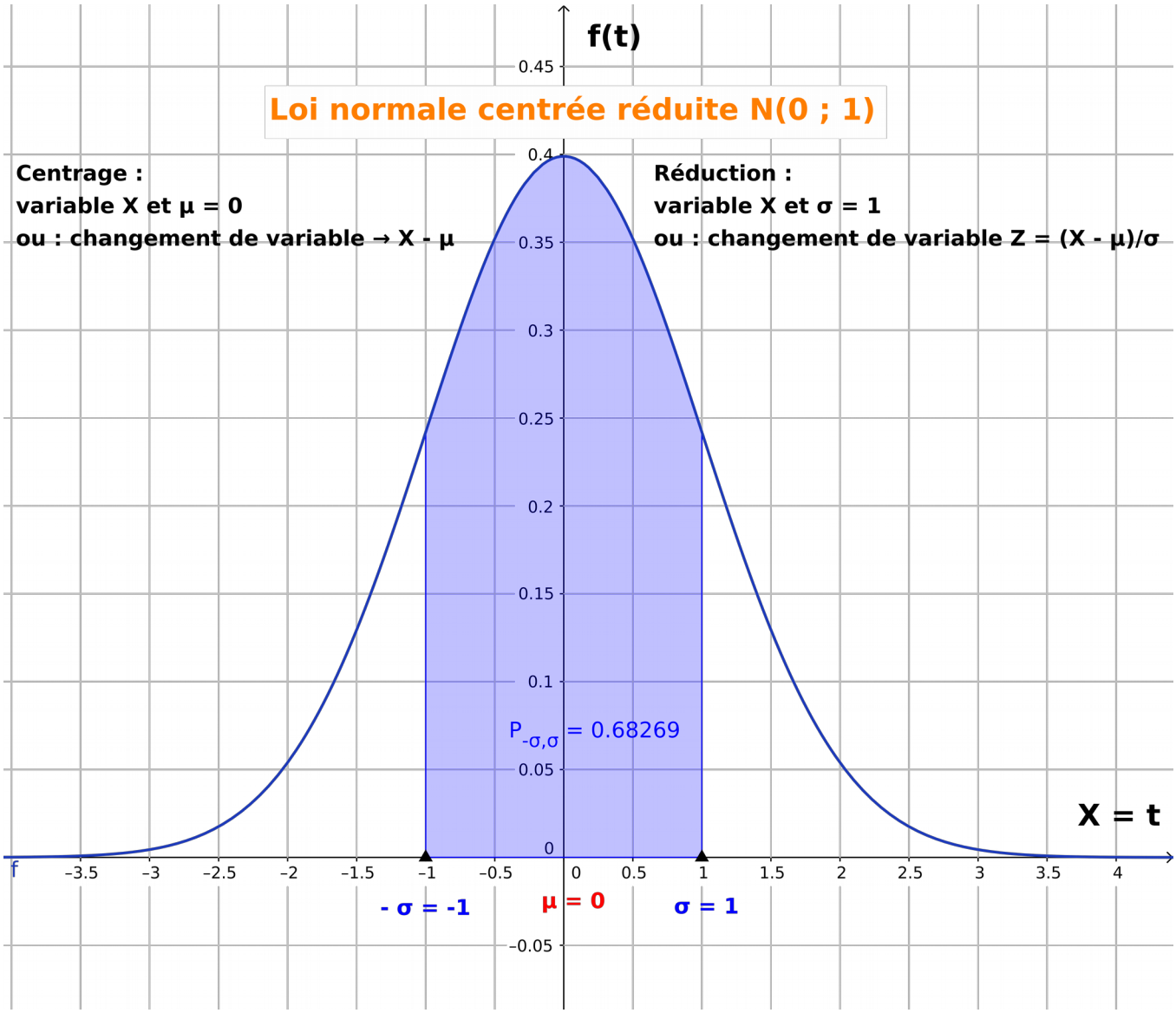
$$P(a_X \leqslant X \leqslant b_X) = P(a_Z \leqslant Z \leqslant b_Z) \approx \int_{a_Z}^{b_Z} \frac{1}{\sqrt{2\pi}}.e^{\frac{-t^2}{2}}.dt$$

