

B0654

Eléments de Statistique

Table des matières

1. Statistique univariée	4
1.1 Définitions	4
1.1.1 Population / Echantillon / Individu.....	4
1.1.2 Caractère.....	4
1.1.3 Série statistique	5
1.2 Effectifs	6
1.2.1 Effectif total	6
1.2.1 Effectif de classe	6
1.3 Les mesures de tendance centrale	8
1.3.1 La moyenne arithmétique.....	8
1.3.2 La médiane	9
1.4 Les mesures de dispersion	10
1.4.1 L'étendue	10
1.4.2 La variance	10
1.4.3 L'écart-type	11
1.5 Quartiles et Déciles	12
1.5.1 Quartiles.....	12
1.5.2 Déciles	13
1.6 Représentations graphiques usuelles	14
1.6.1 Caractère quantitatif.....	14
1.6.2 Caractère qualitatif	16
2. Statistique bivariée.....	19
2.1 Introduction	19
2.2 Ajustement statistique	19
2.2.1 Ajustement linéaire.....	20
2.2.2 Covariance.....	21
2.2.1 Corrélation.....	22
2.2.2 Corrélation et causalité.....	24
2.2.3 Ajustements non-linéaire	25
3. Eléments de probabilité	28
3.1 Définitions	28

3.1.1	Expérience aléatoire	28
3.1.2	Catégorie d'épreuve.....	28
3.1.3	Evénement.....	28
3.2	Diagramme de Venn	29
3.2.1	Introduction	29
3.2.2	Théorie des ensembles	29
3.2.3	Evénements contraires	31
3.2.4	Cardinal d'un événement	32
3.2.5	Fréquence et probabilité	32
3.2.6	Axiomes de Kolmogorov	33
3.2.7	Evénements équiprobables.....	36
3.2.8	Evénements indépendants.....	36
3.2.9	Probabilité conditionnelle	38
3.3	Diagramme en arbre	39
3.4	Analyse combinatoire	40
3.4.1	Factorielle d'un naturel	41
3.4.2	Permutations	41
3.4.3	Permutations avec répétitions	43
3.4.4	Arrangements avec répétitions	44
3.4.5	Arrangements sans répétition.....	45
3.4.6	Combinaisons sans répétition	46
4.	Distributions théoriques	47
4.1	Distribution binomiale	47
4.2	Distribution de Poisson	48
4.3	Distribution normale.....	49
4.4	Tables de probabilité.....	49

1. Statistique univariée

1.1 Définitions

1.1.1 Population / Echantillon / Individu

Un **individu** est un élément sur lequel portent les observations, les mesures ou l'étude statistique.

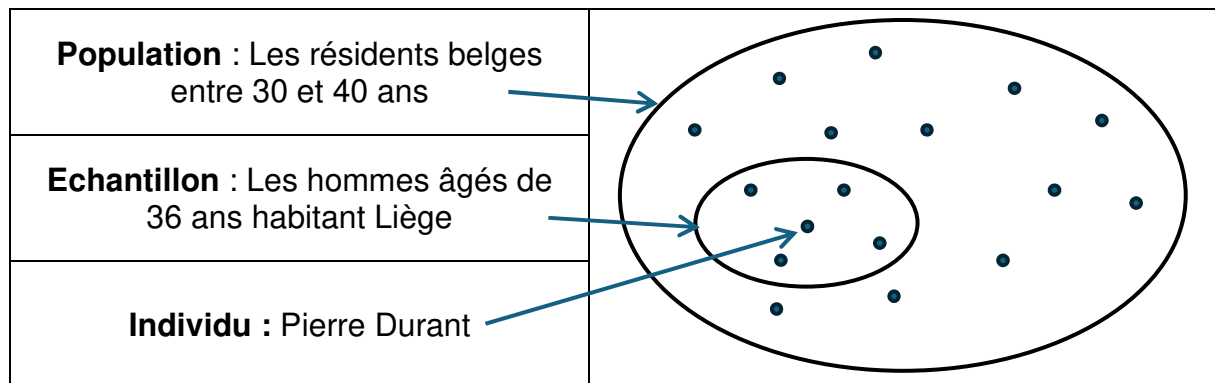
Ex : Pierre Durant, 36 ans, habitant Liège.

La **population** est l'ensemble des individus étudiés.

Ex : Les résidents belges entre 30 et 40 ans.

Un **échantillon** est un sous-ensemble de la population.

Ex : Les hommes âgés de 36 ans habitant Liège.



1.1.2 Caractère

Un **caractère** est une caractéristique étudiée sur les individus considérés

Ex : la taille, les sports pratiqués, la couleur des yeux, ...

Un caractère est **qualitatif** s'il n'est pas mesurable.

Ex : Les sports pratiqués, le prénom,

Un caractère est **quantitatif** s'il est mesurable.

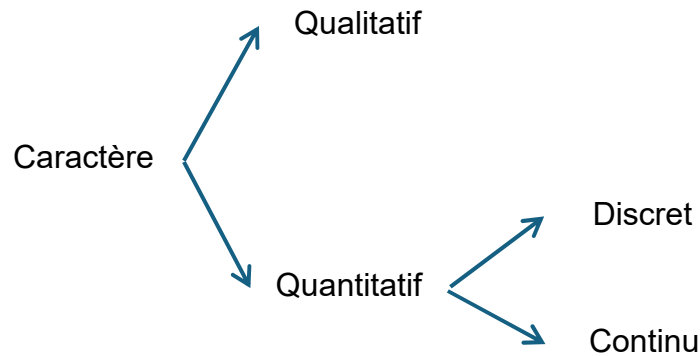
Ex : La taille, le nombre de sports pratiqués,

Un caractère quantitatif est **discret** ne peut prendre qu'un nombre fini de valeurs.

Ex : Le nombre de sports pratiqués,

Un caractère quantitatif est **continu** peut prendre toutes les valeurs dans un intervalle donné.

Ex : La taille,



1.1.3 Série statistique

Une série statistique est un ensemble de données recueillies portant sur un ou plusieurs caractères observés chez les individus d'une population ou d'un échantillon.

Elle représente la suite des valeurs prises par ce caractère.

Nom	Prénom	Nombre de sports pratiqués	Taille
Durant	Pierre	0	180,24 cm
Dupont	Adam	1	178,32 cm
Martin	Lucas	1	182,54 cm
Janssens	Pierre	3	177,98 cm
Alami	Mohamed	3	180,74 cm
Dupond	André	0	181,27 cm
Ilunga	Noah	1	181,84 cm

↑ Série statistique d'un caractère qualitatif	↑ Série statistique d'un caractère quantitatif discret	↑ Série statistique d'un caractère quantitatif continu
---	---	---

1.2 Effectifs

1.2.1 Effectif total

L'effectif total est le nombre d'individus présents dans une population, (l'effectif total est alors noté N) ou dans un échantillon (l'effectif total est alors noté n)

Mathématiquement, le comptage du nombre d'éléments d'un ensemble A se note $\#A$ (cardinal de A), On compte tous les individus même si le caractère se répète.

Considérons, la série statistique précédente et regardons le caractère prénom.

$\{Pierre, Adam, Lucas, Pierre, Mahamed, André, Noah\}$

$$n = \{Pierre, Adam, Lucas, Pierre, Mahamed, André, Noah\} = 7$$

1.2.1 Effectif de classe

Une classe est un sous-ensemble d'une population ou d'un échantillon regroupant des individus qui partagent des caractéristiques similaires. L'effectif de classe est le nombre d'individus dans la classe, il se note n_i ou i est le nom de la classe.

En règle générale :

Pour un caractère qualitatif et pour un caractère quantitatif discret, les classes correspondent aux valeurs possibles.

