PrivateTeacher

Cours Privés de Science

La Loi Normale

L'art de chiffrer le hasard

Sommaire:

Introduction

Les statistiques sont l'art de compter les événements.

Pourquoi donc compter ?
Pour pouvoir prédire !

Admettons que je m'intéresse de savoir quelles sont mes chances de rencontrer un chat noir dans une certaine région.

Je me promène donc, et je compte les chats.

À la fin de la journée, j'ai croisé 14 chats et 5 d'entre eux étaient noirs !

À l'aide de ces observations, je peux dire maintenant que si vous vous promenez dans cette même région, alors vous avez vous aussi 5 chances sur 14 de rencontrer un chat noir.

Il s'agit bien là d'une <u>prédiction</u>.

Le <u>pouvoir prédictif</u> des statistiques cependant, est bien plus grand.

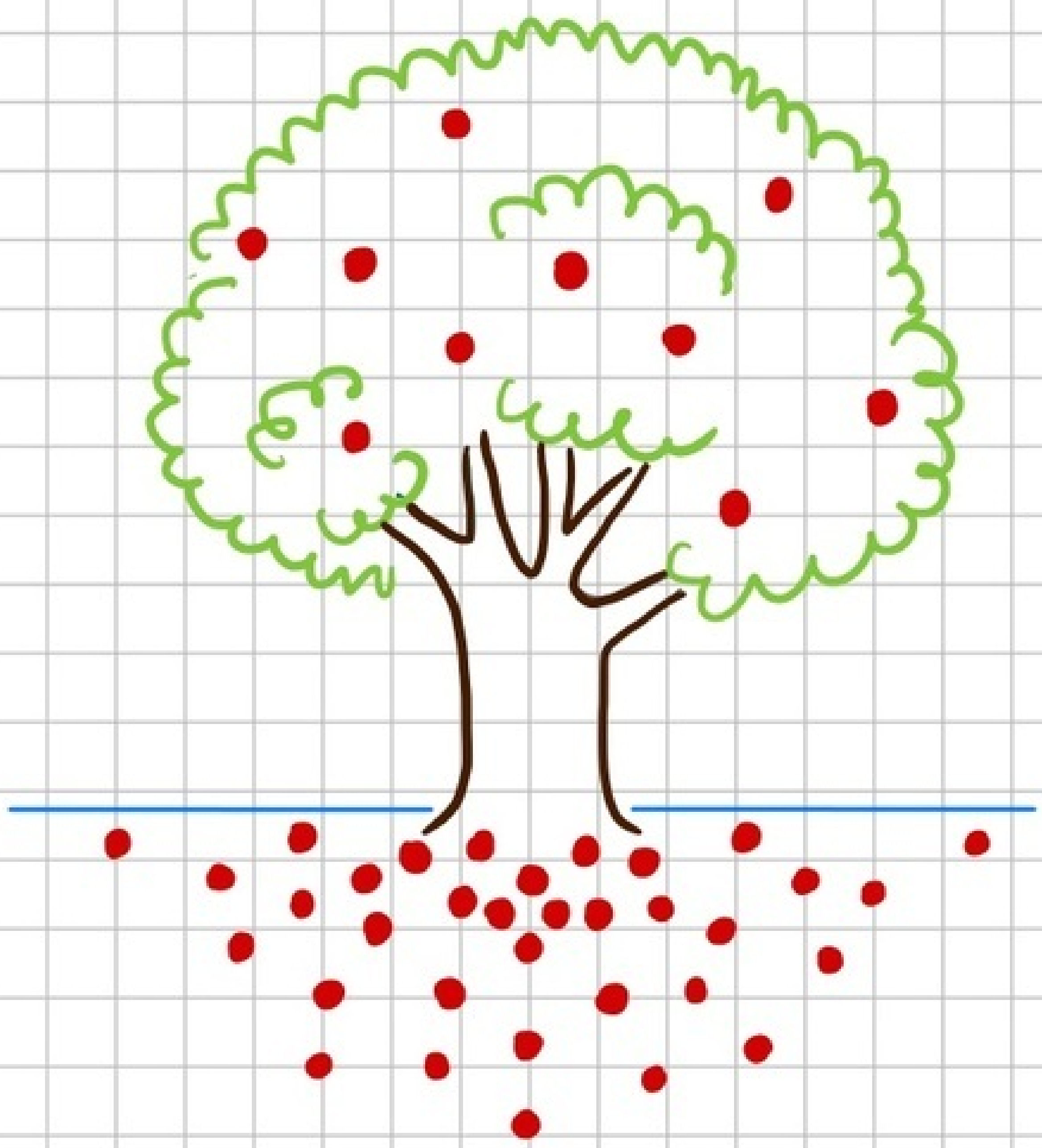
Le pouvoir prédictif des statistiques

Observons des pommes tomber d'un arbre.

Qui peut prévoir à quel endroit précis tombera la prochaine pomme ?

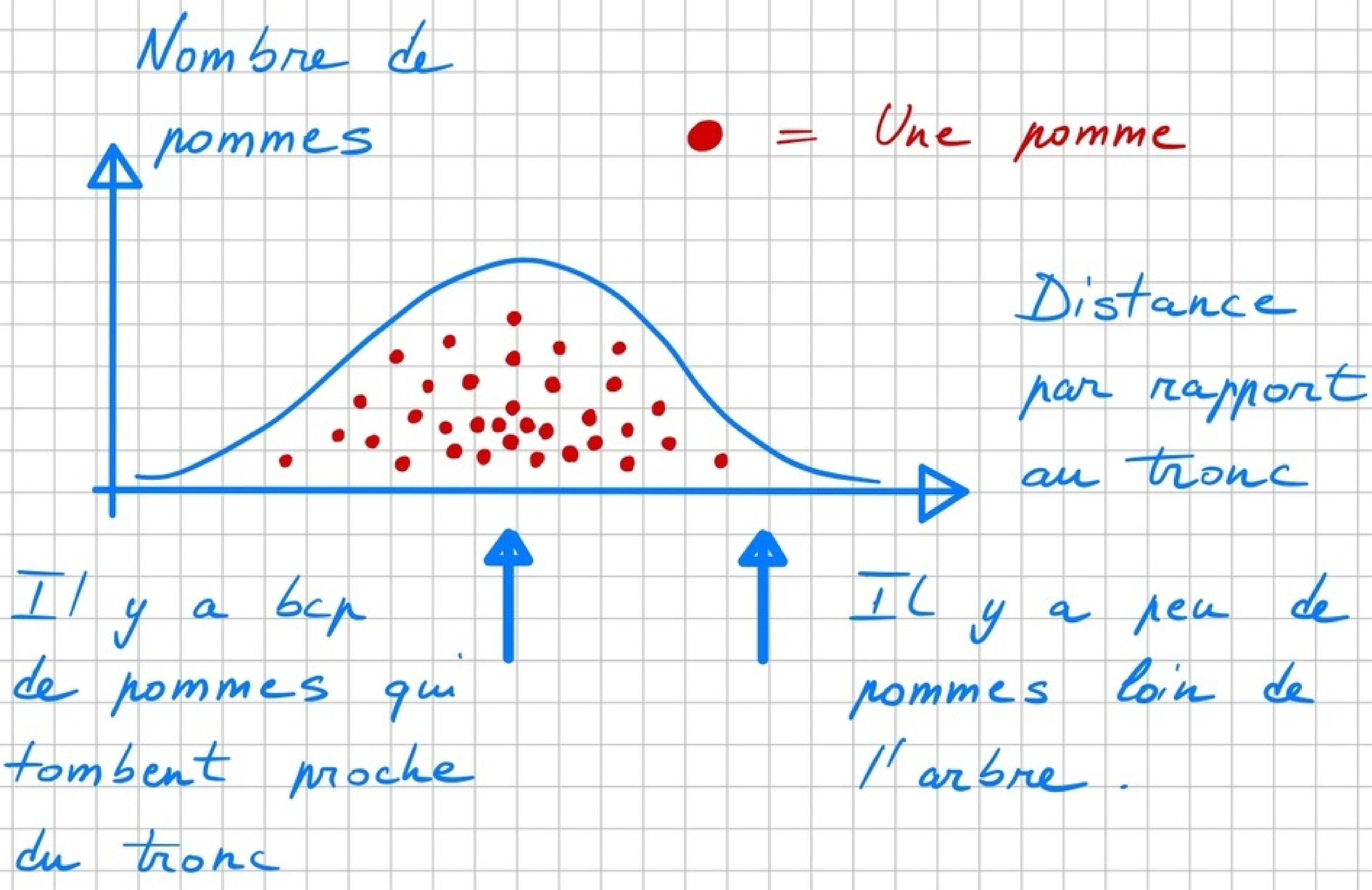
Réponse : Personne !