✗ **Loi normale centrée réduite** notée : **N(0 ; 1)** :

➢ Densité de probabilité de **Laplace-Gauss** :

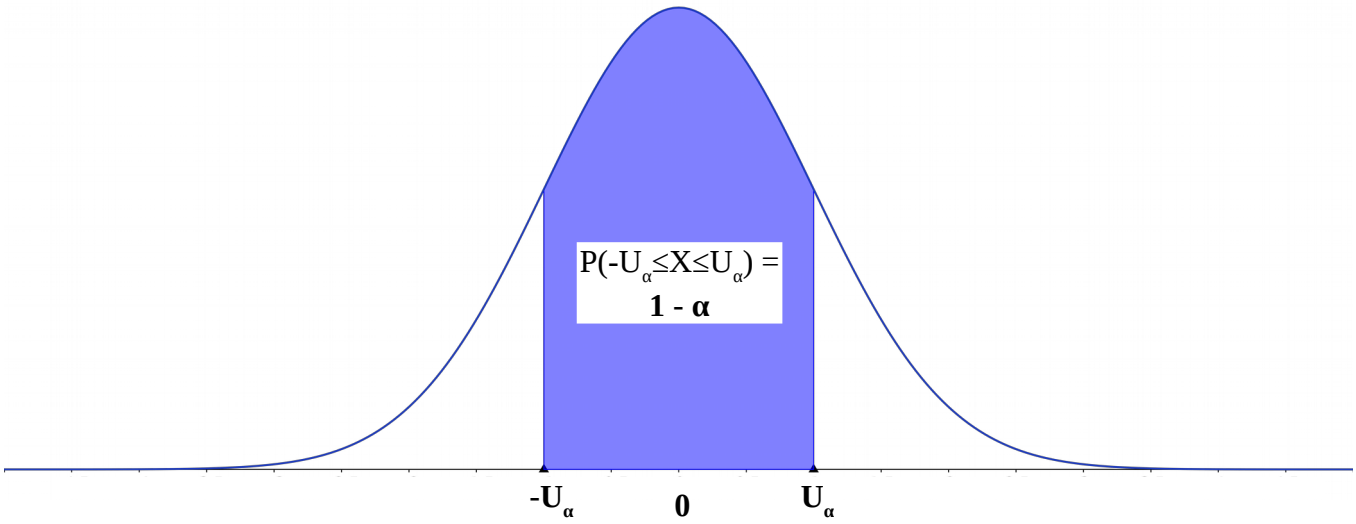
$$f(t) = \frac{1}{\sqrt{2\pi}}.e^{\frac{-t^2}{2}} \quad \text{avec } t \in \mathbb{R}$$

➢ f(t) est représentée par une **« courbe en cloche »** appelée aussi **« courbe de Gauss »** :