Si on reprend l'exemple précédent (qualitatif) :

$\{Pierre, Adam, Lucas, Pierre, Mahamed, André, Noah\}$

Les classes sont : $\{Pierre\}$; $\{Adam\}$; $\{Lucas\}$; $\{Mohamed\}$; $\{André\}$; $\{Noah\}$

Avec :

$$n_{Pierre} = 2 ; n_{Adam} = 1 ; n_{Lucas} = 1 ; \dots$$

Pour un caractère quantitatif discret, prenons le nombre de sports pratiqués :

$\{0; 1; 1; 3; 3; 0; 1\}$

Les classes sont : $\{0\}$; $\{1\}$; $\{2\}$; $\{3\}$

Avec :

$$n_0 = 2 ; n_1 = 3 ; n_2 = 0 ; n_3 = 2$$

Pour un caractère quantitatif continu, la définition des classes est plus complexe. En effet, si on reprend l'exemple de la taille, il y a peu de chance que deux individus aient exactement la même taille et donc l'effectif de classe serait presque toujours 1 ou 0.

Pour pallier ce problème, on va regrouper les individus ayant des tailles similaires. Ce regroupement se fait en identifiant la plus grande et la plus petite valeur du caractère étudié et en divisant l'intervalle entre ces deux extrêmes en une série d'intervalles.

Reprenons notre exemple et considérons 3 classes.

$\{180,24 ; 178,32 ; 182,54 ; 177,98 ; 180,74 ; 181,27 ; 181,84\}$

La taille minimum est : 177,98 cm

La taille maximum est : 182,54 cm

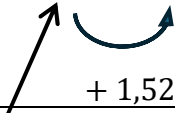
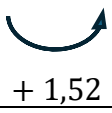
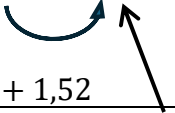
L'amplitude totale vaut : $182,54 - 177,98 = 4,56$ cm

L'amplitude de classe vaut : $\frac{4,56}{3} = 1,52$ cm

On divise par 3 car on considère 3 classes.

Remarque : La détermination du nombre de classes optimal sort du cadre de ce cours et le nombre de classes vous sera toujours donné.

Les classes sont définies par :

c_1	c_2	c_3
$[177,98 ; 179,5[$	$[179,5 ; 181,02[$	$[181,02 ; 182,54]$
		
minimum		maximum

Nous avons donc

$c_1 = \{177,98 ; 178,32\}$

$c_2 = \{180,24 ; 180,74\}$

$c_3 = \{182,54 ; 181,27 ; 181,84\}$

$n_1 = \#c_1 = \#\{177,98 ; 178,32\} = 2$

$n_2 = \#c_2 = \#\{180,24 ; 180,74\} = 2$

$n_3 = \#c_3 = \#\{182,54 ; 181,27 ; 181,84\} = 3$

Les intervalles de classes sont toujours de la forme

$[\dots ; \dots [$

Sauf la dernière qui est de la forme

$[\dots ; \dots]$

Le centre de classe est le centre de l'intervalle de la classe.

Par exemple, le centre de la classe c_1 vaut 178,74 $\left(= \frac{177,98+179,5}{2} \right)$

1.3 Les mesures de tendance centrale

Une tendance centrale n'a de sens que pour un caractère quantitatif.

Les mesures de tendance centrale vont nous donner une idée du centre de la série statistique, c'est-à-dire de la valeur autour de laquelle les données sont réparties.

1.3.1 La moyenne arithmétique

La moyenne arithmétique (m) d'une série statistique est le quotient de la somme des valeurs obtenues par l'effectif total.

$$m = \frac{(x_1 + x_2 + \dots + x_N)}{N} \text{ pour la population}$$

$$m = \frac{(x_1 + x_2 + \dots + x_n)}{n} \text{ pour un échantillon}$$

Avec : m la moyenne arithmétique.

N l'effectif total de la population.

n l'effectif total de l'échantillon.

x_i un caractère quantitatif.

A titre d'exemple, calculons la moyenne de la taille.

Taille = {180,24 ; 178,32 ; 182,54 ; 177,98 ; 180,74 ; 181,27 ; 181,84}

$n = \# \{180,24 ; 178,32 ; 182,54 ; 177,98 ; 180,74 ; 181,27 ; 181,84\} = 7$

$$m = \frac{(180,24+178,32+182,54+177,98+180,74+181,27+181,84)}{7} = 180,42 \text{ cm}$$

1.3.2 La médiane

La médiane (M) d'une série statistique est la valeur centrale qui partage la population en 2 groupes de même effectif : 50% en-dessous et 50% au-dessus.

Pour calculer une médiane, on doit classer la série statistique par ordre croissant et prendre la valeur du milieu du classement pour un effectif impair et la moyenne des deux termes centraux pour un effectif pair.

Reprenons notre exemple.

$Taille = \{180,24 ; 178,32 ; 182,54 ; 177,98 ; 180,74 ; 181,27 ; 181,84\}$

La première étape consiste à classer la série statistique par ordre croissant.

$Taille = \{177,98 ; 178,32 ; 180,24 ; 180,74 ; 181,27 ; 181,84 ; 182,54\}$

- Si l'effectif de la série est impair ($\#Taille = 7$ qui est impair)

La seconde étape consiste simplement à identifier l'élément central de la série.

$M = 180,74 \text{ cm}$

Et dans notre série statistique, il y a bien 3 tailles inférieures à 180,74 cm :

$\{177,98 ; 178,32 ; 180,24\}$

Et 3 tailles supérieures à 180,74 cm :

$\{181,27 ; 181,84 ; 182,54\}$

- Si l'effectif de la série est pair :

Supprimons un terme de notre série pour obtenir un effectif pair. Nous obtenons une nouvelle série, appelons la $Taille'$.

$Taille' = \{177,98 ; 178,32 ; 180,24 ; 180,74 ; 181,27 ; 181,84\}$

Dans ce cas, $\#Taille' = 6$ qui est pair. La détermination de la médiane se fait en calculant la moyenne des deux termes centraux.

$$M = \frac{180,24 + 180,74}{2} = 180,49 \text{ cm}$$

Et dans notre série statistique, il y a bien 3 tailles inférieures à 180,49 cm :

$\{177,98 ; 178,32 ; 180,24\}$

Et 3 tailles supérieures à 180,49 cm :

$\{180,74 ; 181,27 ; 181,84\}$

1.4 Les mesures de dispersion

Les mesures de dispersion vont nous permettre de savoir si les valeurs la série statistique s'écartent plus ou moins de la tendance centrale.

1.4.1 L'étendue

L'étendue d'une série statistique (E) est la différence entre la plus grande et la plus petite valeur de la série.

Reprenons notre exemple.

$Taille = \{180,24 ; 178,32 ; 182,54 ; 177,98 ; 180,74 ; 181,27 ; 181,84\}$

La plus petite taille vaut 177,98 cm et la plus grande 182,54 cm.

$$E = (182,54 - 177,98) = 4,56 \text{ cm}$$

1.4.2 La variance

La variance (V) est la moyenne des carrés des écarts entre chaque valeur de la série et la moyenne.

$$V = \frac{(x_1 - m)^2 + (x_2 - m)^2 + \dots + (x_N - m)^2}{N} \text{ pour la population}$$

$$V = \frac{(x_1 - m)^2 + (x_2 - m)^2 + \dots + (x_n - m)^2}{n} \text{ pour un échantillon}$$

Avec : V la variance.

m la moyenne arithmétique.

N l'effectif total de la population

n l'effectif total d'un échantillon

x_i un caractère quantitatif.

$Taille = \{180,24 ; 178,32 ; 182,54 ; 177,98 ; 180,74 ; 181,27 ; 181,84\}$

$$m = \frac{(180,24 + 178,32 + 182,54 + 177,98 + 180,74 + 181,27 + 181,84)}{7} = 180,42 \text{ cm}$$

$$V = \frac{(180,24 - 180,42)^2 + (178,32 - 180,42)^2 + (182,54 - 180,42)^2 + (177,98 - 180,42)^2 + (180,74 - 180,42)^2 + (181,27 - 180,42)^2 + (181,84 - 180,42)^2}{7}$$

$$V = 2,53 \text{ cm}^2$$

1.4.3 L'écart-type

L'écart-type (σ) est la racine carrée positive de la variance.

$$\sigma = \sqrt{V}$$

Avec : σ l'écart-type.

