64ACG610 – Les outils statistiques de gestion

La statistique est une branche des mathématiques qui a pour objet l'analyse et l'interprétation de données quantifiables.

On a coutume de distinguer **la statistique descriptive et l'inférence statistique.**

La **statistique descriptive**, comme son nom l'indique, sert à décrire les caractéristiques (âge, revenu…) des individus d'une population donnée à l'aide de tableaux, de graphiques, de mesures de tendance centrale (moyenne, médiane…) et de dispersion (écart-type…).
Par exemple, l'étude de la distribution des salaires dans une entreprise rentre dans ce cadre. Calculer la moyenne d'une classe ou tracer la courbe d'évolution du chômage depuis 2020 relèvent aussi de la statistique descriptive.

L'inférence statistique permet de généraliser les données tirées des échantillons à un ensemble plus vaste, la population.
L'inférence comporte deux volets : **l'estimation et le test d'hypothèse.**

L'estimation joue un grand rôle dans les sondages d'opinion et les études de marché. Dans ces deux cas, l'objectif est d'obtenir rapidement une idée précise de l'opinion ou des goûts d'une grande population.
Un recensement étant hors de question en raison de son coût et de la lenteur du processus, on recueille l'information voulue auprès d'un échantillon jugé représentatif et on généralise les résultats obtenus à l'ensemble de la population.
Par exemple lorsque j'essaie de prédire les intentions de vote à partir d'un échantillon de 1000 personnes je fais de la **statistique inférentielle.**

Les **tests d'hypothèses** servent à faire un choix entre deux hypothèses à partir des renseignements tirés d'un échantillon. Par exemple, nous faisons une affirmation (ou hypothèse) sur un paramètre de la population (un responsable pédagogique suggère que le niveau moyen du QI des étudiants qu'il a en charge est de 110), puis nous vérifions si cette hypothèse est vraie en examinant un échantillon aléatoire pris dans la population (hypothèse vérifiée si la moyenne de l'échantillon prélevé est comprise entre 109 et 111 par exemple).

Avant d'aborder sereinement les théories d'échantillonnage, champ d'étude de la statistique inférentielle, nous ferons un rappel des notions (variables aléatoires, loi normale…) qui seront souvent sollicitées.

64ACG610 – Les outils statistiques de gestion

1) Les variables aléatoires

a) Les variables aléatoires

Étant donné une **expérience aléatoire** (dont on ne connaît pas le résultat à l'avance[1]), on appelle **variable aléatoire** toute grandeur numérique dont la valeur dépend de l'issue de l'expérience.

Exemple : Soit un jeu qui consiste à lancer un dé bien équilibré. Lancer le dé une fois constitue une **épreuve** et le résultat de cette épreuve (par exemple obtenir un 2) est un **évènement** (ou éventualité ou issue ou résultat). On a donc :

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

On appelle **univers** (ou ensemble fondamental) l'ensemble de tous les résultats (ou éventualités ou évènements) possibles au cours d'une épreuve.

- si le résultat est 2, 3, 4 ou 5, on gagne 1.50€ ;
- si le résultat est 1 ou 6, on doit verser un montant égal au résultat sur le dé.Ainsi, on paye 1€ si le résultat est 1 et 6€ si le résultat est 6.

Dans cet exemple, la variable aléatoire X représente le gain (ou la perte) du joueur. Ses valeurs possibles sont -1, -6 et 1.50. On notera X (Ω) = {-1, -6, 1.50}. Ainsi X est une variable parce qu'elle peut prendre différentes valeurs. Et X est aléatoire parce que la valeur qu'elle prend dépend du hasard.

En associant à toutes les valeurs d'une variable aléatoire la probabilité qui lui correspond, on définit une **loi de probabilité** (ou fonction de distribution). Dans notre exemple, cette loi est donnée par le tableau ci-dessous :

x_i	-6	-1	1.50	
p_i	1/6	1/6	4/6	1

La loi de probabilité de X est l'ensemble des valeurs possibles de X (les x_i) et les probabilités de chacune d'elles (les p_i).La loi de probabilité permet donc de déterminer la probabilité de chaque modalité de la variable aléatoire.

[1] Le résultat est donc uniquement déterminé par le hasard. Par exemple lancer une pièce à deux faces ou jeter un dé à six côtés sont des expériences aléatoires.

La loi de probabilité *de X* satisfait toujours les conditions suivantes :

$$\text{(i) } p(x_i) \geq 0 \quad \text{et} \quad \text{(ii) } \sum_{i=1}^{n} p(x_i) = 1$$

Nous avons ici défini une **variable aléatoire discrète** car elle ne peut prendre que certaines valeurs bien précises. Ainsi la variable X « nombre de vélos défectueux dans une station de 20 vélib » est discrète car elle n'a que 21 valeurs possibles, de 0 à 20.

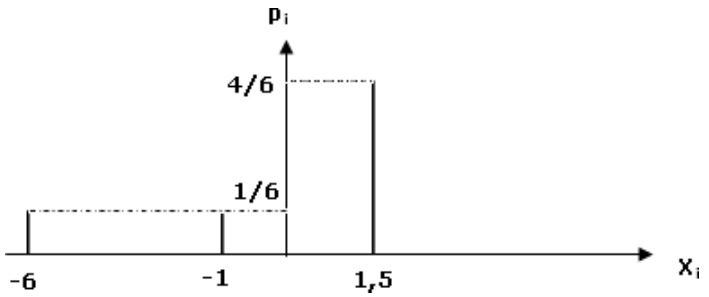
Dans notre exemple, la variable aléatoire *X* prend les trois valeurs -6, -1 et 1.50. Si on pose :

$$x_1 = -6,\ x_2 = -1 \text{ et } x_3 = 1.50$$

On constate que la variable aléatoire ne peut prendre qu'un nombre fini de valeurs. Si on note p_i la probabilité pour que *X* soit égal à x_i. Alors,

$$p_1 = 1/6,\ p_2 = 1/6 \text{ et } p_3 = 4/6$$

Les représentations graphiques d'une loi de probabilité sont obtenues en portant en abscisse les valeurs de la variable aléatoire et en ordonnée les probabilités. La **fonction de distribution** est tout simplement la représentation de la loi sous la forme d'un diagramme en bâtons, la variable aléatoire étant discrète. On observe que les segments verticaux sont proportionnels aux probabilités.



On pourrait, comme on le fait en statistique descriptive, introduire une fonction cumulative, dite **fonction de répartition *F***. Elle indique la probabilité pour que la variable aléatoire *X* prenne une valeur *au plus* égale à une valeur donnée *a* :

$$F(a) = P(X \leq a)$$

Enfin la **variable aléatoire** X est dite **continue** lorsqu'elle peut prendre toutes les valeurs d'un intervalle de ℝPar exemple si le nombre de connexions Internet sur

un site est une variable aléatoire discrète, le temps de connexion est une variable aléatoire continue[2].

Dans le cas d'une variable continue, il est impossible de présenter toutes les valeurs possibles du caractère étudié (théoriquement, il y en a un nombre infini). On est obligé de regrouper les données par **classes** (par exemple des tranches de revenu ou d'âge). On obtient alors une distribution qui peut être représentée par un **histogramme**. C'est un graphique où chaque classe est représentée par un rectangle dont la largeur est proportionnelle à l'amplitude de la classe et dont la hauteur est proportionnelle à l'effectif de la classe.