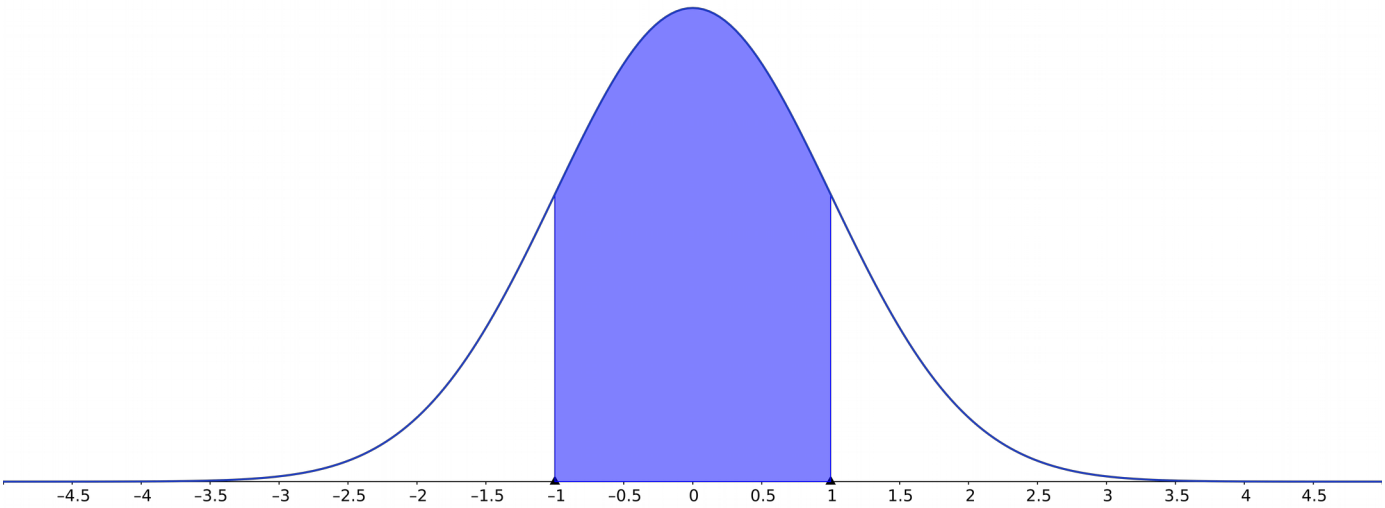
- Cette fonction est paire (symétrie par rapport à (0y)), d'où le **théorème :**

X est une variable aléatoire qui suit la loi normale centrée réduite **N(0 ; 1)**.
Pour tout $\alpha \in]0\ ;\ 1[$, il existe un **unique réel** $U_\alpha > 0$ tel que : $P(-U_\alpha \leqslant X \leqslant U_\alpha) = 1 - \alpha$

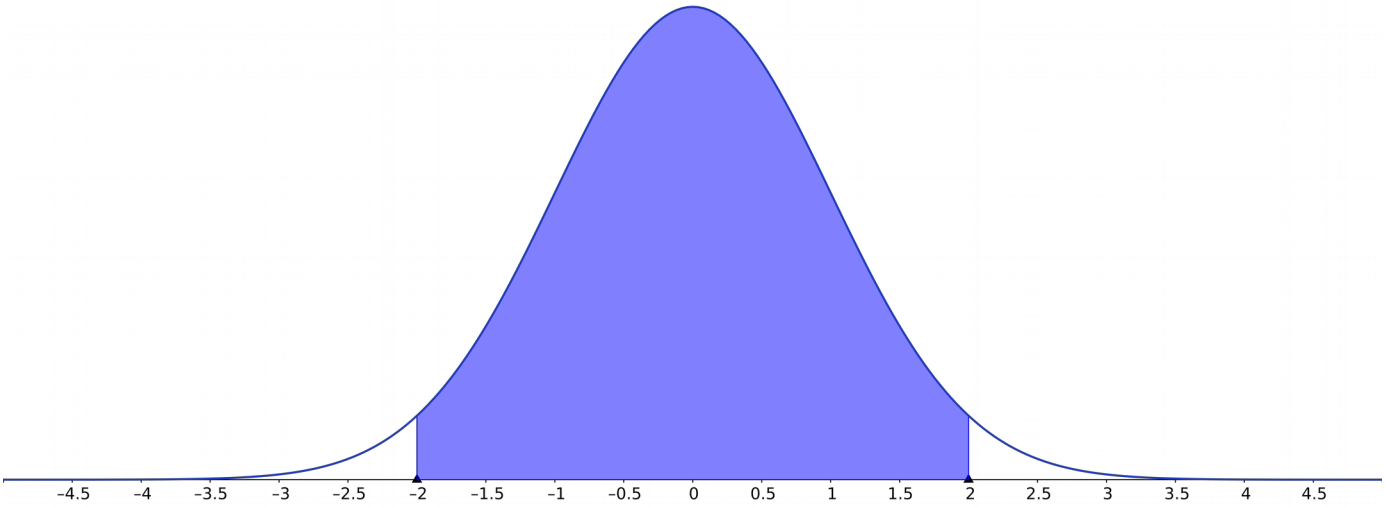


Cas particuliers : $u_{0,05} \approx 1{,}96$ **(95%)** $u_{0,01} \approx 2{,}58$ **(99%)**