V la variance.

Et donc dans notre exemple :

$$\sigma = \sqrt{2,53} = 1,59 \text{ cm}$$

1.5 Quartiles et Déciles

Les quartiles et les déciles sont des extensions de la notion de médiane.

Pour rappel, la médiane (M) d'une série statistique est la valeur centrale qui partage la population en 2 groupes de même effectif : 50% en-dessous et 50% au-dessus.

1.5.1 Quartiles

Le premier quartile (Q_1) d'une série statistique est la valeur de la série qui partage la population en 2 groupes, le premier groupe contenant au moins 25% des effectifs : au moins 25% des effectifs sont inférieurs ou égaux au premier quartile.

Le second quartile (Q_2) d'une série statistique est simplement sa médiane.

Le troisième quartile (Q_3) d'une série statistique est la valeur de la série qui partage la population en 2 groupes, le premier groupe contenant au moins 75% des effectifs : au moins 75% des effectifs sont inférieurs ou égaux au troisième quartile.

Reprenons notre exemple.

$Taille = \{180,24 ; 178,32 ; 182,54 ; 177,98 ; 180,74 ; 181,27 ; 181,84\}$

Et comme pour la médiane, la première étape consiste à classer la série statistique par ordre croissant.

$Taille = \{177,98 ; 178,32 ; 180,24 ; 180,74 ; 181,27 ; 181,84 ; 182,54\}$

$\#Taille = 7$

$25\% * 7 = 1,75$

On prend, le premier entier supérieur à 1,75, soit 2, le premier quartile est le deuxième terme de la série triée.

$Q_1 = 178,32 \text{ cm}$

$75\% * 7 = 5,25$

On prend, le premier entier supérieur à 5,25, soit 6, le troisième quartile est le sixième terme de la série triée.

$Q_3 = 181,84 \text{ cm}$

Est associé aux quartiles une mesure de la dispersion, l'écart interquartile qui est l'écart entre le troisième et le premier quartile ($Q_3 - Q_1$). Il mesure la dispersion des 50 % centraux de la population.

$Q_3 - Q_1 = 181,84 - 178,32 = 3,52 \text{ cm}.$

1.5.2 Déciles

Par analogie, on définit 9 déciles qui partagent la population par tranche de 10%.

Le premier décile (D_1) d'une série statistique est la valeur de la série qui partage la population en 2 groupes, le premier groupe contenant au moins 10% des effectifs : au moins 10% des effectifs sont inférieurs ou égaux au premier décile.

Le second décile (D_2) d'une série statistique est la valeur de la série qui partage la population en 2 groupes, le premier groupe contenant au moins 20% des effectifs : au moins 20% des effectifs sont inférieurs ou égaux au second décile.

Et ainsi de suite avec D_5 qui est égal à la médiane.

De la même manière, on associe aux déciles une mesure de la dispersion qui est très utilisée en économie, l'écart interdécile qui est l'écart entre le neuvième et le premier décile ($D_9 - D_1$).

Dans notre exemple :

$Taille = \{177,98 ; 178,32 ; 180,24 ; 180,74 ; 181,27 ; 181,84 ; 182,54\}$

$\#Taille = 7$

$10\% * 7 = 0,7$

On prend, le premier entier supérieur à 0,7, soit 1, le premier décile est le premier terme de la série triée.

$D_1 = 177,98 \text{ cm}$

$90\% * 7 = 6,3$

On prend, le premier entier supérieur à 6,3, soit 7, le neuvième décile est le septième terme de la série triée.

$D_9 = 182,54 \text{ cm}$

L'écart interdécile = $D_9 - D_1 = 182,54 - 177,98 = 4,56 \text{ cm}$

1.6 Représentations graphiques usuelles

Dans le cadre des statistiques univariées, la représentation graphique la plus importante est le graphique qui représente l'effectif de classe en fonction du caractère.

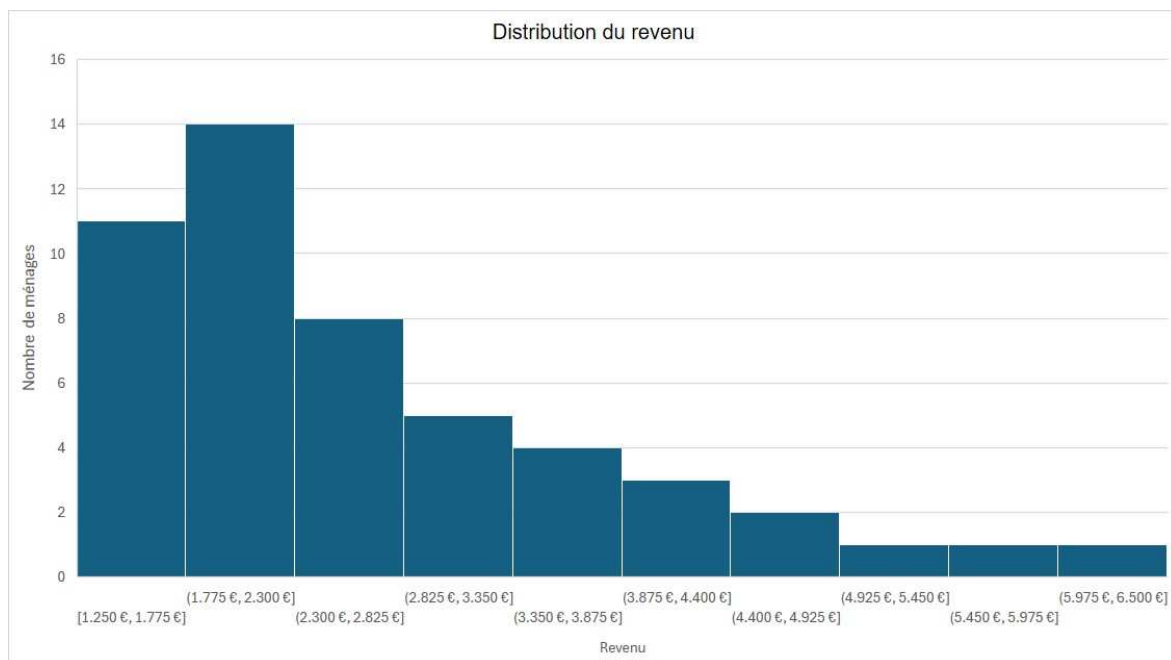
1.6.1 Caractère quantitatif

Pour un caractère quantitatif, on utilise généralement l'histogramme ou la boîte à moustaches (box-plot en Anglais).

Histogramme est un graphique qui donne l'effectif de classe, en valeur absolue ou en pourcentage, en fonction des intervalles de classe ou de la valeur centrale des classes.