Ex : Considérons la répartition de 100 ménages selon le revenu du chef de famille donnée par le tableau suivant :

Revenus	Nombre de ménages	Effectifs corrigés
[3000 ; 5000[	**30**	30/2 = 15
[5000 ; 6000[	**40**	40
[6000 ; 7000[	**20**	20
[7000 ; 9000[	**10**	10/2 = 5

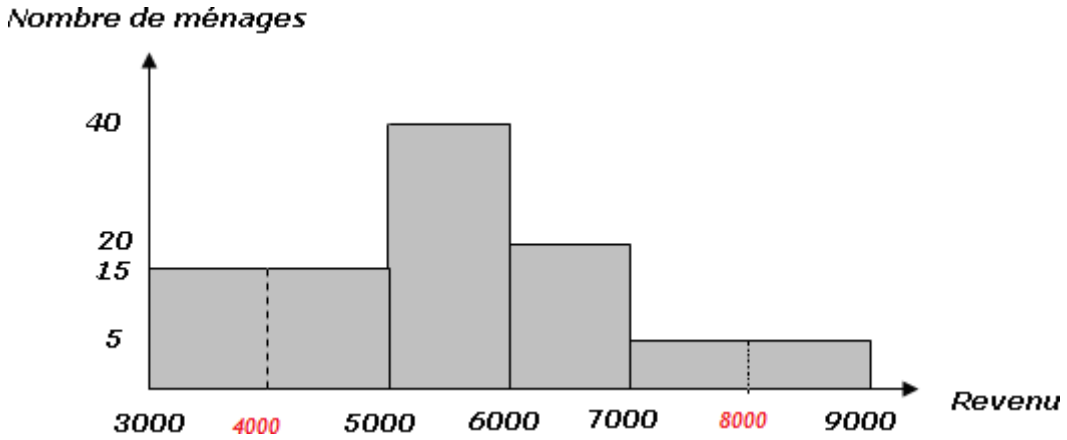
La probabilité que l'on « tire » un individu dont le revenu serait exactement de 5 320.26 euros est pratiquement nulle. Pour éviter ce problème, on a donc préféré considérer des intervalles de revenu, et non des revenus « ponctuels ». Ainsi on cherche la probabilité que le revenu de l'individu tiré soit compris dans l'intervalle [6000 ; 7000[, plutôt que celle qu'il prenne une valeur précise. Ces revenus sont ici regroupés dans des classes d'amplitude 1000 euros. Les classes étant ici inégales, le principe est le suivant : on prend une longueur de classe de référence, en général *la plus petite* (1000 dans le cas précédent) *ou la plus répandue*, et dans chaque classe dont la longueur est égale à k fois cette longueur de référence, on divise les effectifs par k[3].

---

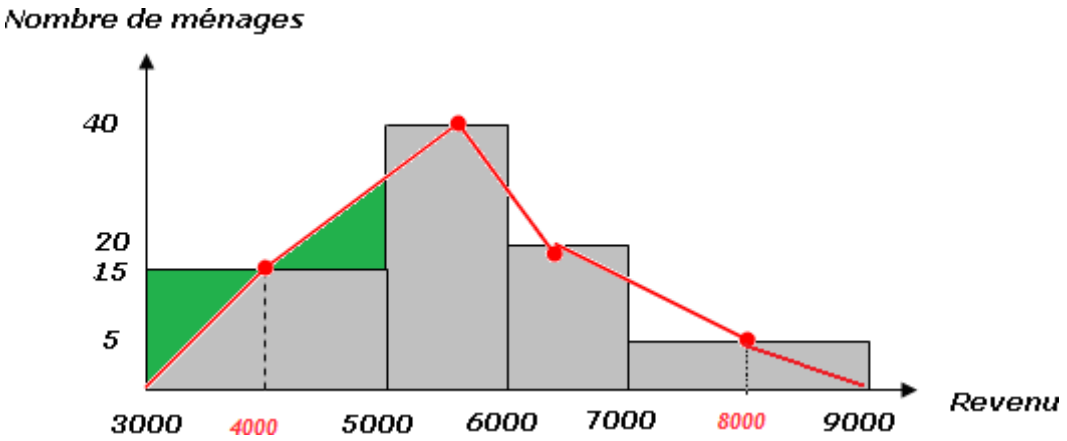
[2] On notera que si X est une variable aléatoire continue, alors $p(X = x_i) = 0$ (et ce, quel que soit x). Prenons l'exemple de paquets de farine dont le poids en farine est une variable aléatoire "Y" prenant ses valeurs dans l'intervalle [950 ; 1100]. Lorsque l'on prend un paquet au hasard, la probabilité d'obtenir un poids de farine rigoureusement égal à 978.25 g par exemple est nulle. En effet, c'est un paquet parmi une infinité, et 1/infini = 0 (nombre cas favorables/nombre cas possibles). Donc l'événement poids = 978.2 g est dit "quasi impossible".

[3] Comme la première et la dernière sont le double des autres, il est normal de diviser leurs effectifs par deux, de façon à rendre comparables les situations (ce qui revient, par exemple, à répartir de façon égale les 30 ménages de la première classe, allant de 3000 à 5000, en deux classes : 15 ménages dans la première, 3000-4000, 15 dans la seconde, 4000-5000 (la hauteur du tuyau correspondant égale à : 30/2 = 15). Même chose, pour la dernière classe.

Dans notre exemple, le plus petit intervalle étant de 1 000, il est pris comme unité. Un intervalle de 2 000 se trouve être égal à deux unités. La valeur du premier intervalle mesurée avec la nouvelle unité est donc égale à 2 ([3 000 ; 5 000[). La surface du rectangle correspondant est bien égale 2 × 15 = 30 soit l'effectif de la classe. D'où la représentation :



Enfin après rectification, on constate que la surface de l'histogramme est bien égale à l'effectif de 100 (2×15 + 1×40 + 1×20 + 2×5). Si nous raisonnions en fréquences, la surface de l'histogramme serait égale à la somme des fréquences, donc à l'unité.



Par ailleurs, le caractère continu de la variable peut être matérialisé par la ligne polygonale rouge des effectifs (ou des fréquences) qui joint les sommets des rectangles au niveau des **centres de classe**. Par construction, l'aire sous le **polygone des effectifs** (ou des fréquences) correspond à l'aire totale de l'histogramme. Á chaque fois, on perd et on gagne simultanément deux triangles égaux, ce que montre clairement la figure précédente.

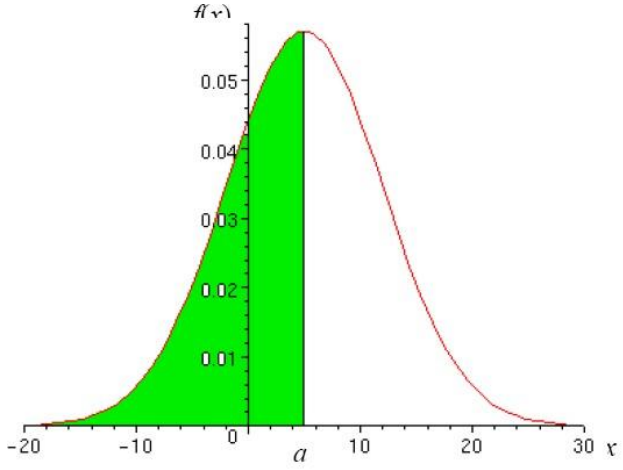
Si l'effectif de la population est important, lorsque l'amplitude de la classe tend vers zéro, c.-à-d. que le nombre de classes devient très important, la ligne

polygonale du dessus tend vers une courbe de plus en plus lisse appelée **courbe des effectifs** (ou courbe des fréquences). C'est la courbe d'une fonction continue f représentative de la distribution de fréquences. Cette fonction *f* prend le nom de **densité de probabilité**. Dans le cas d'une variable continu, la fonction de distribution ou loi de probabilité s'appelle donc densité de probabilité *f(x)*. Par analogie avec l'histogramme, l'aire qui est sous la courbe de densité de probabilité est égale à 1 (somme des probabilités).