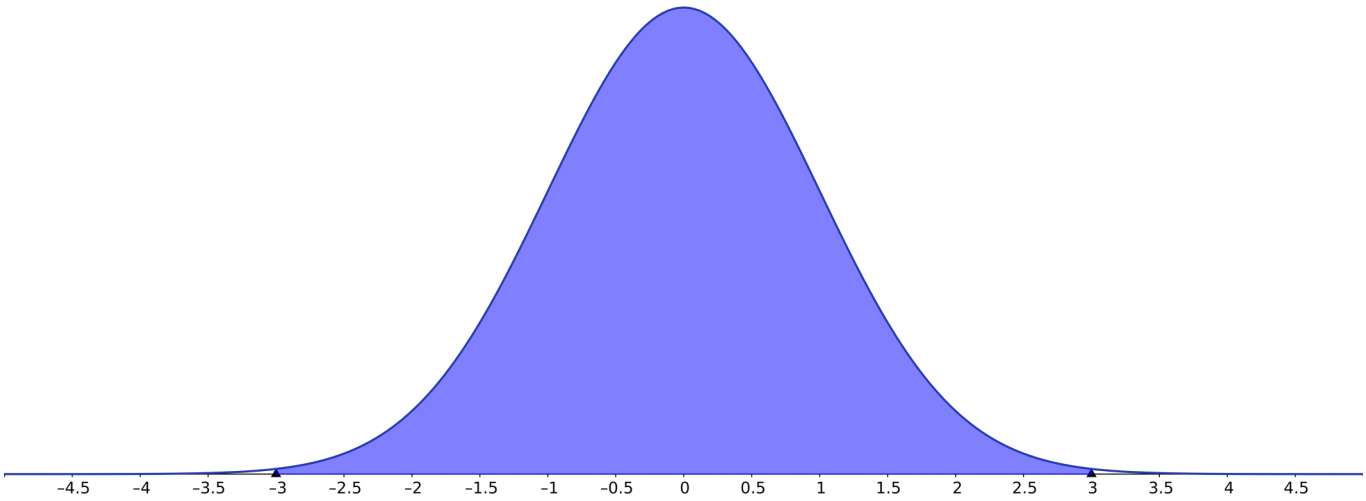
- $\mathbf{P(\mu-\sigma \leqslant X \leqslant \mu+\sigma) \approx 0{,}683}$



- $\mathbf{P(\mu-2\sigma \leqslant X \leqslant \mu+2\sigma) \approx 0{,}954}$



- $\mathbf{P(\mu-3\sigma \leqslant X \leqslant \mu+3\sigma) \approx 0{,}997}$



FLUCTUATIONS ET ESTIMATIONS

✗ **Conditions de validité générales des approximations utilisées ci-dessous** *(celles de la loi normale $N(\mu\ ;\ \sigma^2)$)* **:**

- $\mathbf{n \geqslant 30}$
- $\mathbf{n.p \geqslant 5}$
- $\mathbf{n.(1-p) \geqslant 5}$

  (p n'est alors pas trop proche de 0 ou 1)

✗ **Contexte d'utilisation :**

- Statistiques appliquées à la médecine, l'astronomie, la météorologie…

- *Exemple :* *une urne contient un* ***très grand nombre de boules*** *blanches et de boules noires.*
  Proportion de boules blanches : p
  Proportion de boules noire : 1 – p
  Tirage avec remise (→ événements indépendants) : ***n boules (échantillon)***
  Deux issues *: blanche (x = 1) ou noire (x = 0)*