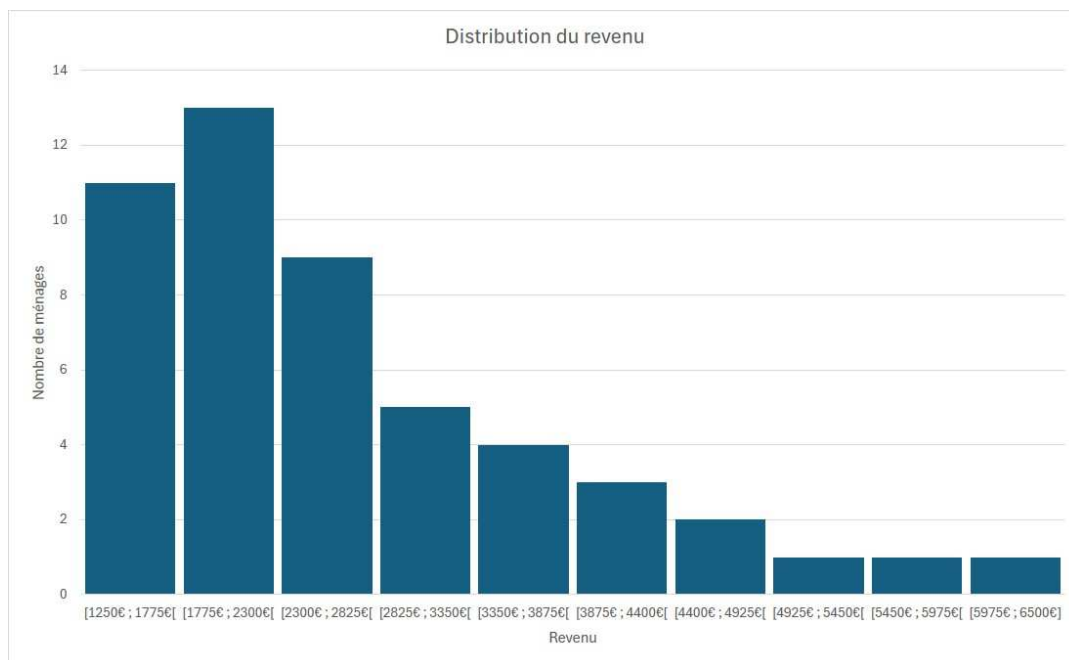
Exemples :

Histogramme Excel par défaut :

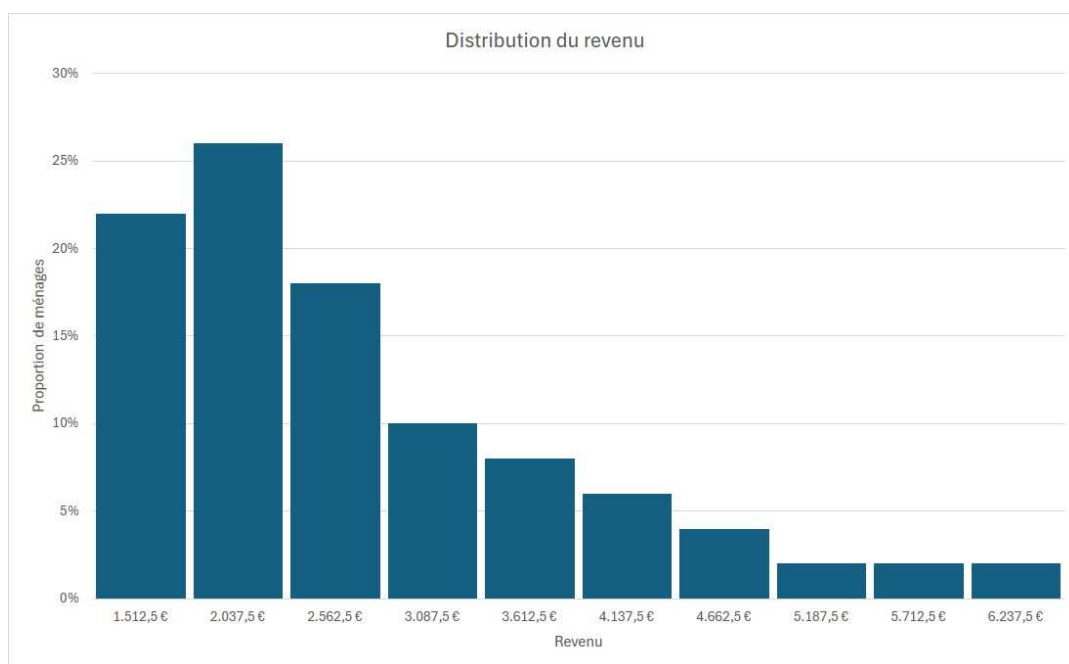


C'est très rapide à faire mais on ne peut pas modifier l'écriture des classes qui est erronée.

Nous apprendrons à faire des histogrammes personnalisés aux TP.



Ou



La boîte à moustaches (Box-plot), permet de comparer différentes séries statistiques.

De haut en bas :

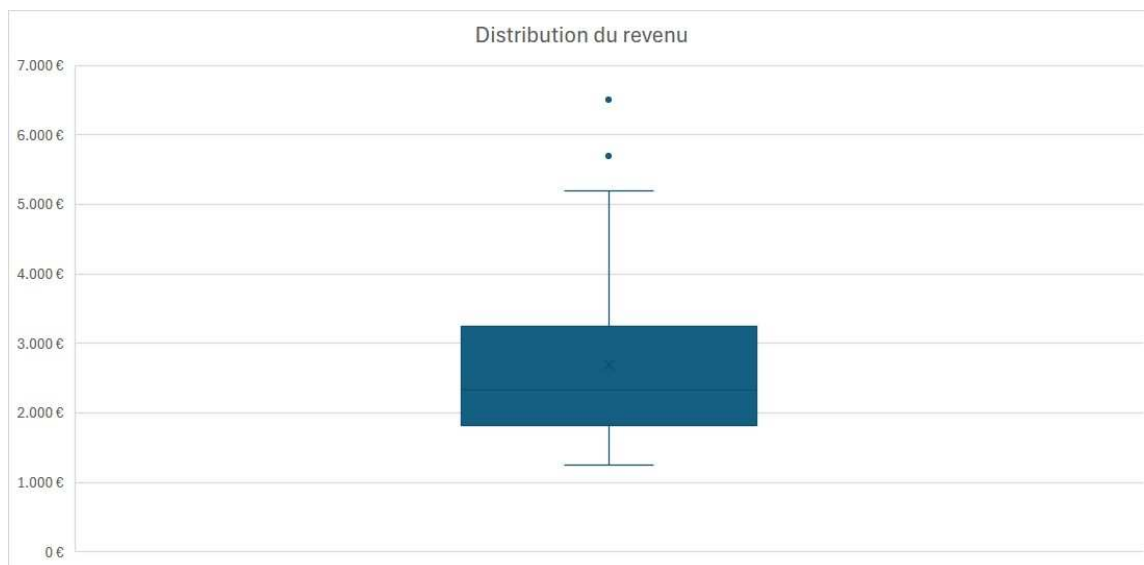
La moustache du dessus le maximum.

Le sommet de la boîte le 3^{ème} quartile.

La ligne dans la boîte, la médiane.

La base de la boîte le 1^{er} quartile.

La moustache du bas le minimum.