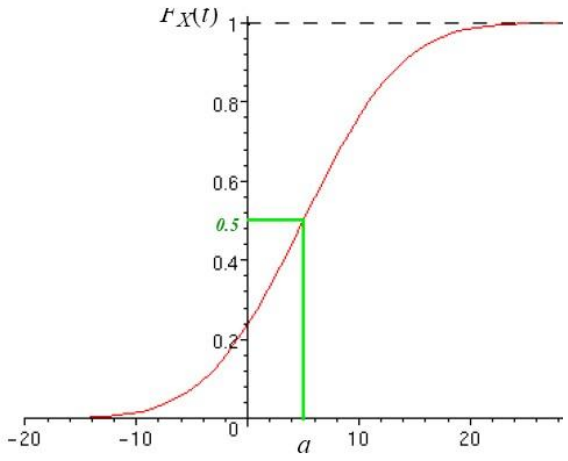
Enfin la **fonction de répartition** correspondrait aux probabilités cumulées associées à la variable aléatoire continue. La probabilité P (X ≤ x_i) est représentée par l'aire située sous la courbe à gauche de l'abscisse x_i (centre de classe)[4].

$$P(X < x_i) = \int_{-\infty}^{x_i} f(x)dx$$

Par exemple dans le schéma suivant, l'aire hachurée en vert sous la courbe de la fonction densité de probabilité correspond à la probabilité P (X < a) et vaut 0.5 car ceci correspond à la moitié de l'aire totale sous la courbe. Cette probabilité correspond à la valeur de la fonction de répartition au point d'inflexion de la courbe.



Fonction densité de probabilité *f(x)*



Fonction de répartition F_X

b) Espérance mathématique, moyenne et variance de variables discrètes

[4] Puisque dans le cas d'une variable aléatoire continue, la densité de probabilité en un point est nulle (c.-à-d. que l'intervalle, base du rectangle, est sans dimension, l'aire est donc nulle), on a alors P (X ≤ x_i) = P (X < x_i).

64ACG610 – Les outils statistiques de gestion

Soit X une variable aléatoire qui possède une loi de probabilité. **L'espérance mathématique** est la somme des valeurs prises par X, pondérée par les probabilités qui leur sont associées :

$$\mathbf{E(X) = x_1\, p(x_1) + x_2\, p(x_2) + \ldots + x_n\, p(x_n) = \sum_{i=1}^{n} x_i\, p(x_i)}$$

Ex : Pour poursuivre l'exemple de l'introduction, l'espérance mathématique de gain est,

$$E(X) = (-1)(1/6) + (1.50)(4/6) + (-6)(1/6) = -1/6 \text{ euros}$$

On peut donc s'attendre à perdre en moyenne 1/6€ par coup, ce qui signifie que le jeu est défavorable au joueur.

La variance est définie comme la différence entre la moyenne (l'espérance) des carrés et le carré de la moyenne. Ainsi :

$$\mathbf{V(X) = E(X^2) - [E(X)]^2}$$

ou

$$V(X) = \sum_{i=1}^{n} x_i^2\, p(x_i) - \left[\sum_{i=1}^{n} x_i\, p(x_i)\right]^2$$

Ex : Pour la variance, nous allons calculer la moyenne des carrés moins le carré de la moyenne. Calculons d'abord la moyenne des carrés :

$$E(X^2) = 6^2 \times 1/6 + -1^2 \times 1/6 + 1.50^2 \times 4/6 = 7.67$$

Variance et écart-type :

$$V(X) = 7.67 - (-1/6^2) = 7.68$$

L'**écart-type** de X, que l'on écrit σ, est la racine carrée (positive ou nulle) de V(X) :

$$\sigma_X = \sqrt{Var(X)}$$

Ex : D'après l'exemple précédent, $\sigma_X = \sqrt{7.68} = 2.77€$

Notons que si l'espérance indique la valeur moyenne espérée d'une série de valeurs, **l'écart-type permet de mesurer la dispersion des valeurs autour de cette moyenne** espérée. S'il y a une grande dispersion (c.-à-d. si l'écart entre les valeurs et la moyenne est important), alors l'écart-type est grand ; si au contraire la dispersion est faible, alors l'écart-type est petit.

Attention aussi aux confusions, courantes chez les étudiants. La première concerne la moyenne, le plus souvent notée $\bar{x}$, et l'espérance mathématique notée E(X). **La moyenne est empirique, l'espérance est théorique**.

Ex : Imaginons un jeu de pile ou face où vous gagnez 1 euro s'il y a face et 0 euros s'il y a pile. La loi de probabilité est donc la suivante :

x_i	1	0	
p_i	1/2	1/2	1

Quel est mon espérance de gain ?

$$E(X) = \sum_{i=1}^{2} x_i\, p(x_i) = (1)(1/2) + (0)(1/2) = 0.5€$$

C'est ce que vous pouvez espérer gagner **avant** de jouer. L'espérance de gain ne dépend pas de votre expérience : elle vaut toujours 0.50 euros. Mais supposons que vous jouez effectivement en faisant 4 lancers, et que vous avez 1 fois face, alors votre gain moyen (ou moyenne des gains) est de 0.25 euros.

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i}{N} = \frac{1+0+0+0}{4} = 0.25€$$

C'est surtout la formule précédente, celle avec $\bar{x}$, qui sera utilisée en inférence statistique. Rappelons en effet que l'inférence statistique est un procédé de généralisation des résultats échantillonnaux à l'ensemble de la population. Ainsi un fabricant d'ampoules électriques qui veut estimer la durée de vie moyenne des ampoules qu'il produit prélèvera un échantillon de sa production et estimera la vie moyenne de l'ensemble des ampoules d'après la durée moyenne observée dans l'échantillon. **On fait de l'inférence statistique lorsqu'on estime la moyenne d'une population à partir de la moyenne d'un échantillon tiré de la population**.