Une fois que toutes les pommes sont tombées pourtant, on observe qu'elles se répartissent autours de l'arbre de la manière suivante :



Il y a beaucoup de pommes proches du tronc, et il y en a de moins en moins à mesure que l'on s'éloigne de l'arbre.

On peut aussi représenter la distribution des pommes graphiquement :



Si l'on observe un autre pommier aucunes pommes ne tombera au même endroit.

Pourtant, la répartition des pommes autours du tronc sera la même !

Il y a là ce que l'on appel une <u>régularité statistique</u>

C'est de cette régularité que les statistiques tirent leur <u>pouvoir prédictif</u>

Distribution de probabilité

Nous avons vu comment le décompte d'événements donne accès à la probabilité de l'observer.

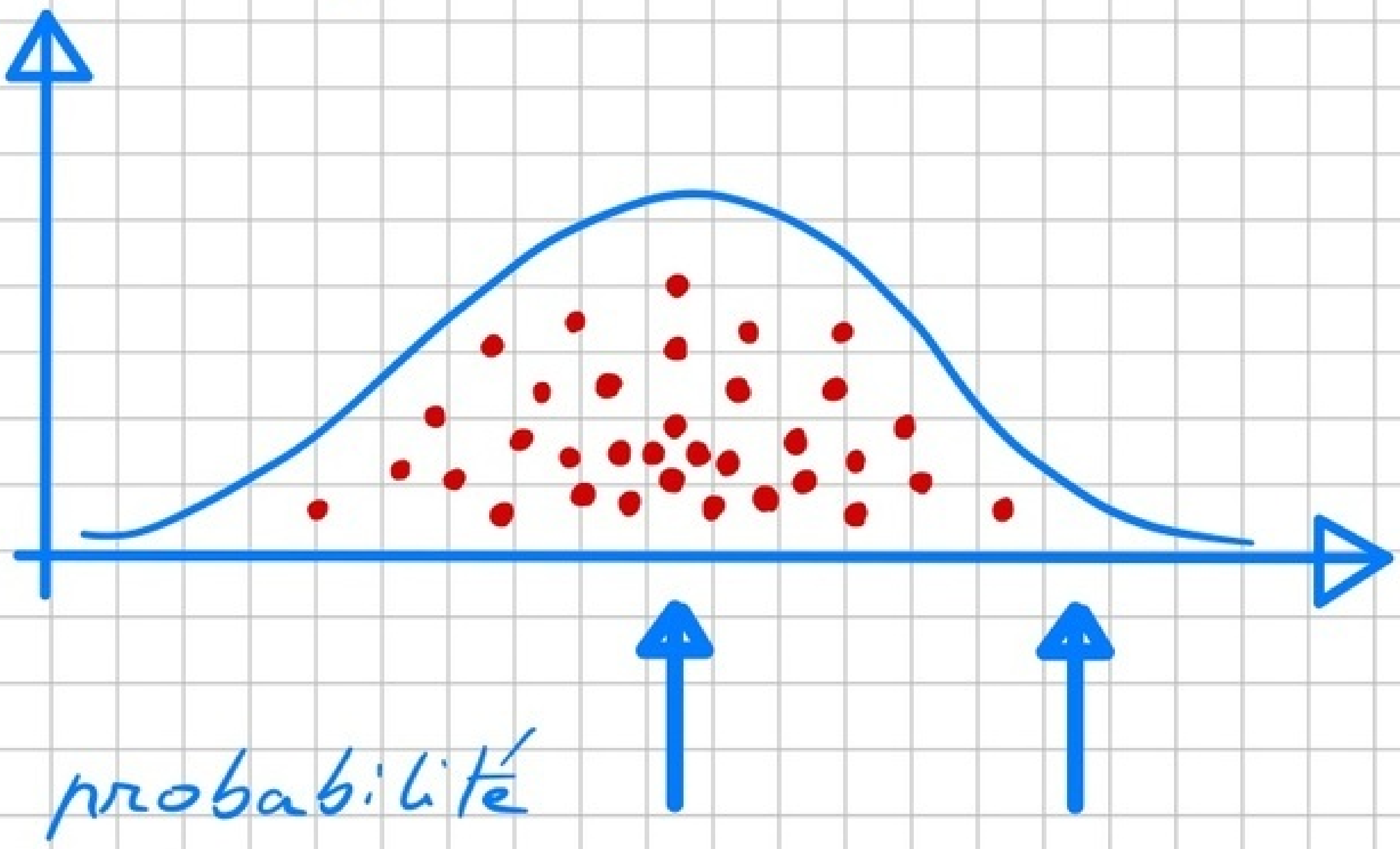
(5 chats noirs sur 14 chats rencontré permet de calculer la probabilité :

$\frac{5}{14} = 0.357$

$= 35.7\ \%$ de chance de rencontrer un chat noir)

On le voit, il suffit de diviser le nombre d'événements qui nous intéresse par le nombre total d'événements

Si l'on dessine maintenant le même graphique que tout à l'heure avec cette fois-ci la probabilité de voir tomber une pomme à une certaine distance de l'arbre plutôt que simplement le nombre de pommes à cet endroit, on a ce que l'on appel une distribution de probabilité.



La probabilité de voir tomber une pomme proche du tronc est importante

Il est très peu probable de trouver une pomme loin de l'arbre.

Une <u>distribution de probabilité</u> permet donc de se représenter un <u>ensemble d'événements</u>

Il existe plusieurs types d'événements différents. Pour cette raison, il existe plusieurs familles de distributions.

La distribution gausienne

La distribution gausienne est une distribution de probabilité qui apparaît lorsque l'on observe des événements aléatoires et indépendants les uns des autres.

Aléatoire, cela signifie que la réalisation d'un événement est possible, mais non certaine.

(la probabilité est une mesure du caractère aléatoire d'un événement)

Indépendant, cela signifie : sans aucun lien les uns avec les autres

(si une pomme tombe à droite, rien ne permet de prédire ou tombera la pomme suivante)

Attention, tous les phénomènes aléatoires dont les résultats sont indépendants les uns des autres ne se distribuent pas nécessairement de façon gausienne. Il existe d'autres distributions.

Forme mathématique

En mathématique, une fonction transforme une valeur x en son image $f(x)$

On note $x \mapsto f(x)$
et on dit : "x a pour image f de x"

Une distribution de probabilité est une fonction qui prend la valeur d'une observation et la transforme en la probabilité de l'observer.

On parle aussi de <u>loi de probabilité</u>

La <u>distribution gausienne</u>, dite aussi distribution normale ou loi normale à la forme mathématique suivante :

$$\boxed{f(x) = C \cdot e^{-\frac{1}{2}x^2}}$$

x représente la valeur d'une observation et $f(x)$ représente la probabilité de l'observer