  Je peux alors pratiquer :
  ➔ *ou bien* ***un échantillonnage*** *:*

  - ***Si je connais p****, je calcule l'****intervalle de fluctuation asymptotique*** ***(centré sur p)*** *de la fréquence, au* ***seuil 1 – α*** *(= 0,95 généralement). Concrètement, je peux donc calculer de façon théorique l'****intervalle dans lequel 95 % des fréquences doivent se trouver****.*

    Je peux en parallèle mesurer la ***fréquence de succès « apparition des boules blanches »*** *pour comparer.*

  - ***Si je ne connais pas p mais peux faire une hypothèse sur sa valeur****, permettant de calculer de même un* ***intervalle de fluctuation asymptotique*** ***(centré sur la valeur hypothétique de p)*** *au* ***seuil 1 – α*** *(= 0,95 généralement), je vais alors pouvoir le comparer à la fréquence réelle. Je dois donc mesurer ou connaître la* ***fréquence de succès « apparition des boules blanches »****. Suivant la correspondance ou non entre l'intervalle et la fréquence connue, je ferai une* ***prise de décision*** *sur la* ***validité de l'hypothèse sur p****.*

  ➔ *ou bien* ***une estimation*** *:*

  Je ne connais ni ne peux faire d'hypothèse sur la valeur de p*. Je dois d'abord mesurer ou connaître la* ***fréquence de succès « apparition des boules blanches »****. Je choisis ensuite le* ***niveau de confiance 1 – α*** *(= 0,95 généralement), puis je calcule à l'aide de la fréquence connue un* ***intervalle de confiance*** *(encadrant p) afin d'****estimer la valeur de p****.*

  ➔ *NB :* ***plus n augmente, plus les intervalles*** *de fluctuation ou de confiance* ***se resserrent****.*

- **Variable aléatoire X_n** = k dénombrant le nombre de succès k *(exemple : « tirer une boule blanche »)* parmi un échantillon de n essais, et qui suit une **loi binomiale B(n ; p)**.

- **Variable aléatoire fréquence de succès $F_n = \dfrac{X_n}{n}$** pour un **« schéma de *Bernoulli (?)* de paramètres n et p »**.

✗ **Échantillonnage :**

➔ **Théorème :**
Soit $\alpha \in]0 ; 1[$.
Soit X_n une variable aléatoire qui suit une loi binomiale **B(n ; p)**.

La **probabilité** que la **fréquence F_n** prenne ses valeurs dans l'intervalle
$I_n = \left[p - u_\alpha . \sqrt{\frac{p(1-p)}{n}} \; ; \; p + u_\alpha . \sqrt{\frac{p(1-p)}{n}} \right]$ se rapproche de **1 – α** quand la taille de l'échantillon **n devient grande**.

On note : $\lim\limits_{n \to +\infty} P(F_n \in I_n) = 1 - \alpha$
I_n est appelé « intervalle de fluctuation asymptotique de la fréquence F_n au seuil 1 - α ».

NB :
- *en général,* ***α = 0,05*** *, donc* ***1 - α = 0,95*** *(seuil de 95%) et* $u_\alpha \approx$ ***1,96****.*
- ***« n devient grand »*** *revient simplement à vérifier les conditions sur n et p validant la loi normale*
- *Ici, on ne cherche pas à calculer F_n mais à trouver un intervalle I_n dans lequel elle a 1 - α chances de se trouver.*

Exemple : *« Un élève a manqué les cours et n'a rien révisé à l'issue d'un chapitre en histoire. L'interrogation comportera heureusement un QCM de 40 questions ! Mais quelles sont ses « chances » de bien répondre en se reposant sur le « pur » hasard, sachant que chaque question comporte 1 seule bonne réponse et 4 propositions de réponse ? »*

✔ *Probabilité de bien répondre à une question :* $p = \frac{1}{4} = 0{,}25$.