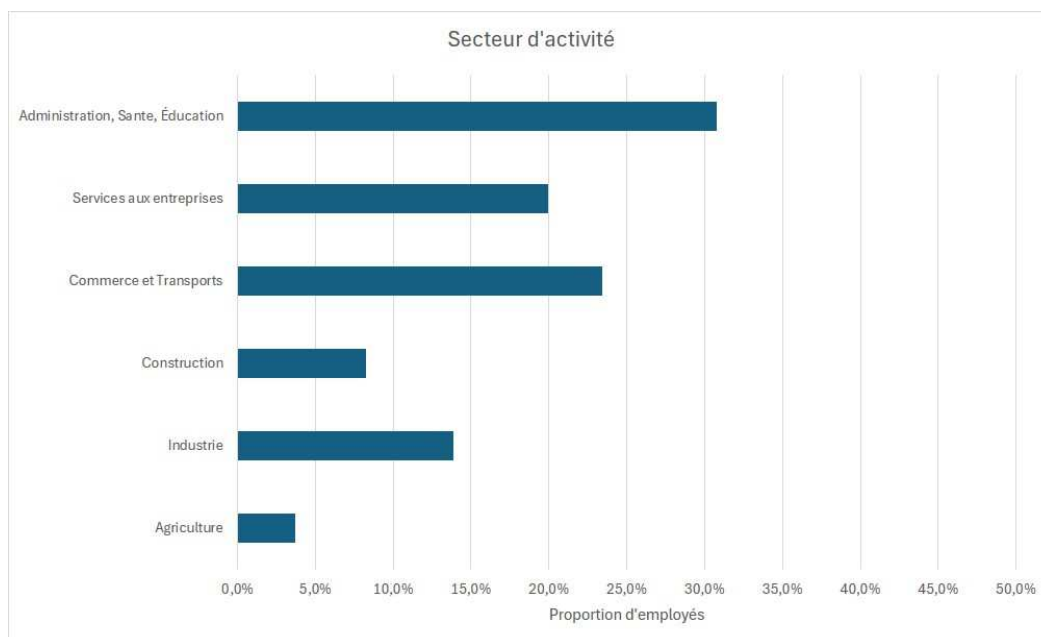
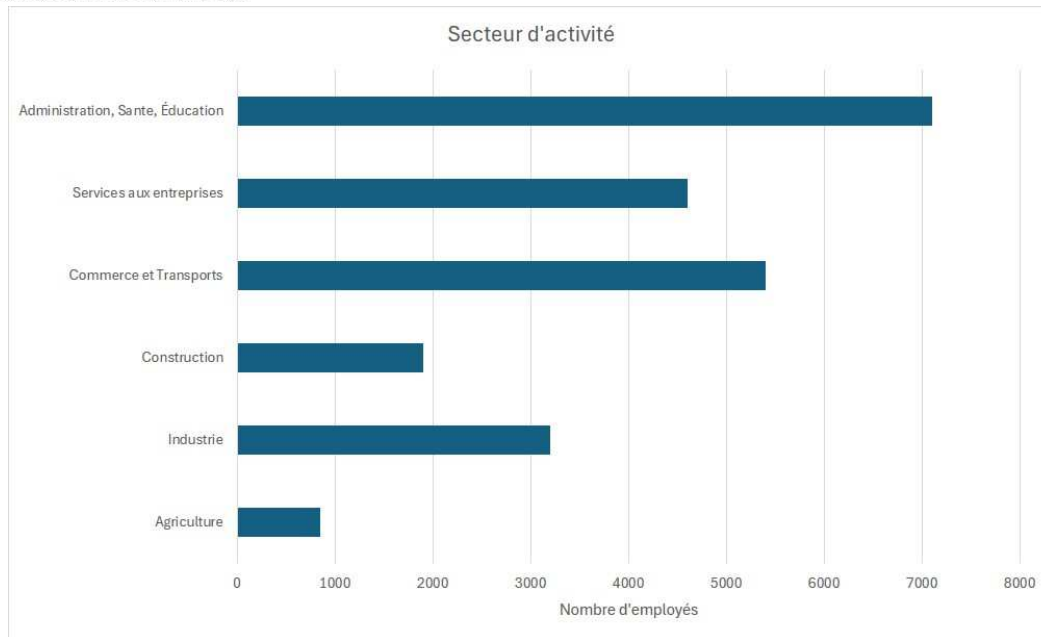
Attention Excel considère qu'une valeur est aberrante (outliers) si elle plus grande que $Q_3 + 1,5 \cdot (Q_3 - Q_1)$ ou plus petite que $Q_1 - 1,5 \cdot (Q_3 - Q_1)$.

Il y a moyen de construire une boîte à moustache personnalisée mais c'est relativement complexe...

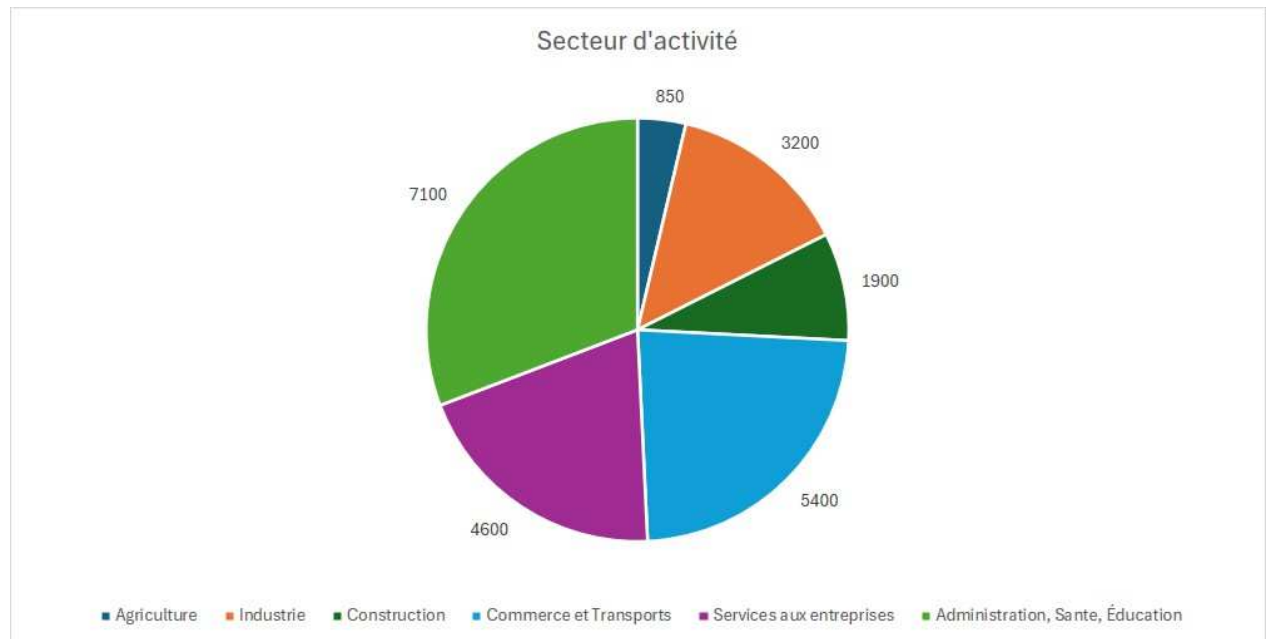
1.6.2 Caractère qualitatif

Pour un caractère qualitatif, on utilise généralement les diagrammes en bâtons (Bar Chart) ou les diagrammes circulaires (Camembert, Pie Chart)

Le diagramme en bâtons est similaire à l'histogramme mais il permet d'avoir des catégories dont le nom est généralement plus long pour un caractère qualitatif. Il représente l'effectif des différentes classes, soit en valeur absolue soit en pourcentage.



Le diagramme circulaire, permet de visualiser en même temps la proportion relative et le nombre absolu. Il est très efficace sauf si le nombre de classes est trop grand ou si certaines classes sont très petites par rapport aux autres.