Quant à C, il s'agit d'une constante

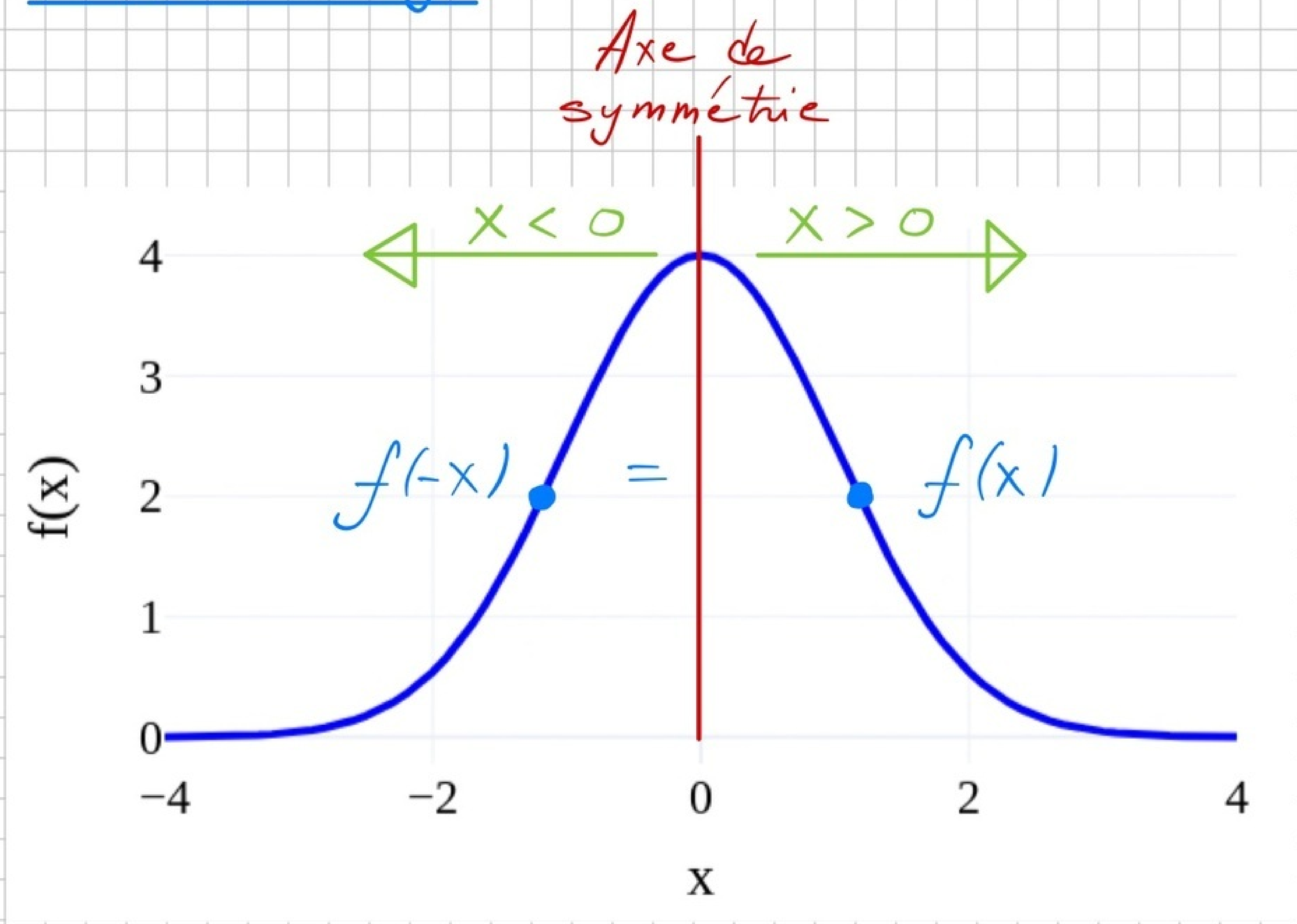
Regardons cette fonction de plus près pour comprendre ce qu'elle signifie :

On remarque tous d'abord que la variable x est élevée au carré.

Tous comme la fct $f(x) = x^2$, la fonction gausienne retourne donc la même valeur pour les x positifs que pour les x négatifs.

On dit que ces deux fonctions sont <u>pairs</u>

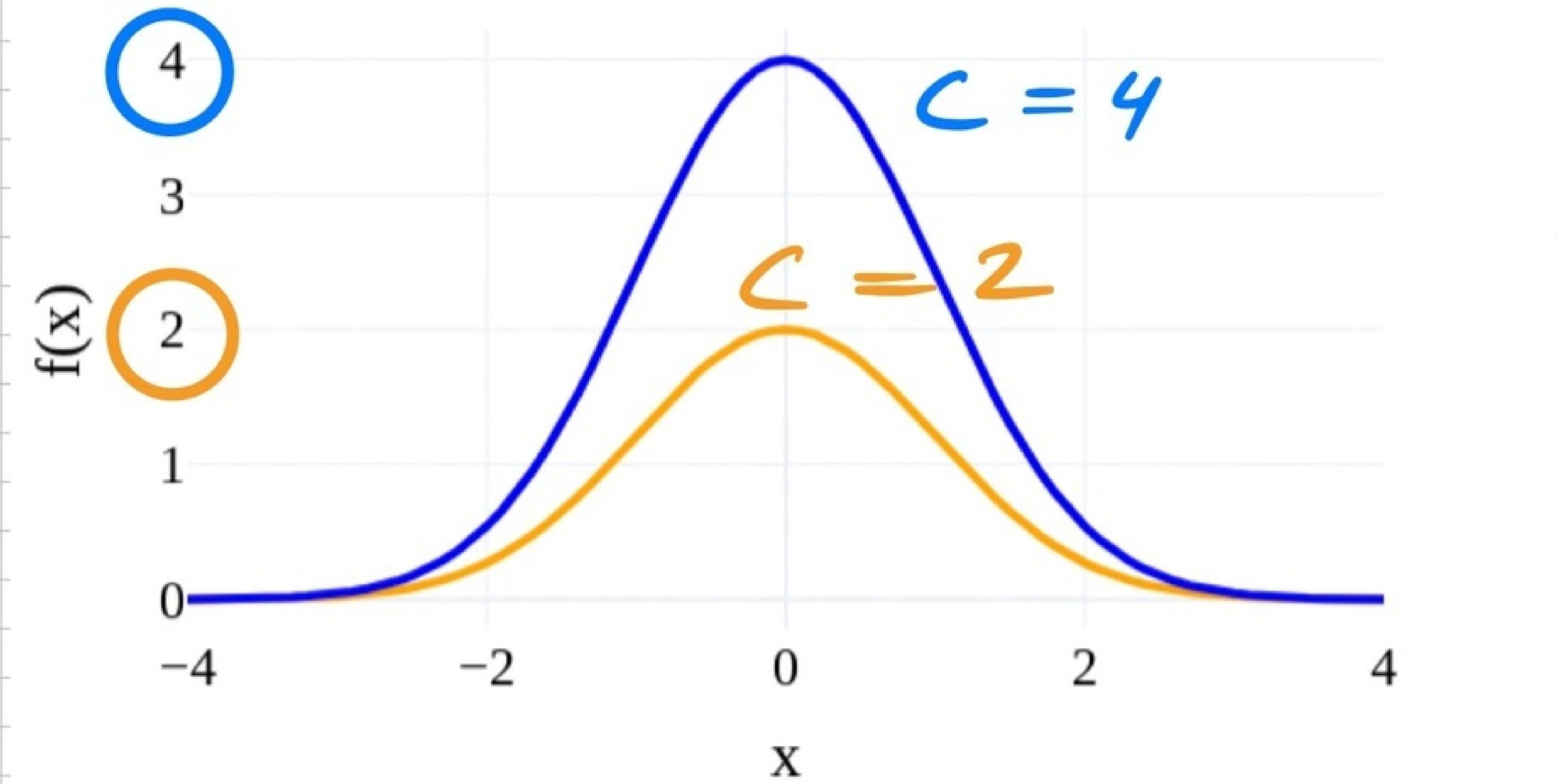
Graphiquement, cela se traduit par une courbe <u>symmétrique par rapport à l'axe des y</u>



Le facteur C est une constante qui multiplie toute la fonction :

$$f(x) = 4 \cdot e^{-\frac{1}{2}x^2}$$

$$f(x) = 2 \cdot e^{-\frac{1}{2}x^2}$$



En $x = 0$, $\exp(-\frac{1}{2}x^2) = \exp(0) = 1$

lorsque $A = 2$, la fct retourne donc $2 \cdot \exp(0) = 2 \cdot 1 = 2$

Dans le cadre des statistiques, on souhaite que l'aire totale sous la courbe soit égale à 1.

En effet, la somme des probabilités de tous les événements vaut 100 % (= 1)

Afin que l'aire sous une fonction gausienne vaille 1, on choisit

$$C = \frac{1}{\sqrt{2\pi}}$$

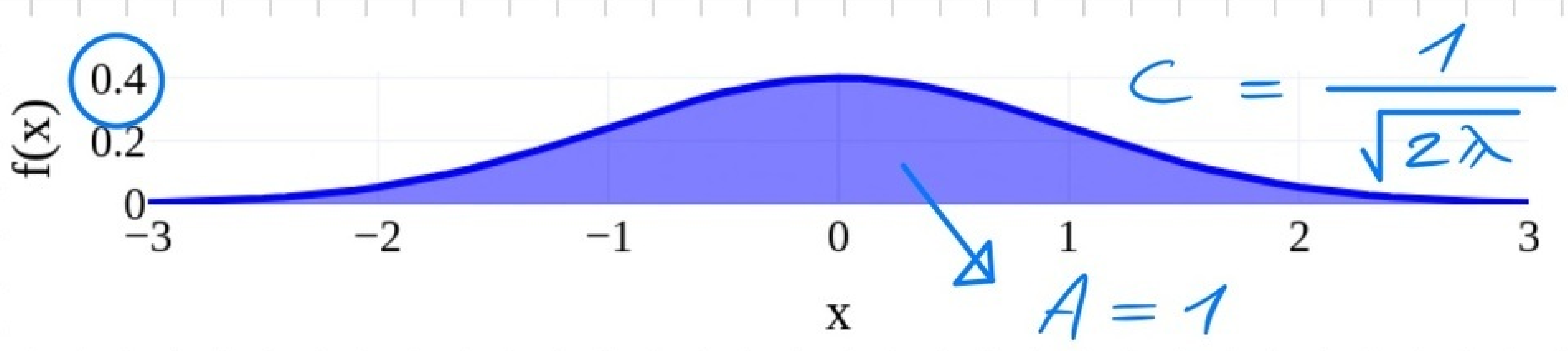
De cette façon, la fonction gaussienne suivante :

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}$$

représente une distribution de probabilité.

x représente la valeur d'une observation, et f(x) représente la probabilité de l'observer.