La seconde confusion concerne les indicateurs de dispersion. Lorsqu'on se projette dans le futur, qu'on raisonne en termes probabilistes, la variance se calcule de la façon suivante :

$$V(X) = E(X^2) - [E(X)]^2$$

Lorsqu'on raisonne sur des données passées, on sollicite les formules suivantes :

$$\sigma^2 = \frac{\sum x^2}{N} - \left(\bar{x}\right)^2 \text{ ou } \sigma^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{N}$$

Ex : Calculons l'écart type de la série observée : 4, 7, 11, 12. La moyenne vaut :

$$\bar{x} = \frac{4 + 7 + 11 + 12}{4} = 8.5$$

Soit le tableau suivant :

x_i	4	7	11	12	**Total**
$(x_i)^2$	16	49	121	144	330

$$V(X) = \frac{330}{4} - (8.5)^2 = 10.25 \Rightarrow \sigma(X) = \sqrt{10.25} = 3.2$$

64ACG610 – Les outils statistiques de gestion

I. Notions d'analyse combinatoire

a) Les permutations et les arrangements

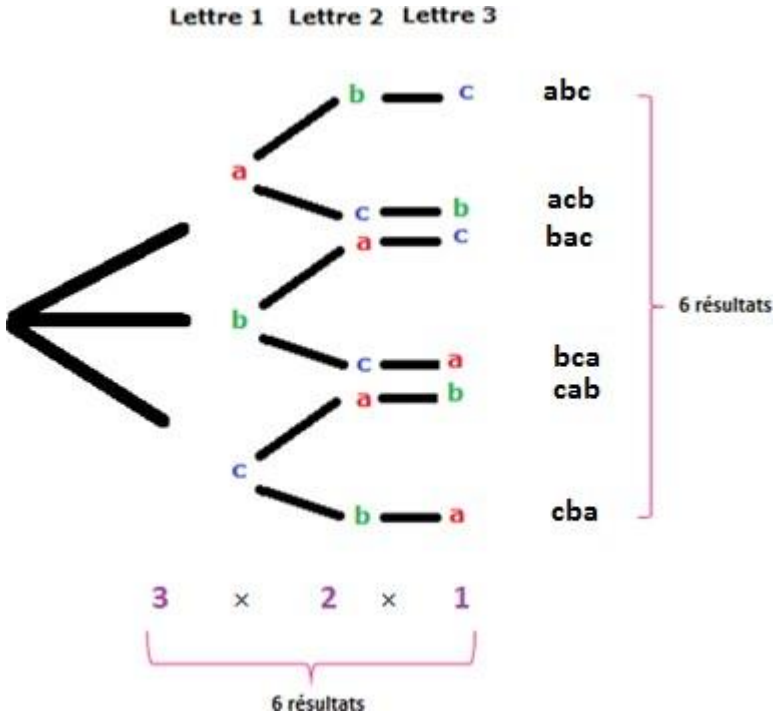
Disposant d'un ensemble de *n* objets *distincts*, une **permutation** est un rangement **ordonné sans répétition** de ces *n* objets.

Le nombre total de permutations distinctes, noté n!, est égal à :

$$P_n = n! = n\,(n-1)\,(n-2)\,(n-3) \times \ldots \times 2 \times 1$$

Par convention, 0 ! = 1.

Ex : Le nombre de permutations de 3 éléments d'un ensemble E = {a,b,c} est égal à 3 ! = 3 × 2 × 1 = 6 comme le montre l'arbre ci-dessous.



Dans l'exemple, l'ensemble des permutions à trois éléments est {(a,b,c);(b,a,c);(b,c,a);(a,c,b);(c,a,b);(c,b,a)}.

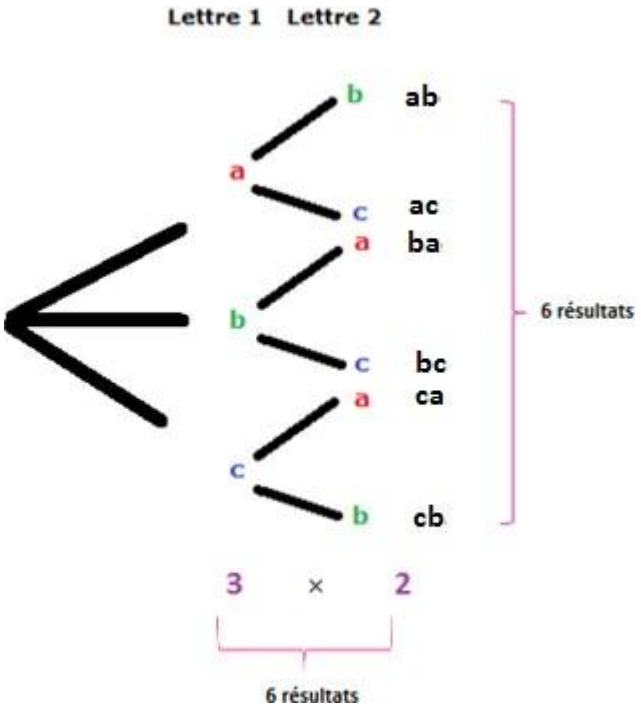
Dans un ensemble de *n* objets *distincts*, un **arrangement** est une disposition ordonnée de *r* objets choisis parmi ces *n* objets.

La formule pour le nombre de permutations (sans répétition) de *n* objets pris *r* à la fois est la suivante :

64ACG610 – Les outils statistiques de gestion

$$A_n^r = \frac{n!}{(n-r)!}$$

Ex : La question « Combien d'arrangements de deux lettres peut-on obtenir avec les lettres a, b, c ? » implique que *ab* est différent de *ba*. Les deux mêmes lettres sont présentes, mais l'ordre d'apparition est différent.



Dans l'exemple, le nombre d'arrangements serait ab, ac, ba, bc, ca, cb. Il y a donc 3 × 2 arrangements (trois choix pour la première lettre et deux choix pour la seconde). Avec la formule :

$$A_3^2 = \frac{3!}{(3-2)!} = 3 \times 2 = 6$$

Pour les arrangements, la formule avec remise est : **Nombre d'arrangements = n^r**, *n* est le nombre total d'éléments et *r* est le nombre d'éléments sélectionnés.

Ex : On choisit au hasard et avec remise deux lettres parmi l'ensemble E = {a,b,c} et on s'intéresse au nombre d'arrangements possibles.

Nombre d'arrangements = 3 × 3 = 3^2

On reprend les six résultats précédents et on ajoute aa, bb et cc.

b) Les combinaisons

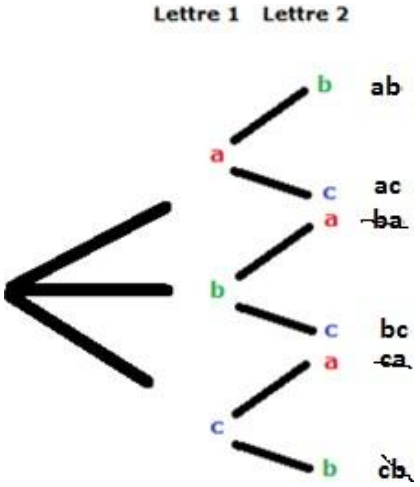
Dans un ensemble de *n* objets *distincts*, une **combinaison** est une disposition non ordonnée de *r* objets choisis parmi les *n*. Contrairement à l'arrangement, l'ordre des objets au sein d'un groupement n'est pas retenu dans les combinaisons. Tout comme dans l'arrangement, les répétitions ne sont pas permises. Une combinaison sera donc le résultat d'un **tirage sans ordre et sans répétition** (on dit aussi **sans remise**).

La formule générale donnant le nombre de combinaisons de *n* objets pris *r* à la fois C_n^r est donc donnée par la formule suivante :

$$C_n^r = \frac{n!}{r!(n-r)!}$$

Le nombre de combinaisons de *r* individus parmi *n* est égal au nombre d'arrangements de *r* individus parmi *n* divisé par *r!*.