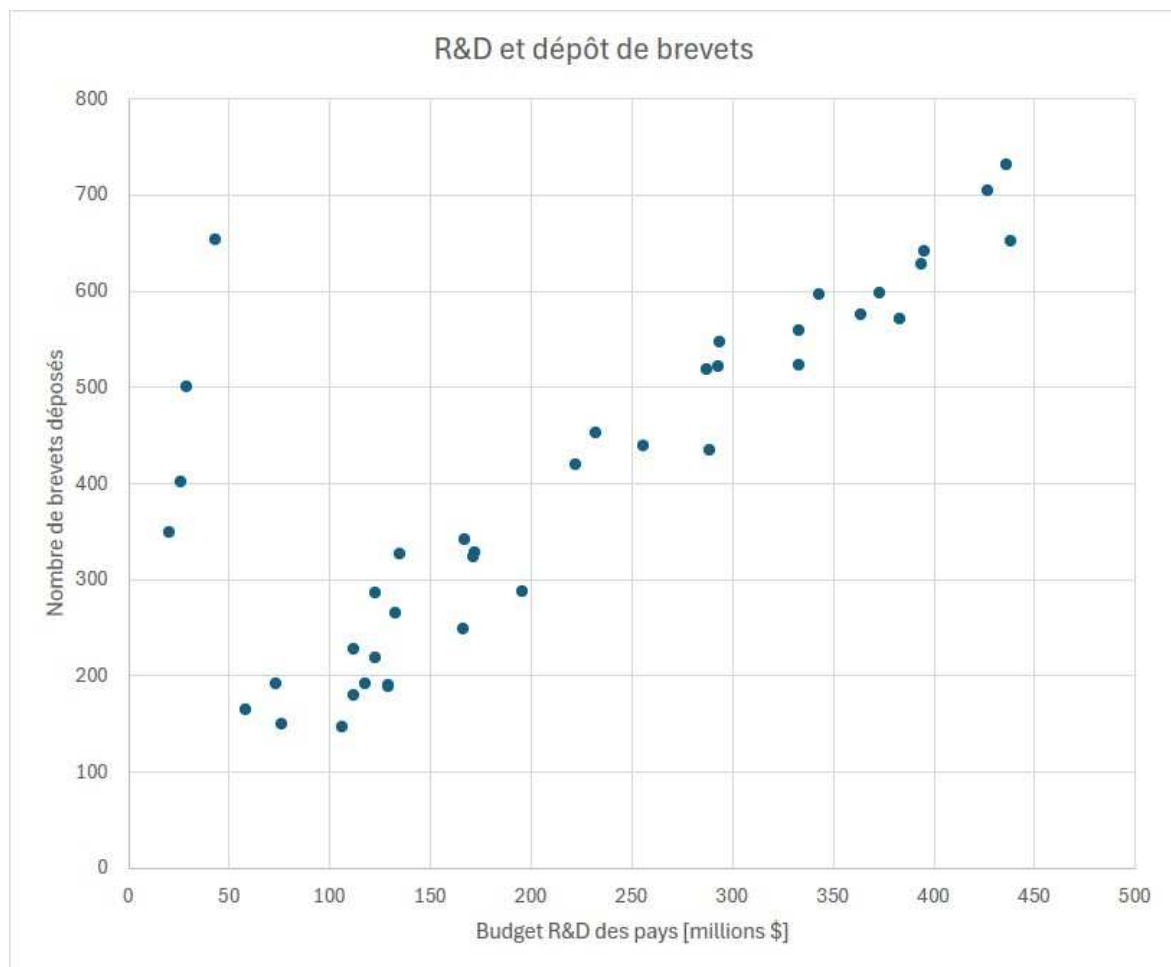


2. Statistique bivariable

2.1 Introduction

La statistique bivariable s'applique à une population (ou un échantillon) dans lequel on a mesuré deux caractères quantitatifs. L'objet de cette statistique est d'analyser la relation éventuelle entre ces deux caractères.

La représentation graphique privilégiée en statistique bivariable est le nuage de points (scatterplot). Cette représentation consiste à représenter les variables dans un repère orthogonale. Les individus de la population sont représentés par un point dont les coordonnées sont les deux caractères mesurés de l'individu.



2.2 Ajustement statistique

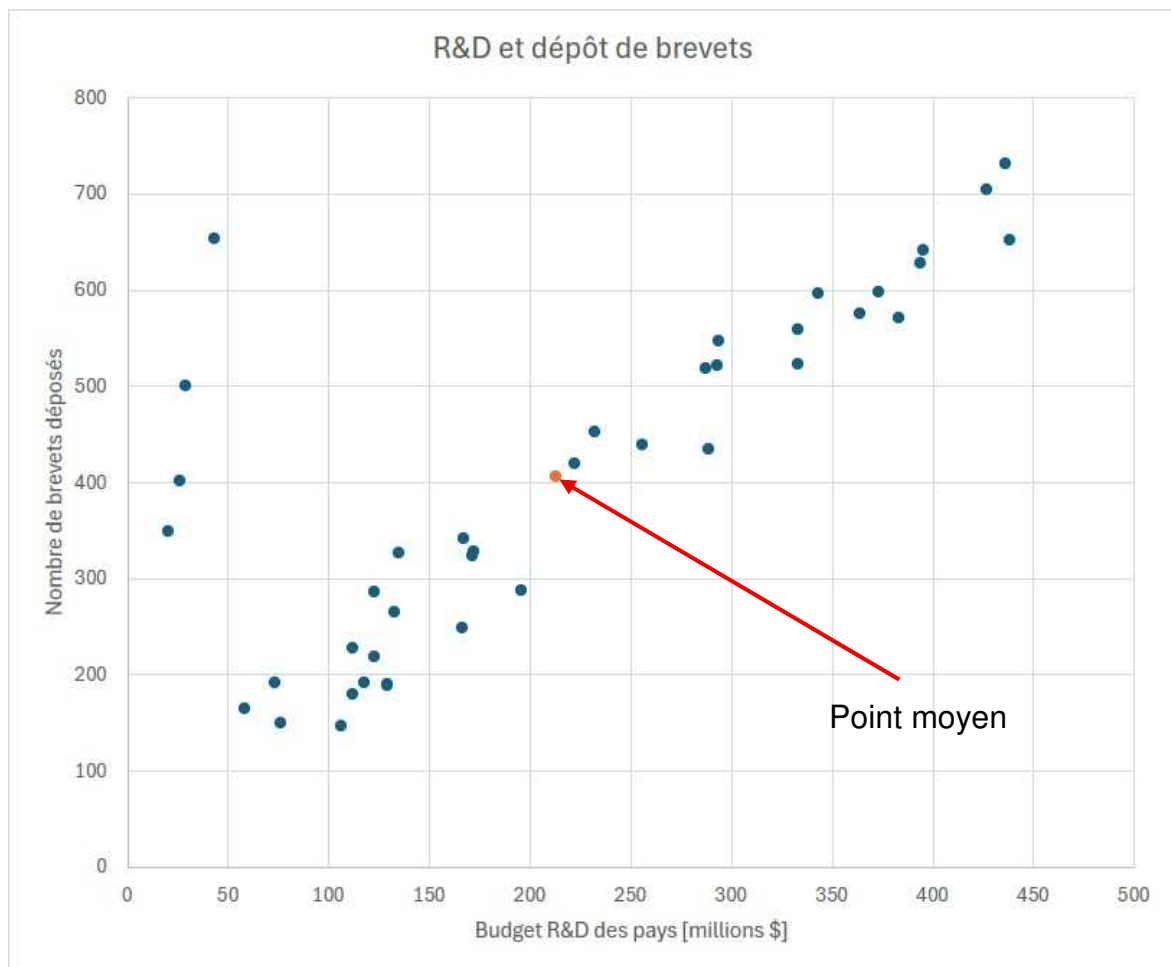
Un nuage de points n'est généralement pas le graphique d'une fonction. Or manipuler des fonctions est souvent plus facile. Les statisticiens essayent donc de trouver une fonction dont le graphique passe le plus près possible des points du nuage. La recherche d'une telle fonction porte le nom d'ajustement statistique.