Sur un graphique, la distribution gausienne se présente sous la forme d'une courbe très aplatie :



L'aire A sous la courbe s'écrit

$$A = \int_{-\infty}^{\infty} f(x)\, dx$$

Terminologie et notation

On note X (grand x)
la <u>mesure du charactère observé</u>.
Cela peut être une taille, un poids,
une couleur ou tout autre
grandeur qui caractérise la population
que l'on souhaite étudier.

X se nomme une <u>variable aléatoire</u>
car elle prend ses valeurs
dans un processus dont l'issue
est <u>incertaine</u>.

On note x (petit x)
la <u>réalisation</u> de la variable aléatoire X

La probabilité de voir se réaliser
une valeur x de la variable aléatoire X
se note <u>$P(X=x)$</u> et se dit :
" probabilité que la variable aléatoire X
prenne une valeur x)

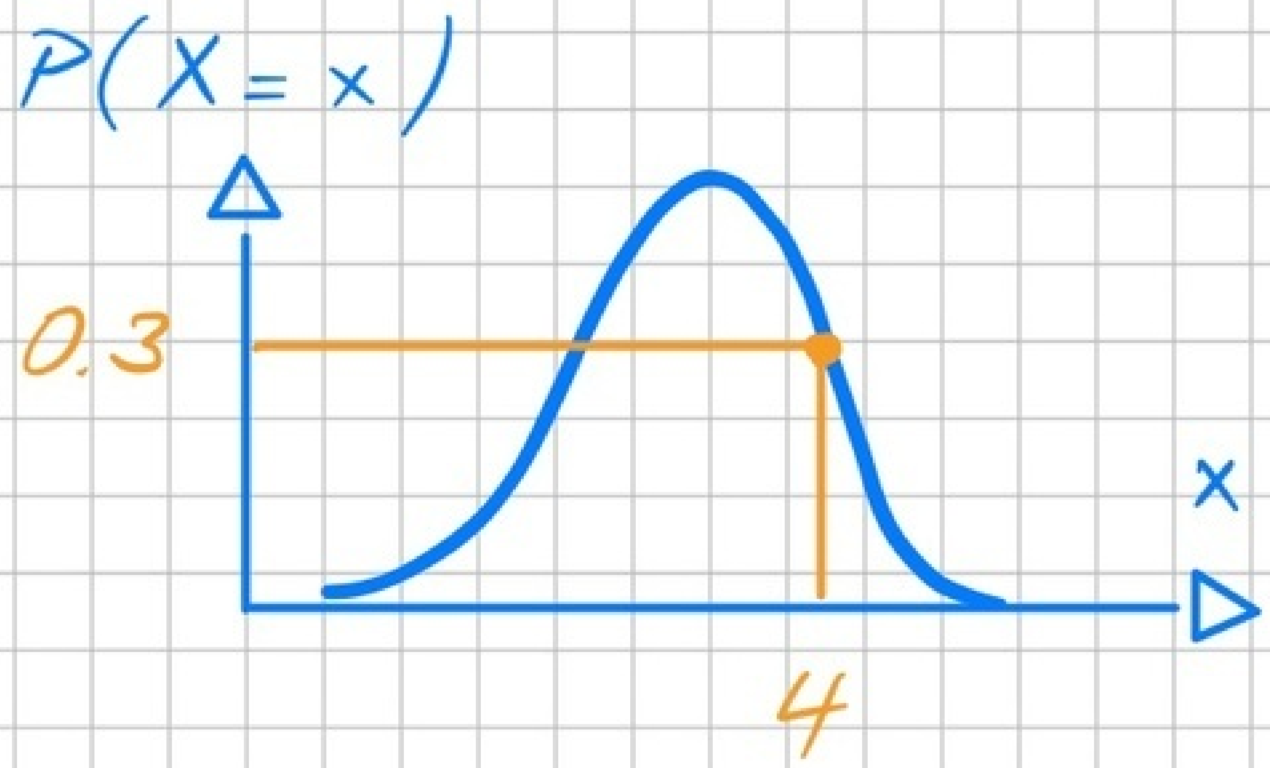
Sur un
graphique :

Calcul de probabilité

Toute l'utilité d'une distribution de probabilité est de permettre de calculer ... des probabilités !

Oui !

Si l'on s'intéresse à la probabilité d'un seul événement, il suffit de la lire sur le graphique :



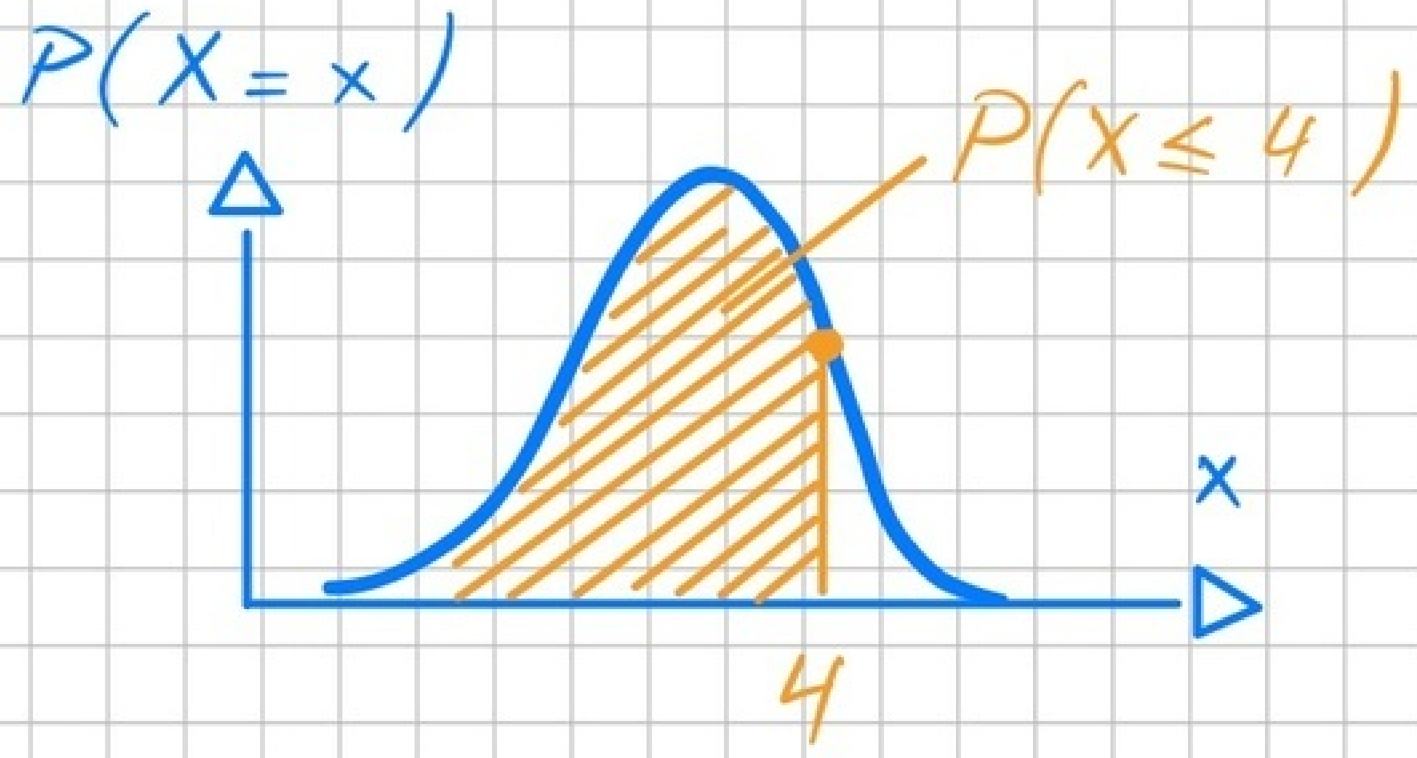
Ici, la probabilité d'observer la valeur 4 est donnée par $f(4) = 0.3$ $(= 30\ \%)$

On note $P(X = 4) = 0.3$

Dans la pratique, il arrive très souvent que l'on s'intéresse à la probabilité d'observer une valeur <u>au moins aussi grande que 4</u>.