✔ ***Conditions de validité des approximations :***
- *échantillon de taille* $n = 40 \geqslant 30$
- $\mu = n.p = 40 \times \frac{1}{4} = 10 \geqslant 5$
- $n.(1-p) = 40 \times \frac{3}{4} = 30 \geqslant 5$

Les conditions sont vérifiées (nous pouvons donc utiliser les théorèmes de ce chapitre)

✔ *Cherchons maintenant, non pas la fréquence F_n, mais l'intervalle I_n dans lequel F_n a une probabilité 1 - α = 1 – 0,05 = 0,95 de se trouver, soit 95 % de chances :*

$$I_{40} \approx \left[0{,}25 - 1{,}96 . \sqrt{\frac{0{,}25 \times 0{,}75}{40}} \; ; \; 0{,}25 + 1{,}96 . \sqrt{\frac{0{,}25 \times 0{,}75}{40}} \right]$$
$$\boldsymbol{I_{40} \approx [0{,}116 \; ; \; 0{,}384]}$$

✔ *Conclusion :*
$$\boldsymbol{P(F_{40} \in I_{40} = [0{,}116 \; ; \; 0{,}384])} \approx \lim_{n \to +\infty} P(F_n \in I_n) = 1 - \alpha = \boldsymbol{0{,}95}$$
En clair, ***il a 95 % de chances d'avoir entre 11,6 % et 38,4 % de bonnes réponses…***

➔ **<u>Propriété (Règle de prise de décision)</u> :**
Soit **f** la fréquence du caractère étudié d'un échantillon de **taille n**.
Soit l'**hypothèse : « la proportion de ce caractère dans la population est p »**.
Soit $\mathbf{I_n}$ l'intervalle de fluctuation asymptotique au **seuil 0,95**.

- Si $\mathbf{f \in I_n}$, alors on **<u>accepte l'hypothèse</u>** faite sur la proportion p.
- Si $\mathbf{f \notin I_n}$, alors on **<u>rejette l'hypothèse</u>** faite sur la proportion p.

<u>NB</u> :

- *On parle de « prise de décision » parce que l'on est en mesure de faire une hypothèse sur la proportion p, ce qui permet de calculer I_n .*

- *On note ici la fréquence « f » (et non « F_n ») car f peut être obtenue par un étude d'un échantillon restreint sur l'échantillon total de taille n (prélèvements de contrôle sur une chaîne de production, ...).*

Exemple : *Reprenons l'exemple précédent, <u>mais</u> supposons maintenant que le QCM de 40 questions comporte par question aléatoirement 3, 4 ou 5 propositions de réponse. Sachant que cette variété dans les proposition semble être plutôt homogène, on formule l'**hypothèse suivante :** $p = \frac{1}{4} = 0{,}25$ .Or nous savons cette fois que l'élève a finalement obtenu une fréquence $f = \frac{2}{5}$ de bonnes réponses, soit 2 fois sur 5 en moyenne.*

✔ ***Conditions de validité des approximations :*** *vérifiées comme précédemment.*

✔ *La fréquence f est connue, mais cherchons l'intervalle I_n dans lequel f a une probabilité 1 - α = 1 – 0,05 = 0,95 de se trouver, soit 95 % de chances :*

$$I_{40} \approx \left[0{,}25 - 1{,}96 \cdot \sqrt{\frac{0{,}25 \times 0{,}75}{40}} \; ; \; 0{,}25 + 1{,}96 \cdot \sqrt{\frac{0{,}25 \times 0{,}75}{40}}\right]$$

$$\mathbf{I_{40} \approx [0{,}116 \; ; \; 0{,}384]}$$

✔ *Prise de décision :*

$$\mathbf{f = 0{,}4 \notin I_n \approx [0{,}116 \; ; \; 0{,}384]}$$

L'hypothèse p = 0,25 est donc rejetée.

➔

✗ **Estimation :**

➔ **<u>Théorème</u> :**
Soit X_n une variable aléatoire qui suit une loi binomiale **B(n ; p)**.
$\mathbf{F_n = \frac{X_n}{n}}$ est la fréquence associée à X_n.