Ex : La question « Combien de combinaisons de deux lettres peut-on obtenir avec les lettres a, b, c ? » implique que *ab* est identique à *ba*. Autrement dit, ce sont deux arrangements (ou permutations) mais une même combinaison puisqu'on ne tient pas compte de l'ordre.



Dans l'exemple, le nombre de combinaisons serait ab, ac, et bc.

$$C_3^2 = \frac{3!}{2!(3-2)!} = \frac{6}{2} = 3$$

On sait que a, b donnent deux permutations (ab, ba) mais une combinaison ab. Ainsi deux lettres dans l'ensemble {a, b, c} donnent deux permutations, mais

seulement une combinaison. Ici il y a au total 6 permutations, donc pour trouver le nombre de combinaisons, nous devons diviser A_3^2 par P_2 (c.-à-d. 2 !).

64ACG610 – Les outils statistiques de gestion

2) La loi normale

a) Principes généraux

L'étude de certains phénomènes fait apparaître une « permanence statistique ». C**ette reproductibilité suggère l'existence de « lois statistiques »**, c.-à-d. de modèles théoriques que suivent de façon plus ou moins proche les phénomènes étudiés.
Ces lois qui font appel à la notion de probabilité sont nombreuses ; cependant, un pe**tit nombre d'entre elles permet de rendre compte de la majeure partie des phénomènes statistiques.**

C'est le cas en particulier de la loi normale (ou loi de Laplace-Gauss).

La loi normale est très intéressante parce que la très grande majorité des phénomènes naturels tendent vers cette distribution quand on prend un grand nombre de mesures.

Les contextes d'utilisation sont très nombreux, par exemple :

- La distribution de la durée de vie des ampoules électriques,
- Les distributions du poids des baleines
- Les notes d'un examen, La température

La plupart des caractéristiques physiques mesurables respectent la loi normale. **On observe la distribution normale partout dans la nature**.

De façon générale, lorsque l'on a affaire à une variable aléatoire continue dont la valeur dépend d'un grand nombre de causes indépendantes dont les effets s'additionnent et dont aucune n'est prépondérante (conditions de Borel), on montre que la distribution des valeurs de cette variable aléatoire suit une loi de Laplace-Gauss.

Ainsi la taille corporelle d'un animal dépend des facteurs environnementaux (disponibilité pour la nourriture, climat, prédation, etc.) et génétiques. Dans la mesure où ces facteurs sont indépendants et qu'aucun n'est prépondérant, on peut supposer que la taille corporelle suit une loi normale.

La loi normale est entièrement déterminée par deux paramètres :

- sa moyenne "m"
- et son écart-type "σ".

64ACG610 – Les outils statistiques de gestion

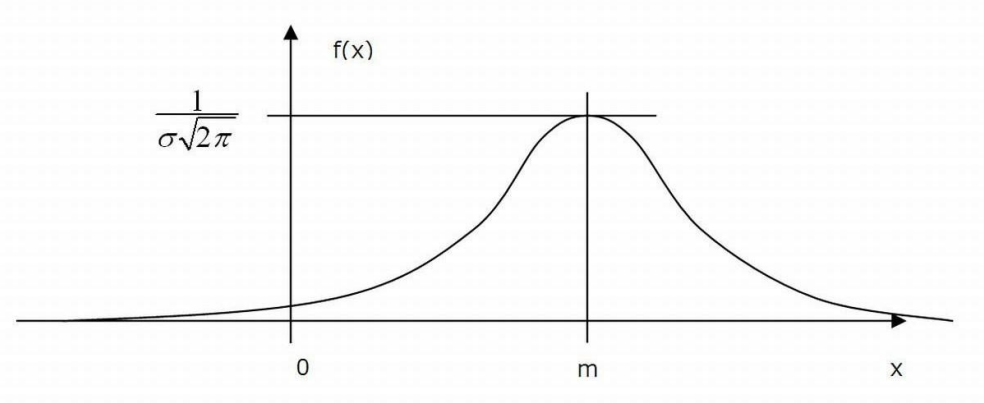
Si une variable aléatoire continue "X" obéit à une loi normale, alors "X" admet la ***fonction de densité*** suivante :

$$f(x)=\frac{1}{\sigma.\sqrt{2\,.\,\pi}}\;e^{-\frac{1}{2}\left[\frac{x-m}{\sigma}\right]^2}$$

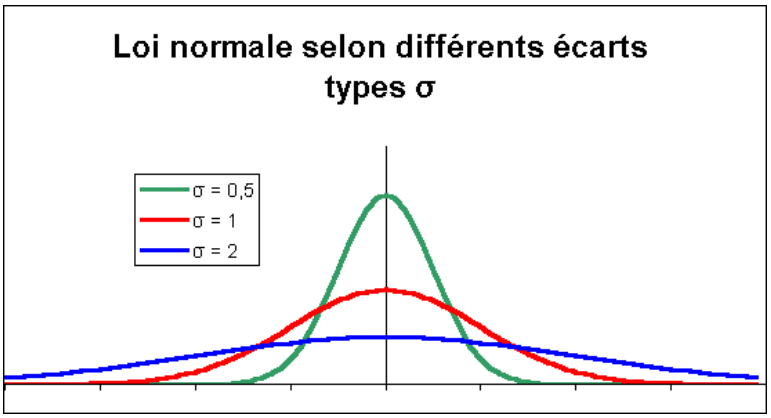
Avec, $-\infty < x < +\infty$ et $\sigma > 0$

$$\text{Et : } \int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{+\infty} \frac{1}{\sigma.\sqrt{2\,.\,\pi}}\;e^{-\frac{1}{2}\left[\frac{x-m}{\sigma}\right]^2} dx = 1$$

Ainsi l'expression, f(x) est la courbe de fréquence ou **densité de probabilité**[5], et la représentation graphique de cette fonction a la forme d'une « courbe en cloche » comme le montre le graphique suivant.



Dans une distribution normale, **la moyenne arithmétique, la médiane et le mode sont donc confondus**[6]. On a donc une courbe symétrique, aplatie aux deux extrémités et telle que la fréquence des données décroît à mesure qu'on s'éloigne du centre, que ce soit vers la gauche ou vers la droite. C'est là l'idée que l'on a généralement de la normalité : beaucoup de données se situent autour de la moyenne et le nombre de données diminue à mesure qu'on s'éloigne de celle-ci.



Bien que toutes les courbes soient symétriques, leur évasement correspond à l'écart-type. Plus il est élevé, plus les mesures sont dispersées autour de la valeur moyenne et plus la courbe s'aplatit. Plus il est faible (les valeurs sont rassemblées autour du centre), plus la courbe est pointue en sa moyenne.

---

[5] Rappelons qu'une fonction *f* est une densité de probabilité sur un intervalle si elle est continue (on trace la fonction d'un seul trait), positive (sa courbe est au-dessus de l'axe des abscisses), et si l'aire sous la courbe est égale à 1.

[6] Dans une distribution, le **mode** (Mo) correspond à la valeur ou la catégorie qui possède l'effectif le plus élevé. La **médiane** (Md) correspond à la valeur qui se trouve au centre d'une distribution de données qui sont ordonnées en ordre croissant ou décroissant.