2.2.1 Ajustement linéaire

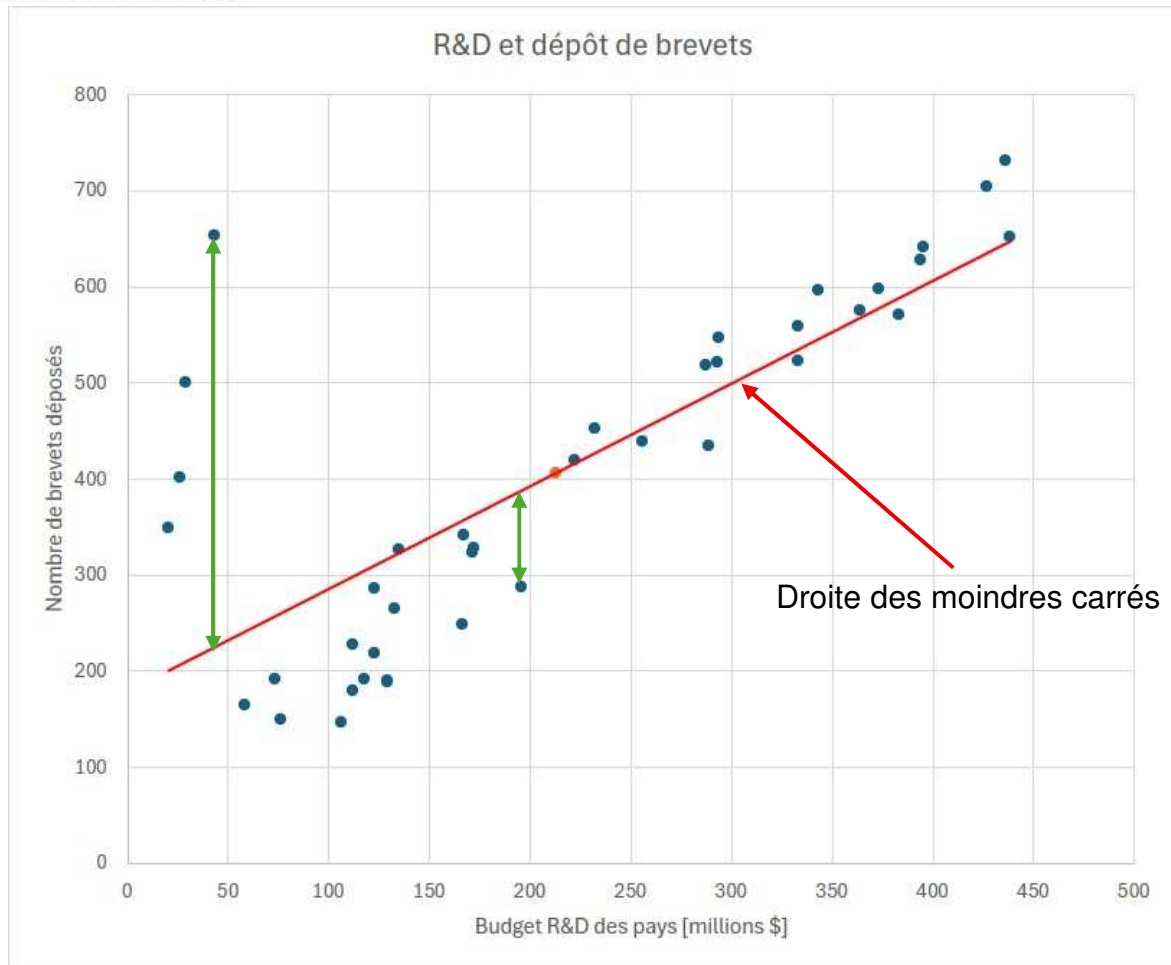
L'ajustement statistique le plus commun est l'ajustement linéaire. Dans celui-ci, la fonction recherchée est une fonction du premier degré c'est-à-dire une fonction dont le graphique est une droite.

On appelle point moyen ou barycentre un point, qui a priori, n'appartient pas à la série statistique mais dont les coordonnées sont les valeurs moyennes des deux caractères.

Dans notre exemple, la moyenne des brevets déposés est $m_{brevets} = 406,7$ et la moyenne des budgets R&D est $m_{R\&D} = 212,6$ millions \$.



On appelle droite des moindres carrés, la droite qui passe par le point moyen et pour laquelle la somme des carrés de toutes les distances verticales des points du nuage à la droite est la plus petite.



La droite des moindres carrés est ajustée pour que la somme des distances en vert (pour l'ensemble des points) soit minimale.

2.2.2 Covariance

Nous n'entrerons pas dans les détails mais la pente de la droite dépend d'un paramètre appelé covariance $cov(x, y)$. Où x et y sont les deux caractères de notre série statistique.

$$cov(x, y) = \frac{(x_1 - m_x) \cdot (y_1 - m_y) + (x_2 - m_x) \cdot (y_2 - m_y) + \dots + (x_N - m_x) \cdot (y_N - m_y)}{N}$$

Avec :

m_x la moyenne du paramètre x (Le budget R&D dans notre exemple)

m_y la moyenne du paramètre y (Le nombre de brevets)

N l'effectif total de notre série.

La pente de la droite étant donnée par le rapport entre la covariance et la variance du paramètre x .

2.2.1 Corrélation

Comme nous l'avons mentionné dans l'introduction, l'objet de la statistique bivariée est de déterminer s'il existe une relation entre les deux caractères étudiés.

On va donc « mesurer » l'intensité de la relation existante entre deux caractères.

Cette mesure se fait à l'aide d'un paramètre nommé coefficient de corrélation ou coefficient de Pearson.

$$r(x, y) = \frac{cov(x, y)}{\sqrt{V_x \cdot V_y}} = \frac{cov(x, y)}{\sigma_x \cdot \sigma_y}$$

Avec

$r(x, y)$ le coefficient de corrélation

$cov(x, y)$ la covariance

V_x la variance du caractère x

V_y la variance du caractère y

σ_x l'écart type du caractère x

σ_y l'écart type du caractère y

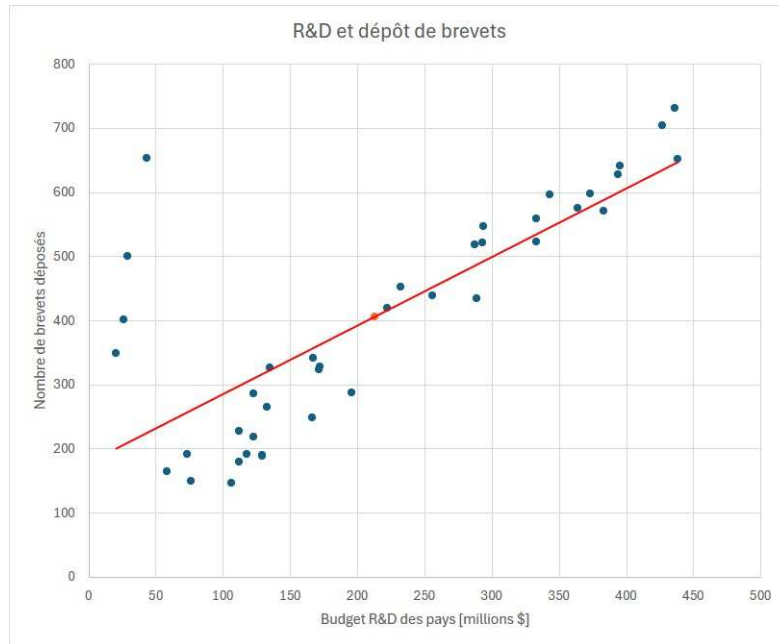
Plus le coefficient de corrélation est proche de 1 ou de -1, plus les deux caractères étudiés sont corrélés, c'est-à-dire « liés ».

De même, plus le coefficient de corrélation est proche de 1 ou de -1, plus le nuage de points ressemble à une droite.

Si le coefficient de corrélation est positif, la corrélation est positive (la droite 'monte') s'il est négatif, elle est négative (la droite 'descend').