On note cela $\underline{P(X \leq 4)}$

Graphiquement, la probabilité $P(X \leq 4)$ représente l'aire sous la gaussienne entre $-\infty$ et 4



Cette aire est donnée par l'intégrale de la fonction gausienne $f(x)$ entre $-\infty$ et 4.

$$P(X \leq 4) = \int_{-\infty}^{4} f(x)\,dx$$

Or, il se trouve que la fonction gaussienne ne se laisse pas intégrer facilement. Il existe même un théorème qui démontre qu'il est impossible d'écrire l'intégrale d'une gausienne (sa primitive) à l'aide des fonctions usuelles simples (théorème de Liouville)

Pour cette raison, on a recours à une <u>intégration numérique</u>.

Table des quantiles

L'intégration numérique est une technique qui permet d'intégrer une fonction sans connaître sa primitive.

Elle repose largement sur la puissance de calcule de nos ordinateurs.

Il fut un temps ou ces calculs demandaient du temps et coûtaient donc cher. Pour cette raison, ils avaient été fait une fois pour toute et les résultats ont été publiés dans une table.

Cette table résume à elle seule tout ce qu'il faut savoir pour calculer n'importe quelle probabilité qui suivrait une distribution normale.

On lui donne le nom de <u>table des quantiles</u>

Je vous la donne ici, c'est cadeau !

Table des quantiles de la distribution normale

X	0	1	2	3	4	5	6	7	8	9
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

Voici maintenant comment utiliser cette table :

On lit la valeur x de la variable aléatoire X dans la première colonne, puis sur la première ligne comme ceci :

(1)

On lit d'abord les dixièmes ici

(1)

$\boxed{x = 0.27}$

(2)

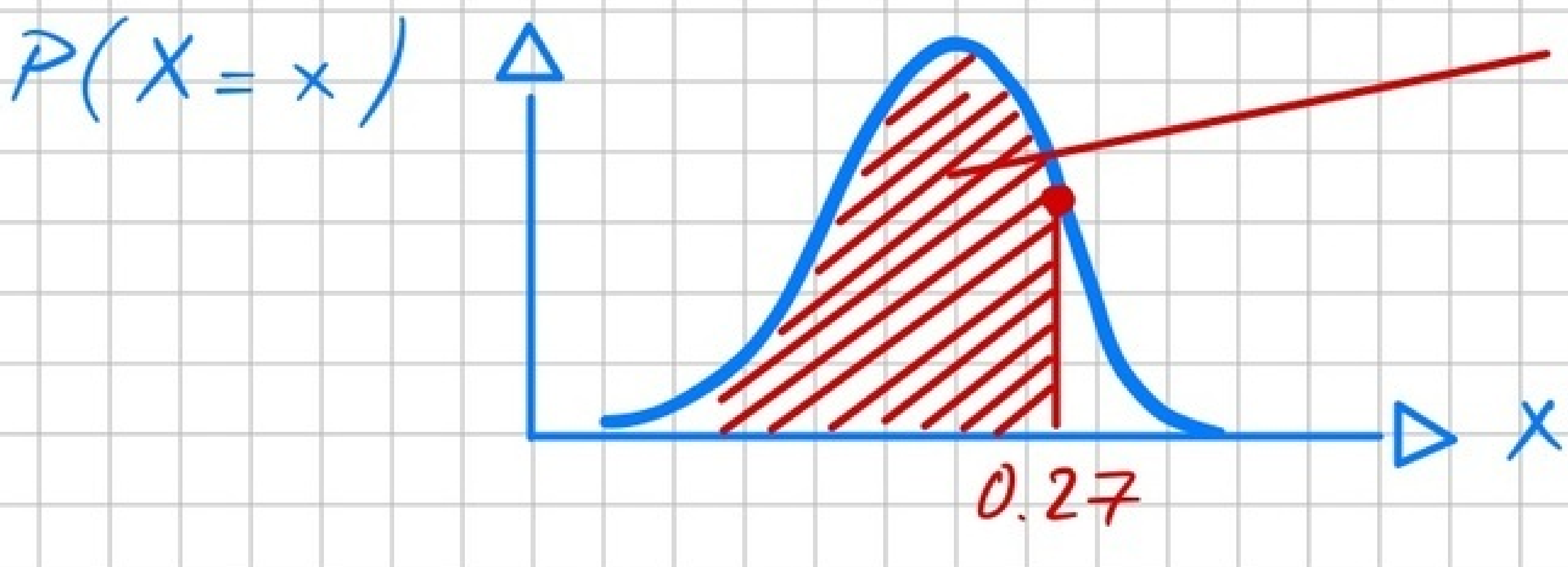
Puis les centièmes là

X	0	1	2	3	4	5	6	7	8	9
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	[illegible]	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	$\boxed{.6064}$	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	[illegible]	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879

À la valeur $x = 0.27$ correspond la valeur 0.6064

On note $\phi(0.27) = 0.6064$

Il s'agit de la probabilité d'observer une valeur plus petite ou égale à 0.27



$\phi(0.27)$
$= P(X \leq 0.27)$
$= 0.6064$

Propriétés de symmétrie

Nous l'avons vu, la fonction gausienne est symmétrique autours de 0.

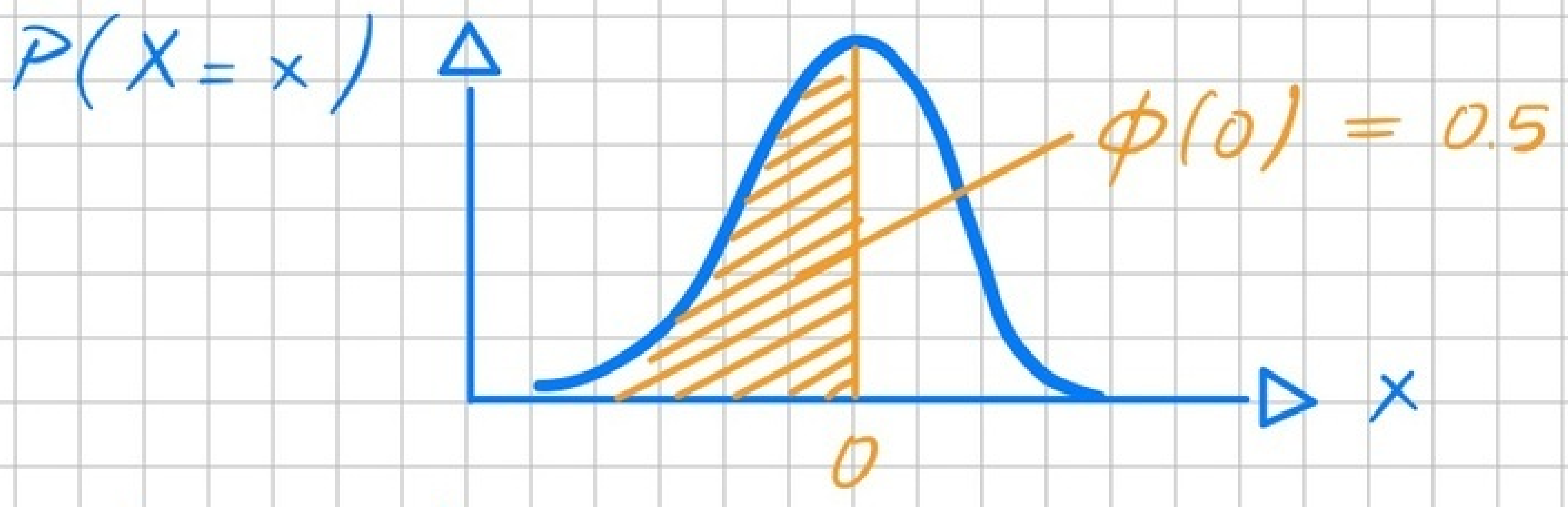
Cette parité est causée par le x élevé à la puissance 2 :