64ACG610 – Les outils statistiques de gestion

La distribution normale tient compte de deux paramètres la moyenne *m* et l'écart-type *σ*. Et puisqu'il existe un nombre infini de combinaisons moyenne et écart-type, il existe un nombre infini de lois normales.

Comme pour toute variable aléatoire continue, c'est la fonction de répartition F (x) qui sert à calculer les probabilités. Ainsi la probabilité que la variable aléatoire X ait une valeur inférieure à "x" est donnée par la ***fonction de répartition*** suivante :

$$F(x) = p(X < x) = \int_{-\infty}^{x} \frac{1}{\sigma . \sqrt{2 . \pi}} e^{-\frac{1}{2}\left[\frac{t-m}{\sigma}\right]^2} dx$$

Mais compte tenu de la complexité de la formule ci-dessus, le calcul des valeurs à partir de la fonction de répartition de la variable normale X qui suit N (m, σ) est quasi impossible. C'est la raison pour laquelle on effectue un changement de variable pour passer à une loi normale particulière appelée **loi normale centrée réduite**.

Soit "T" la variable aléatoire obtenue à partir de X :

$$T = \frac{X - m}{\sigma}$$

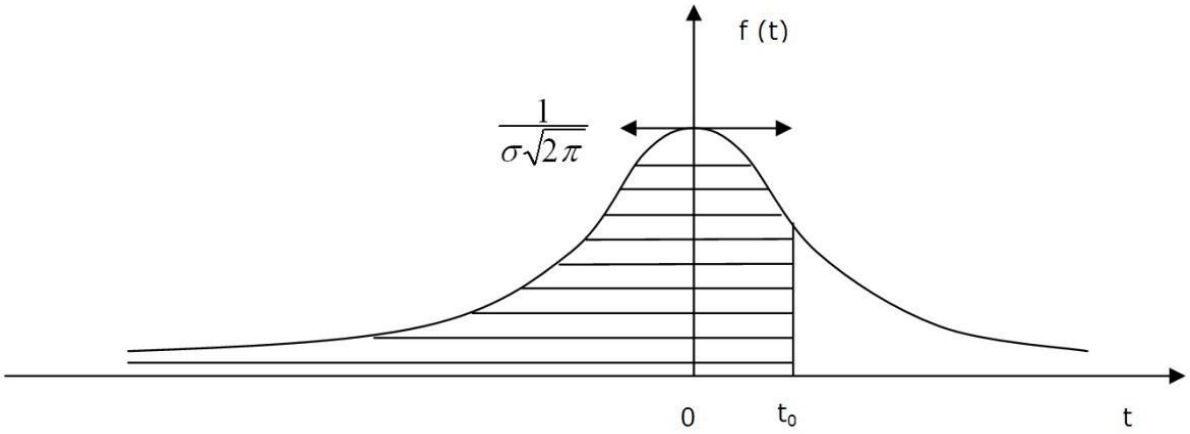
L'espérance et l'écart-type d'une variable suivant une loi normale centrée réduite sont :

- E(T) = 0 ;
- σ (T) = 1.

La loi de T, notée N (0 ; 1) admet une fonction de densité définie par :

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

La représentation graphique de cette fonction a la forme de la courbe ci-dessous, dite fonction de densité de la loi normale centrée réduite.



Evidemment puisque la moyenne est nulle, il en est de même pour le mode et la médiane.

Enfin, la surface sous une courbe de distribution normale représente la probabilité qu'un phénomène se produise dans cet intervalle. Ainsi la probabilité qu'une variable prenne une valeur inférieure à t_0 est égale à l'aire sous la courbe entre $-\infty$ et la valeur t_0. Ainsi **l'aire totale comprise entre la courbe de densité de probabilité et l'axe des abscisses est égale à 1**. La fonction de répartition est notée :

$$\Pi(t) = F(T) = P(T \leq t) = \int_{-\infty}^{t} f(t)dt = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

Nous verrons que $\Pi(t) = F(T) = P(T \leq t)$ indique que cette fonction est tabulée pour des valeurs positives de *t*.

a) Calculs de probabilités avec la loi normale

Comment calculer la probabilité pour que X soit < à 25 000 ?

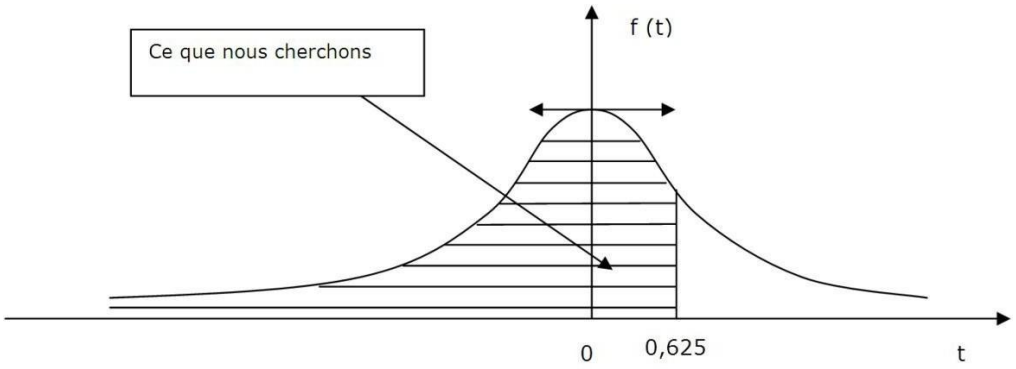
On cherche p(X < 25 000) sachant que la variable aléatoire continue "X" suit une loi normale de paramètres "m"= 20 000 et "σ"= 8 000. On ne peut pas répondre à cette question directement puisqu'il faudrait utiliser pour cela la formule suivante :

$$F(x) = p(X < x) = \int_{-\infty}^{x} \frac{1}{\sigma . \sqrt{2} . \pi} e^{-\frac{1}{2}\left[\frac{t-m}{\sigma}\right]^2} dt$$

Or grâce au changement de variable de Laplace-Gauss, nous allons pouvoir *centrer réduire* la variable X et la remplacer par "T". Il vient :

$$p(X < a) = p\left(T < \frac{a - m}{\sigma}\right) \Rightarrow p(X < 25\,000) = p\left(T < \frac{25\,000 - 20\,000}{8\,000}\right) = p(T < 0{,}625)$$

Ce que nous recherchons graphiquement :



Pour trouver cette probabilité nous allons donc utiliser une des tables de loi normale. Il existe en effet une table de la loi normale centrée réduite P (T < t) qui nous donne ce que nous recherchons, à savoir l'aire grisée de la courbe précédente. Comme son nom l'indique cette table donne les probabilités : p(T < t) = Π(t).