Pour **n suffisamment grand**, p appartient à l'intervalle : $\mathbf{J_n = \left[F_n - \frac{1}{\sqrt{n}} \; ; \; F_n + \frac{1}{\sqrt{n}}\right]}$ avec une probabilité supérieure ou égale à **0,95**.

Remarque :

- *L'intervalle J_n se rapproche de l'intervalle de fluctuation I_n , à la différence près qu'on considère que :*
  - ✔ $U_\alpha = U_{0,05} \approx 2$
  - ✔ $p(1 - p) \approx 0,25$ *(dans les conditions évoquées ci-dessus)*
  - ✔ *la proportion p (issue des probabilités) est remplacée par la fréquence F_n (issue des statistiques).*

- *Il est plus facile de s'appuyer sur la définition ci-dessous pour faire une estimation →*

➔ **Définition :**
Soit **f une fréquence observée** du caractère étudié sur un **échantillon de taille n** *(dans le cadre de la loi binomiale B(n ; p)).*

On appelle **« intervalle de confiance »** de la proportion p au **niveau de confiance 0,95** l'intervalle : $\mathbf{J_n = \left[f - \frac{1}{\sqrt{n}} \; ; \; f + \frac{1}{\sqrt{n}} \right]}$.

Remarque :

- *p étant parfaitement inconnu ici, il n'est pas possible de vérifier les conditions énoncées sur n et p rappelées précédemment. Cependant, il faudra les vérifier en remplaçant la proportion p par la fréquence observée f :*

$$n \geqslant 30$$
$$n.f \geqslant 5$$
$$n.(1 - f) \geqslant 5$$

- *Un niveau de confiance 0,95 signifie qu'on affirme à juste titre que $p \in J_n$ dans 95 % des cas.*

- *Il n'est pas possible de situer p dans J_n , donc on ne peut affirmer ici que J_n est centré sur p.*

Exemple : *« Un institut de sondage interroge 1 052 personnes entre les deux tours de l'élection présidentielle sur leur intention de vote.*
614 déclarent avoir l'intention de voter pour Michelle Coluchi.
En supposant que les votes seront conformes aux intentions, la candidate a-t-elle raison de croire qu'elle sera élue ? »

- ✔ *Fréquence de succès observée dans les intentions :* $f = \frac{614}{1052} \approx 0,584$
- ✔ *Échantillon de taille* $n \geqslant 1052$ *(65 millions d'habitants en France, dont plus de la moitié ont droit de vote ; mais il y a aussi les abstentions…).*
- ✔ *La proportion p finale des votes favorables n'est pas connue en revanche.*
  Je vais donc calculer son ***intervalle de confiance J_n au seuil 0,95****.*
- ✔ *Mais je dois vérifier d'abord le* ***respect des conditions sur n et f :***
  - $n \geqslant 1052 \geqslant 30$
  - $n.f \geqslant \frac{1052 \times 614}{1052} = 614 \geqslant 5$
  - $n.(1-f) \geqslant \frac{1052 \times 438}{1052} = 438 \geqslant 5$ → *Conditions* ***vérifiées****.*
- ✔ *D'où :* $J_{n = 1052} = \left[\frac{614}{1052} - \frac{1}{\sqrt{1052}} \; ; \; \frac{614}{1052} + \frac{1}{\sqrt{1052}} \right] \approx [0,553 ; 0,614]$

  $p \in J_{n \geqslant 1052}$ ***donc p a 95 % de chances d'être supérieur à 0,50 (d'autant que J_n se resserre quand n augmente !) : la candidate peut croire en ses chances !***

Références en ligne **:**

- **Yvan MONKA – Math-et-tiques :**
  https://www.youtube.com/watch?v=4Y12jMMYyVM

- **jaicompris Maths :**
  https://www.youtube.com/watch?v=TnkHUF1Kxeo

- (...)