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

Nous savons également que l'aire totale sous la courbe vaut 1 :

$$A_{tot} = \int_{-\infty}^{\infty} f(x)\, dx = 1$$

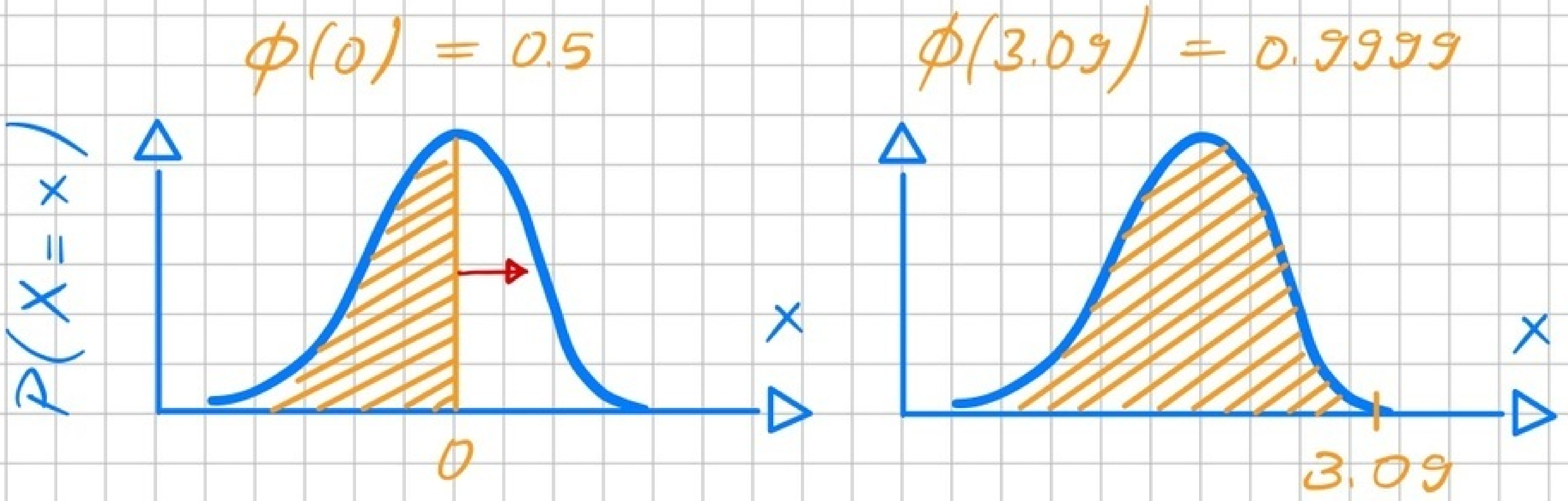
Puisque la fonction est symmétrique autours de 0, on s'attend à retrouver la moitié de l'aire totale à gauche de 0 et l'autre moitié à droite.



Ainsi, lorsqu'on cherche $\phi(0)$ dans la table, on trouve bien 0.5

Si l'on continue de lire la table au delà de $x = 0$, on constate que l'aire sous la courbe ne cesse d'augmenter jusqu'à $x = 3.09$.

À cet endroit en effet, c'est toute la distribution qui est derrière, et l'aire sous la courbe vaut $0.9999 \simeq 1$ comme on s'y attend pour une distribution de probabilité.



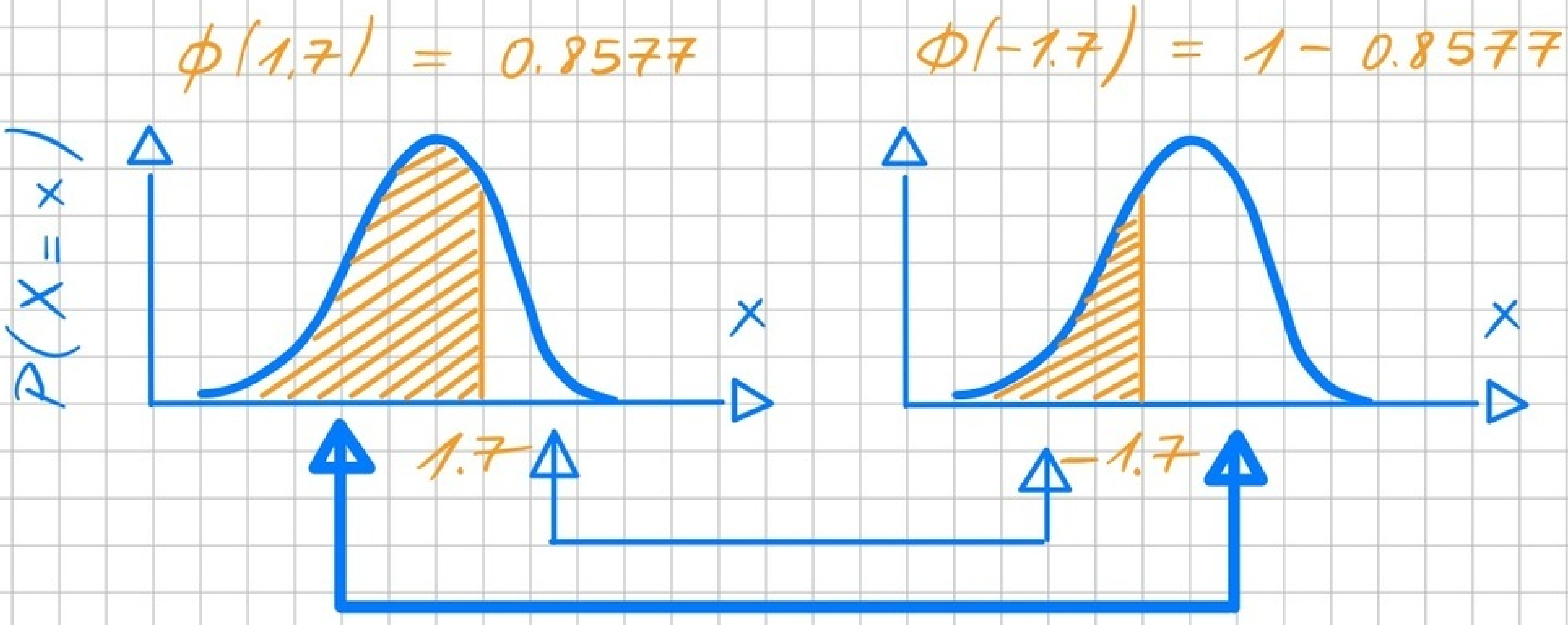
On peut à présent comprendre facilement une propriété de symétrie importante.

Il s'agit pour cela de bien s'imaginer que l'aire sous la courbe augmente vers la droite de 0 de la même manière qu'elle diminue vers la gauche.

Pour cette raison, les valeurs négatives, (à gauche de 0) ne se trouvent pas dans les tables. On peut les calculer facilement à l'aide de l'égalité suivante :

$$\boxed{\phi(-x) = 1 - \phi(x)}$$

Il s'agit là d'une <u>propriété de symmétrie</u> importante que l'on peut représenter de la manière suivante :



Dans la réalité

La gausienne que nous avons étudiée jusqu' ici est un objet mathématique abstrait.

Nous allons voir maintenant comment s'en servire concrètement pour modéliser des observations faites sur le terrain.

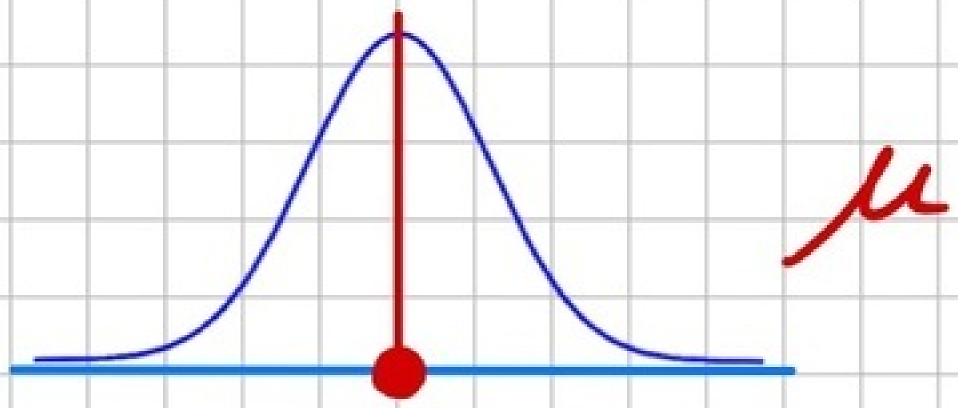
Chaque observation cependant sera différente des autres et on ne peut pas imaginer qu'elle donneront toute la même distribution.

Pour reprendre l'exemple de l'arbre, on s'immagine bien que : des pommes ne se répartissent pas autours d'un pommier de la même manière que des pives autours d'un sapin.