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500 0	0,504 0	0,508 0	0,512 0	0,516 0	0,519 9	0,523 9	0,527 9	0,531 9	0,535 9
0,1	0,539 8	0,543 8	0,547 8	0,551 7	0,555 7	0,559 6	0,563 6	0,567 5	0,571 4	0,575 3
0,2	0,579 3	0,583 2	0,587 1	0,591 0	0,594 8	0,598 7	0,602 6	0,606 4	0,610 3	0,614 1
0,3	0,617 9	0,621 7	0,625 5	0,629 3	0,633 1	0,636 8	0,640 6	0,644 3	0,648 0	0,651 7
0,4	0,655 4	0,659 1	0,662 8	0,666 4	0,670 0	0,673 6	0,677 2	0,680 8	0,684 4	0,687 9
0.5	0,691 5	0,695 0	0,698 5	0,701 9	0,705 4	0,708 8	0,712 3	0,715 7	0,719 0	0,722 4
0,6	0,725 7	0,729 1	0,732 4	0,735 7	0,738 9	0,742 2	0,745 4	0,748 6	0,751 7	0,754 9
0,7	0,758 0	0,761 1	0,764 2	0,767 3	0,770 3	0,773 4	0,776 4	0,779 3	0,782 3	0,785 2

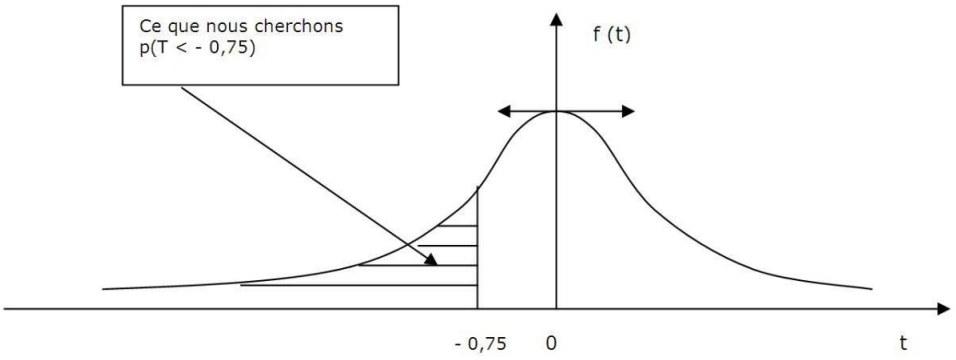
Dans ce cas, Π(0.62) = P (T < 0.62) = 0.7324.

Comment calculer la probabilité pour que X soit < à 14 000 ?

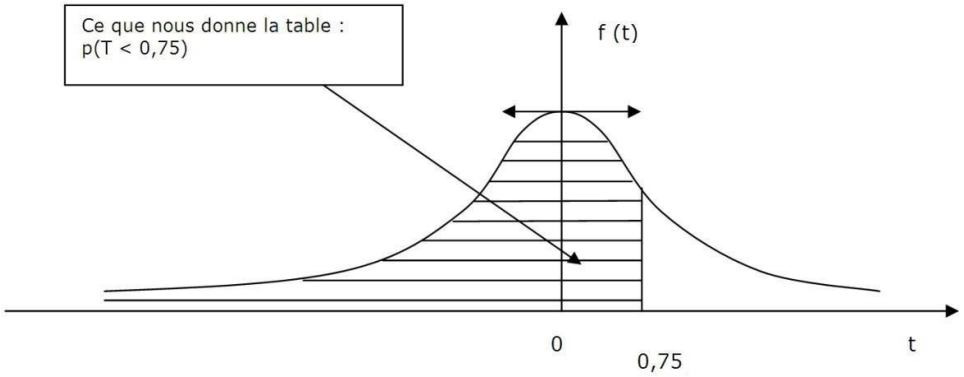
On cherche p(X < 14 000) sachant que la variable aléatoire continue "X" suit une loi normale de paramètres "m" = 20 000 et "σ" = 8 000. Grâce au changement de variable, nous allons pouvoir *centrer réduire* la variable X et la remplacer par "T". Il vient :

$$p(X < 14\,000) = p\left(T < \frac{14\,000 - 20\,000}{8\,000}\right) = p(T < -0{,}75)$$

La courbe suivante nous donne schématiquement ce que nous recherchons :



La table p (T< t) ne donne les probabilités que pour t > 0 :



On note toutefois que si "t" est < 0 $\Rightarrow \Pi(-t) = p(T < -t) = 1 - p(T < t)$.

Compte tenu de la symétrie (par rapport à t = 0) de la table centrée réduite on voit bien en comparant les deux schémas que :

$$p(T < -0.75) = 1 - p(T < 0.75)$$

Donc en appliquant ce principe, il vient :

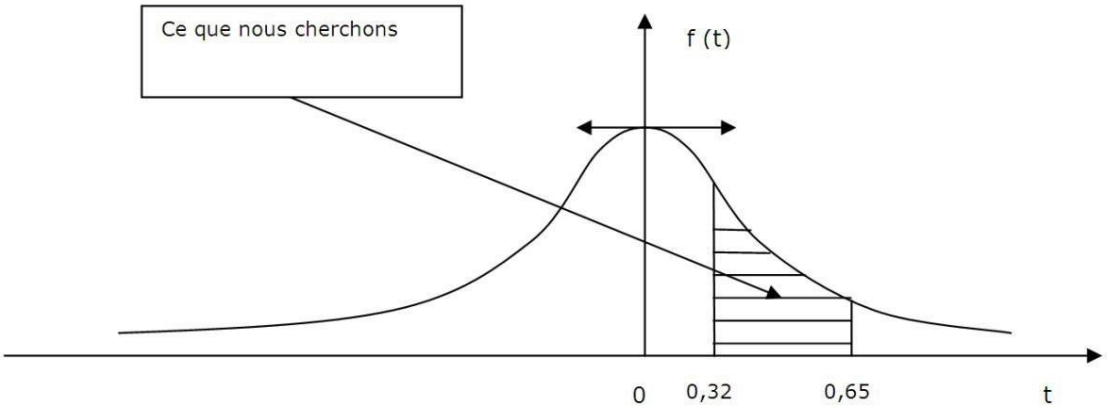
$$p(T < -0.75) = p(T > 0.75) = 1 - p(T < 0.75) = 1 - 0.7734 = 0.2266$$

Comment calculer la probabilité pour que X soit comprise entre 22 560 et 25 200 ?

Cette fois on cherche p(22 560 < X < 25 200) sachant que la variable aléatoire continue "X" suit une loi normale de paramètres "m" = 20 000 et "σ" = 8 000. Grâce au changement de variable, nous allons pouvoir centrer réduire la variable X et la remplacer par "T".

$$\Rightarrow p(22\,560 < X < 25\,200) = p\left[\left(\frac{22\,560 - 20\,000}{8\,000}\right) < T < \left(\frac{25\,200 - 20\,000}{8\,000}\right)\right]$$

$$\Rightarrow p(22\,560 < X < 25\,200) = p(0{,}32 < T < 0{,}65)$$



En regardant le schéma, on voit bien que :

p(0.32 < T < 0.65) = p (T < 0.65) – p (T < 0.32). Or ici 0.32 et 0.65 sont des valeurs positives. Nous pouvons donc trouver le résultat directement dans la table ∏ (t) = p (T < t). Il vient :

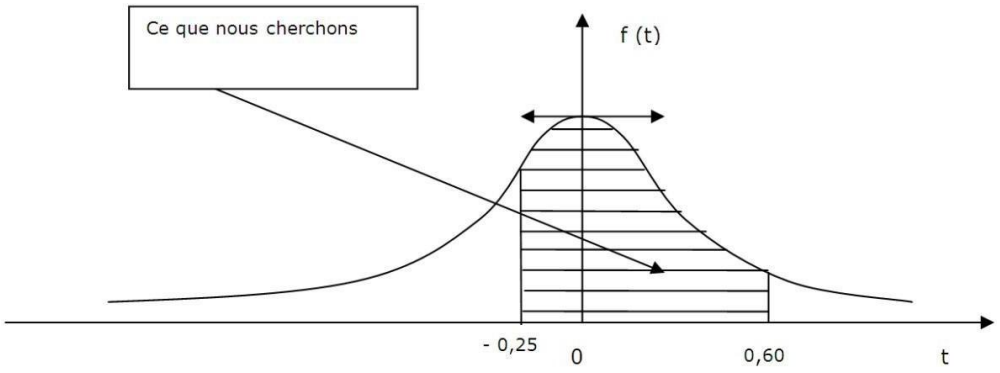
$$p\,(T < 0.65) - p\,(T < 0.32) = 0.7422 - 0.6255 = 0.1167 = 11.67\%$$

Comment calculer la probabilité pour que X soit compris entre 18 000 et 24 800 ?