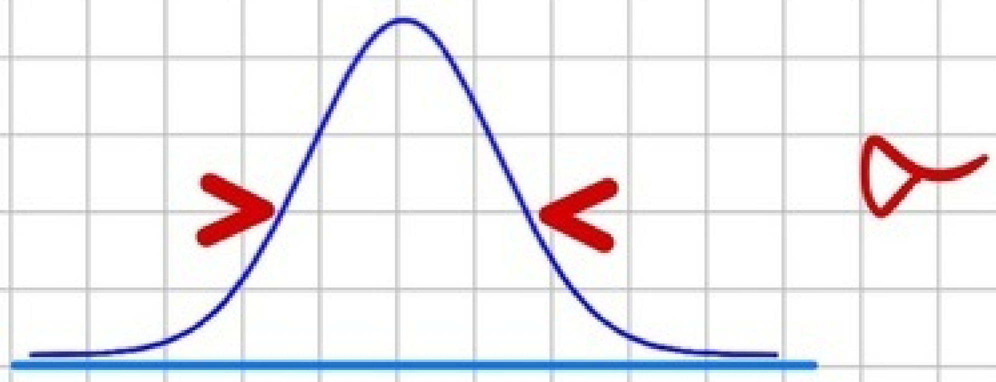
Dans le language des statistiques, on dit que ces fruits ne se distribuent pas de la même manière.

Pour adapter la distribution gausienne à la réalité, on utilise deux valeurs :

1) La position sur x du maximum



2) La largeur de la distribution



Ces deux grandeurs ont une signification concrète et portent donc un nom.

1) On appel <u>moyenne</u> la position sur x du maximum
et on la note μ (mu)

Il s'agit de la valeur autour de laquelle est <u>centrée</u> la gausienne.

2) On appel <u>écart-type</u> l'étendue de la distribution
et on la note σ (sigma)

Il s'agit d'une mesure de la <u>dispersion</u> des valeurs de x autours de la moyenne

Ces deux valeurs permettent d'ajuster la gausienne dans le but de représenter des observations réelles.

Elles peuvent donc varier librement d'une situation à l'autre.
Pour cette raison, on les appel paramètres de la fonction.

On se souvient de la forme mathématique de la fonction gausienne de base :

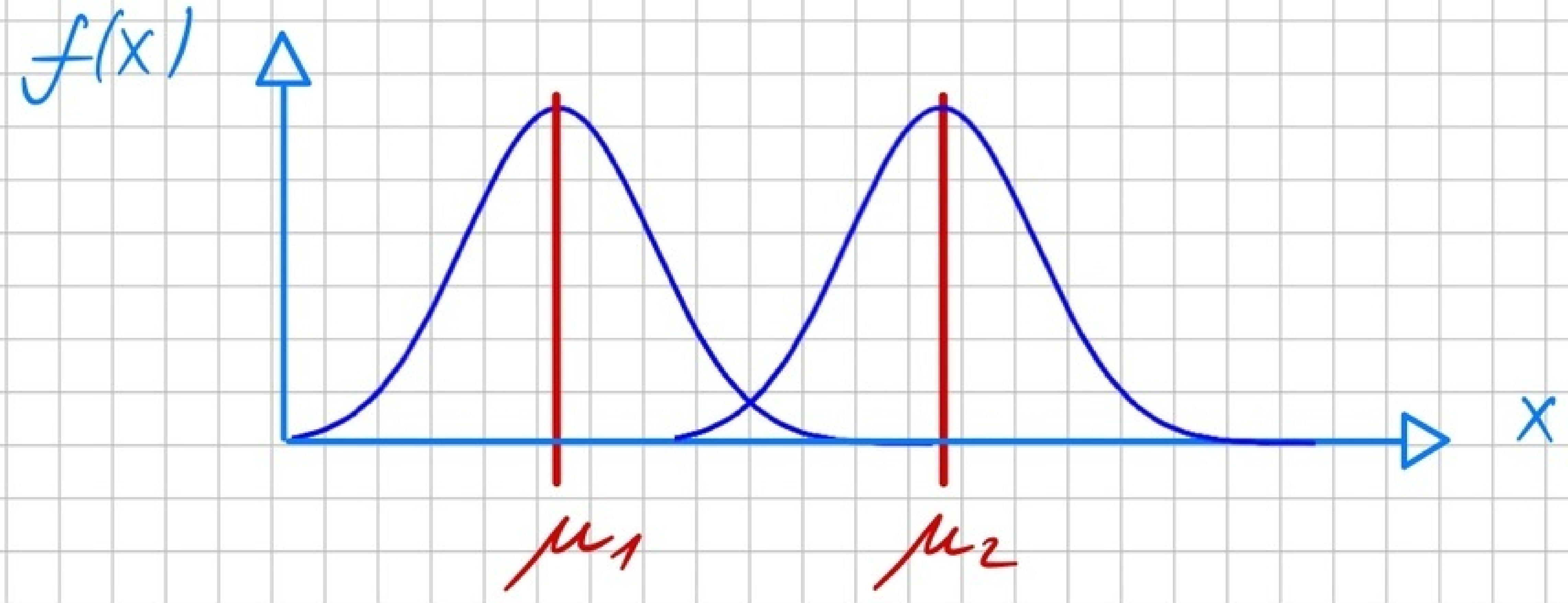
$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

Voici maintenant comment ses deux paramètres la modifient pour la rendre ajustable :

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

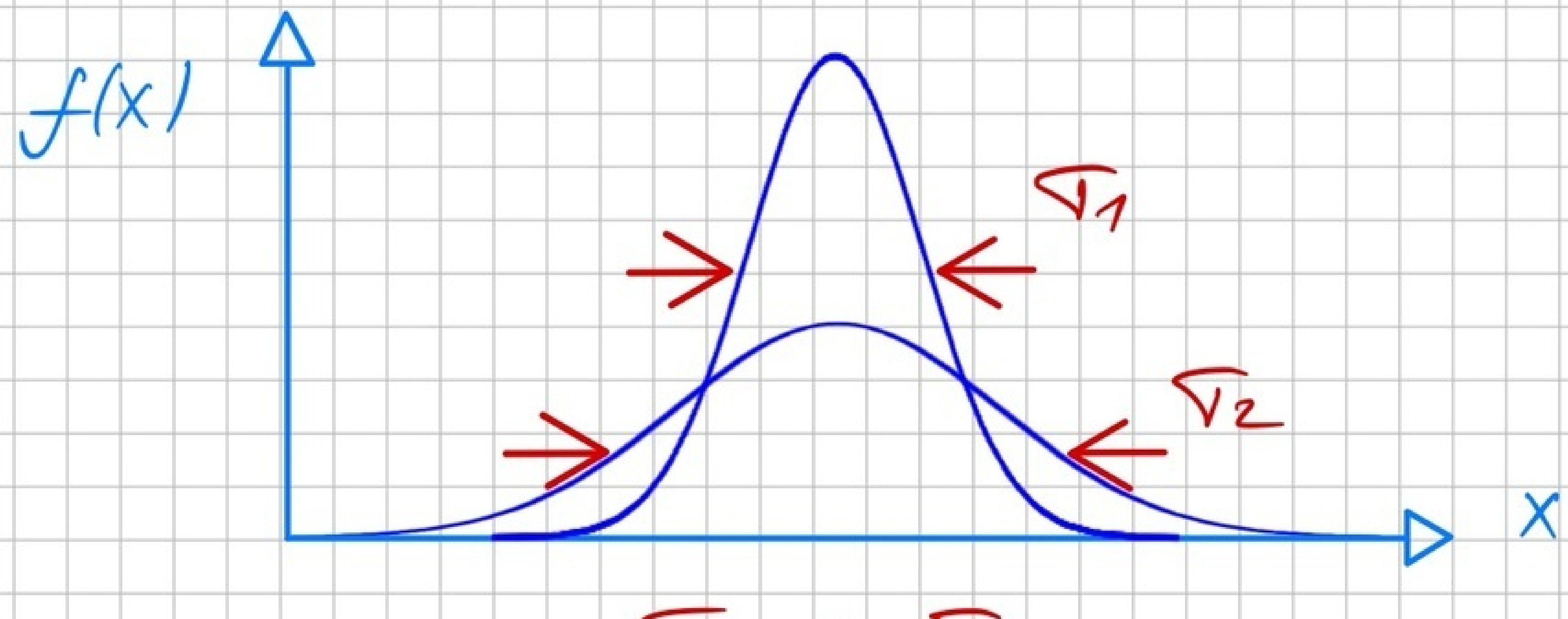
Nous avons là l'expression mathématique de la loi de probabilité gausienne utilisée par les statisticiens et statisticiennes.

Si l'on cherche à se représenter l'action des paramètres sur la fonction, voici ce que l'on obtient

1) 

$\mu_1 < \mu_2$

La moyenne déplace horizontalement la gaussienne (sur l'axe des x)

2) 

$\sigma_1 < \sigma_2$

L'écart-type élargit ou rétrécit la distribution autour de la moyenne. L'aire sous la courbe est conservée et reste donc égale à 1.

Centrer - Réduire

Ces deux paramètres, μ et σ, nous permettent donc d'ajuster la distribution à chaque situation.

Ils sont donc différents d'une situation à l'autre.

On peut dès lors très bien envisager le processus inverse : normaliser chaque distribution particulière vers la même distribution de base.

On appel ce processus "centrer-réduire"

Il permet de se reporter à la table des quantiles pour calculer des probabilités à la main (sans ordi)

1) Centrer consiste à soustraire la moyenne μ à X

2) Réduire consiste à diviser X par son écart-type σ

Ainsi, si une variable aléatoire X suit une loi normale de moyenne μ et d'écart-type σ, on note :

$$X \sim N(\mu; \sigma^2)$$

Alors la variable X^* centrée-réduite s'obtient en effectuant les opérations 1) et 2)

$$X^* = \frac{X-\mu}{\sigma}$$

Cette nouvelle variable à la particularité de suivre une loi normale de moyenne $\mu = 0$ et d'écart-type $\sigma = 1$.

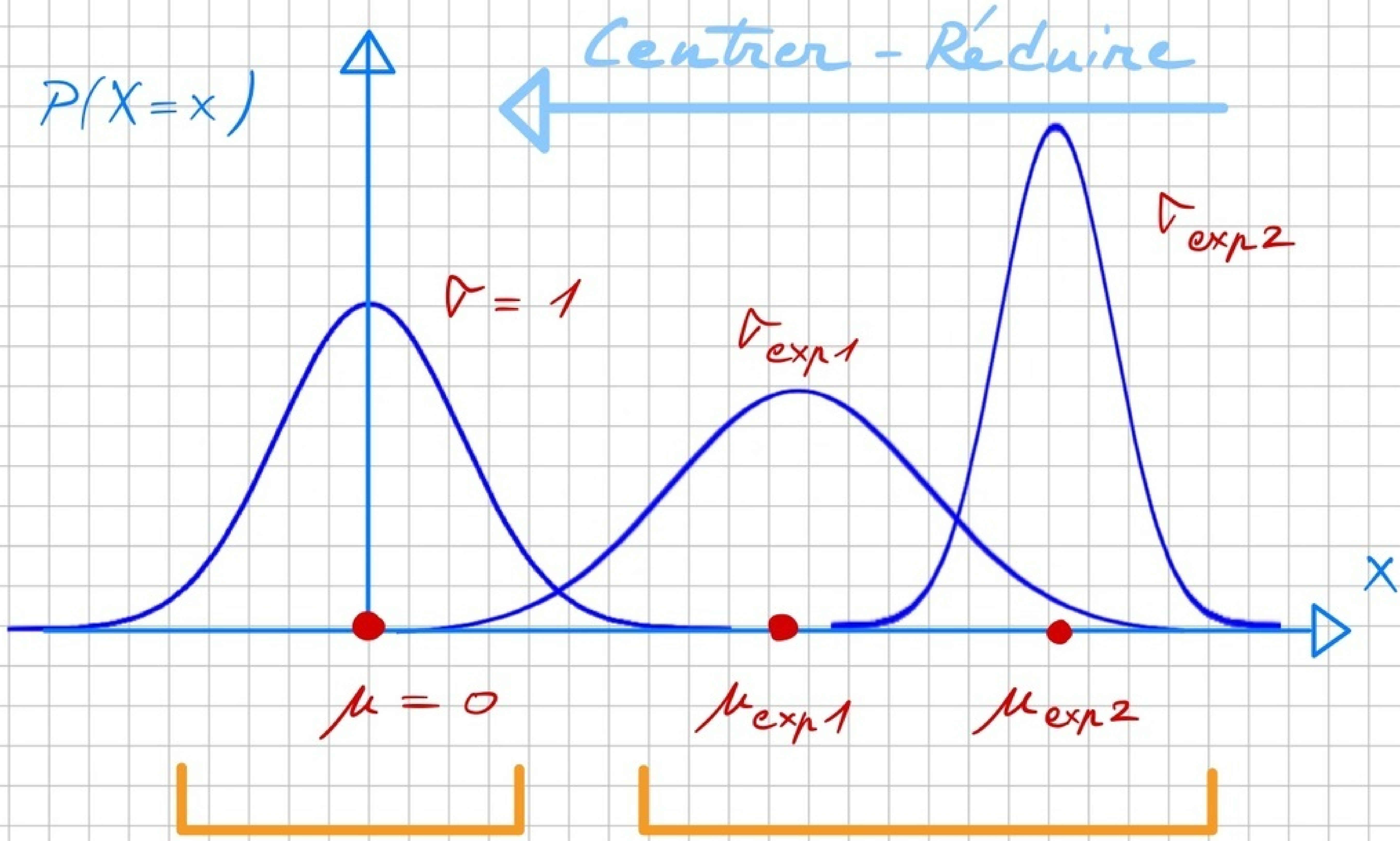
On note : $X^* \sim N(0; 1)$

C'est la distribution qui a servit à calculer la table des quantiles.

La grandeur σ^2 se nomme <u>variance</u>.

Tout comme l'écart-type σ, la variance est une mesure de la dispersion.

On peut représenter graphiquement le processus de centrer - réduire de la manière suivante :



Une seule et même distribution centrée - réduite

$\mu = 0$

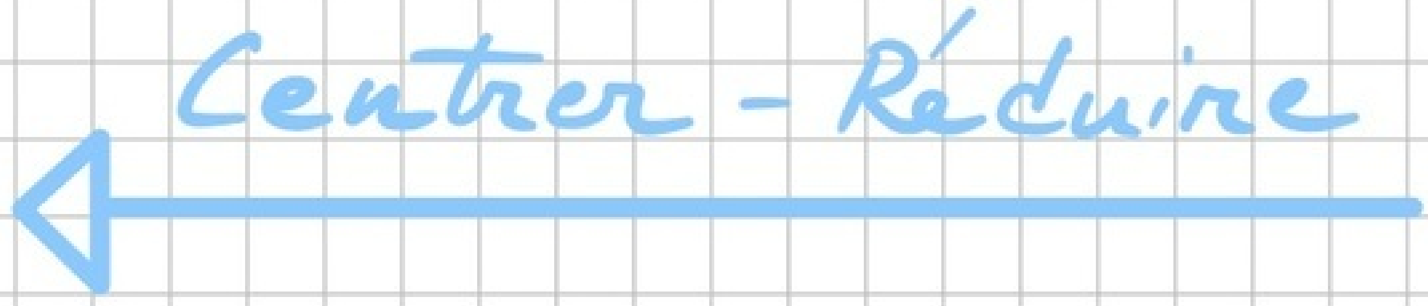
$\sigma = 1$

Deux distributions expérimentales différentes

$\mu = \mu_{exp}$

$\sigma = \sigma_{exp}$

X	0	1	2	3	4	5	6	7	8	9
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990



Les Statistiques
Expliquées Simplement

Cours Privés Individuels

- Progrès rapide
- Réponses précises
- Cours sur mesure

60 CHF / heure

Cours Privés Collectifs

- Travail d'équipe
- Réponse aux questions
- Cours préparé d'avance
- Dès deux personnes

40 CHF / h / Pers.

Modules à Choix

- Maîtriser 1 sujet précis
- Formation intensive
- Exercices pratiques
- Groupe de 5 pers. max
- Durée: 2 heures

40 CHF / heure

Pack 12 Heures

- Se mettre à jour
- Suivi personnalisé
- Méthode de travail
- Exces de renforcement
- 10h + 2h Offertes

50 CHF / heure

Voici mes disponibilités

Genève UNIMAIL :

Lundi et Vendredi
14h00 – 18h00

Lausanne ANTHROPOLE :

Mardi, Mercredi et Jeudi
14h00 – 18h00

Visio TELEGRAM :

Lundi – Vendredi
9h00 – 11h00
18h00 – 20h00

Réservation Calendrier
(Min 2 jours à l'avance)

Horaires WEEKEND

J'ai la possibilité de mettre à disposition des périodes durant le weekend si la situation le demande.

Mes tarifs sont alors majoré de 50%