Cette fois on cherche p(18 000 < X < 24 800) sachant que la variable aléatoire continue "X" suit une loi normale de paramètres "m" = 20 000 et "σ" = 8 000. Grâce au changement de variable, nous allons pouvoir *centrer réduire* la variable X et la remplacer par "T".

$$\Rightarrow p(18\,000 < X < 24\,800) = p\left[\left(\frac{18\,000 - 20\,000}{8\,000}\right) < T < \left(\frac{24\,800 - 20\,000}{8\,000}\right)\right]$$

$$\Rightarrow p\,(-0{,}25 < T < 0{,}60)$$



En regardant le schéma, on voit bien que :

$$p(-0.25 < T < 0.60) = p(T < 0.60) - p(T < -0.25)$$

Or ceci revient à dire : $p(-0.25 < T < 0.60) = p(T < 0.60) - [1 - p(T < 0.25)]$.
Donc en lisant sur la table, il vient :

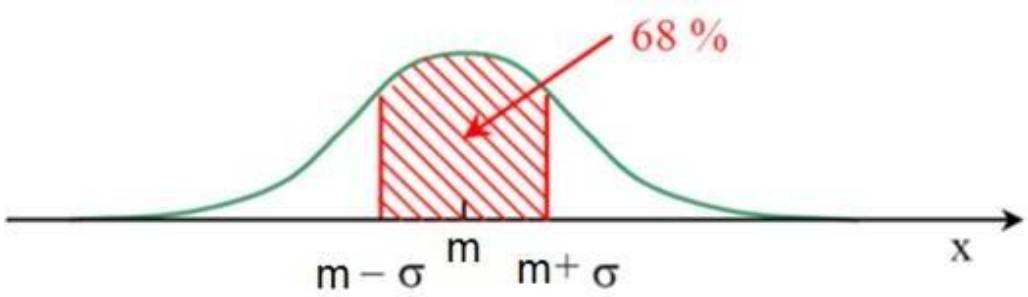
$$p(T < 0.60) - [1 - p(T < 0.25)] = 0.7257 - (1 - 0.5987) = 0.3244 = 32.44\%$$

b) Propriétés fondamentales de la loi normale

Une variable aléatoire X normale de moyenne m et d'écart type σ prend une valeur à moins d'un écart type de sa moyenne environ 68% des fois. Mathématiquement :

$$P(m - \sigma \leq X \leq m + \sigma) = 0.6826$$

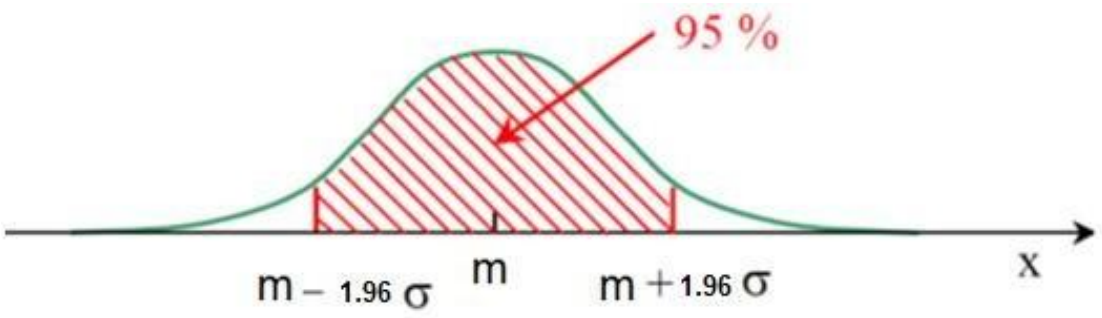
Graphiquement :



Une variable aléatoire X normale de moyenne m et d'écart type σ prend une valeur à moins de 1.96 écarts type de sa moyenne environ 95% des fois. Mathématiquement :

$$P(m - 1.96\sigma \leq X \leq m + 1.96\sigma) = 0.954$$

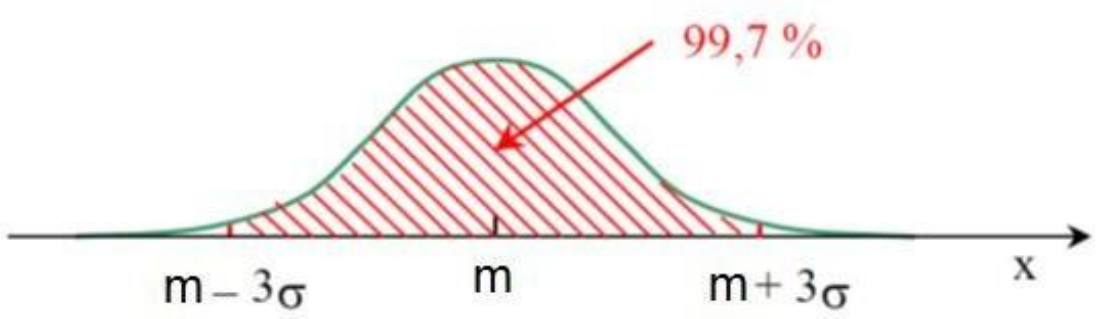
Graphiquement :



Une variable aléatoire normale de moyenne *m* et d'écart type *σ* prend une valeur à moins de trois écarts types de sa moyenne plus de 99% des fois.

$$P(m - 3\sigma \leq X \leq m + 3\sigma) = 0.997$$

Graphiquement :



Ex : Pour le Q.I. avec m = 100 et σ = 15, à peu près 68% des valeurs de la population se retrouvent entre 85 et 115, environ 95% des valeurs de la population se retrouvent entre 70 et 130, presque 100% des valeurs de la population se retrouvent entre 55 et 145.

Si X est une variable aléatoire normale, alors on sait que (X – m)/σ est la variable aléatoire centrée réduite de moyenne 0 et d'écart-type 1. On aura alors les égalités suivantes :

$$P(-1 \leq T \leq +1) = 0.6826$$
$$P(-1.96 \leq T \leq +1.96) = 0.95$$
$$P(-3 \leq T \leq +3) = 0.9974$$

Ex : Calculons $P(-1 \leq T \leq +1)$

$$P(-1 \leq T \leq +1) = p(T < 1) - [1 - p(T < 1)]$$
$$= 2\,p(T < 1) - 1$$
$$= 2\,\Pi(1) - 1$$

Sur la table on peut lire $\Pi(1) = 0.84134$. D'où :

$$P(-1 \leq T \leq +1) = (2 \times 0.84134) - 1 = 0.6826$$

64ACG610 – Les outils statistiques de gestion