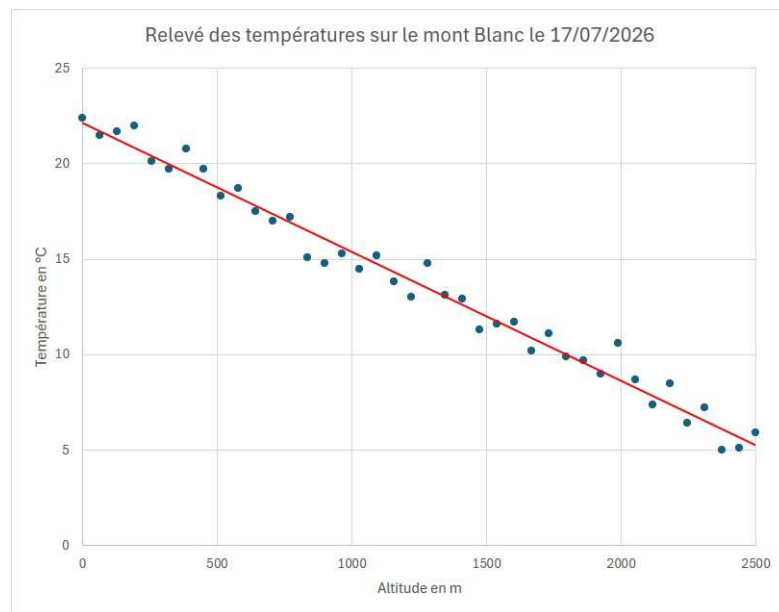
Exemples :

Corrélation positive



$$r(\text{Budget R\&D}, \text{Brevets}) = 0,77$$

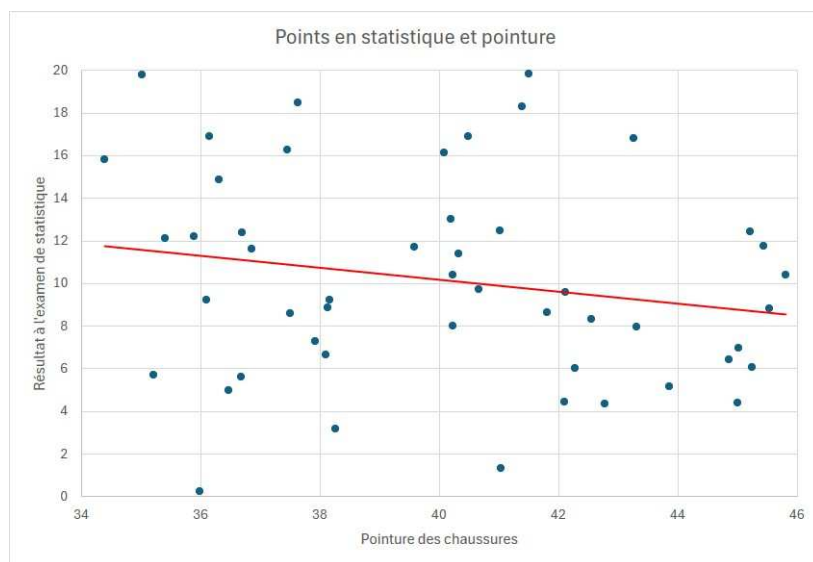
Corrélation négative



$$r(\text{Altitude}, \text{Température}) = -0,98$$

A l'inverse, si le coefficient de corrélation est proche de zéro, les caractères ne sont pas liés et le nuage de points ne ressemble pas à une droite.

Pas de corrélation



$$r(\text{Pointure}, \text{Résultat examen}) = -0,16$$

Type de corrélation	Positive	Négative
Corrélation parfaite	$0,98 < r < 1$	$-1 < r < -0,98$
Corrélation forte	$0,80 < r < 0,98$	$-0,98 < r < -0,80$
Corrélation moyenne	$0,60 < r < 0,80$	$-0,80 < r < -0,60$
Corrélation faible	$0,35 < r < 0,60$	$-0,60 < r < -0,35$
Corrélation nulle	$-0,35 < r < 0,35$	

2.2.2 Corrélation et causalité

Malgré la présence d'une corrélation importante entre deux caractères, il faut éviter toute interprétation abusive.

Une forte corrélation n'implique pas nécessairement une relation directe de cause à effet entre les deux variables.

Considérons la droite de régression qui met en relation la vente de produits solaires et la vente de crèmes glacées ou encore celle qui compare la vente de lunettes de soleil et la consommation de boissons rafraîchissantes. Elles possèdent toutes les deux un excellent coefficient de corrélation et pourtant aucun lien de cause à effet ne lie les deux phénomènes étudiés.

En réalité, ces deux variables dépendent d'une troisième : la température extérieure.

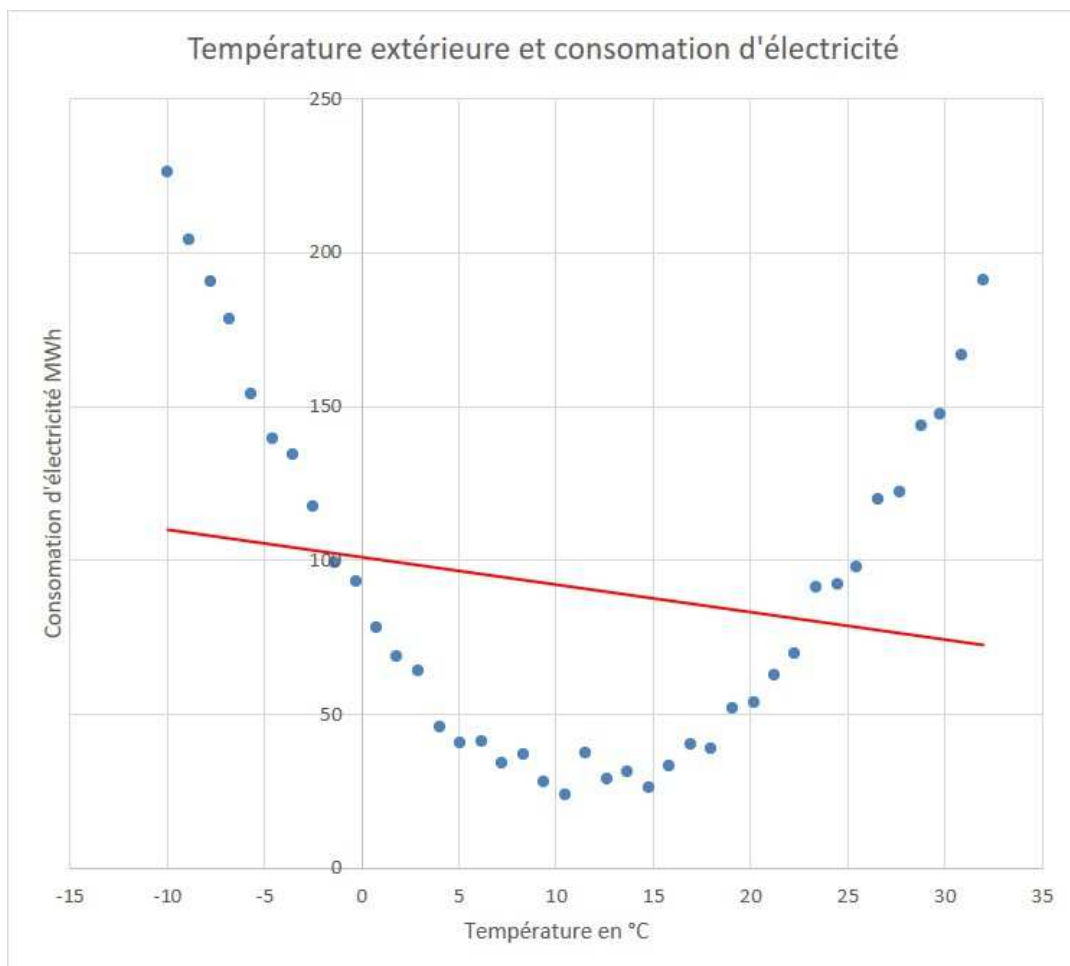
Il se peut aussi que deux variables fortement corrélées n'aient aucun lien direct.

Par exemple, au fil des années, à cause d'un phénomène appelé la subsidence, l'altitude moyenne de Pays-Bas baisse et dans le même temps la population des Pays-Bas augmente. Cette excellente corrélation ne révèle, bien évidemment, aucune relation entre le poids des Hollandais et l'enfoncement du Pays....

2.2.3 Ajustements non-linéaire

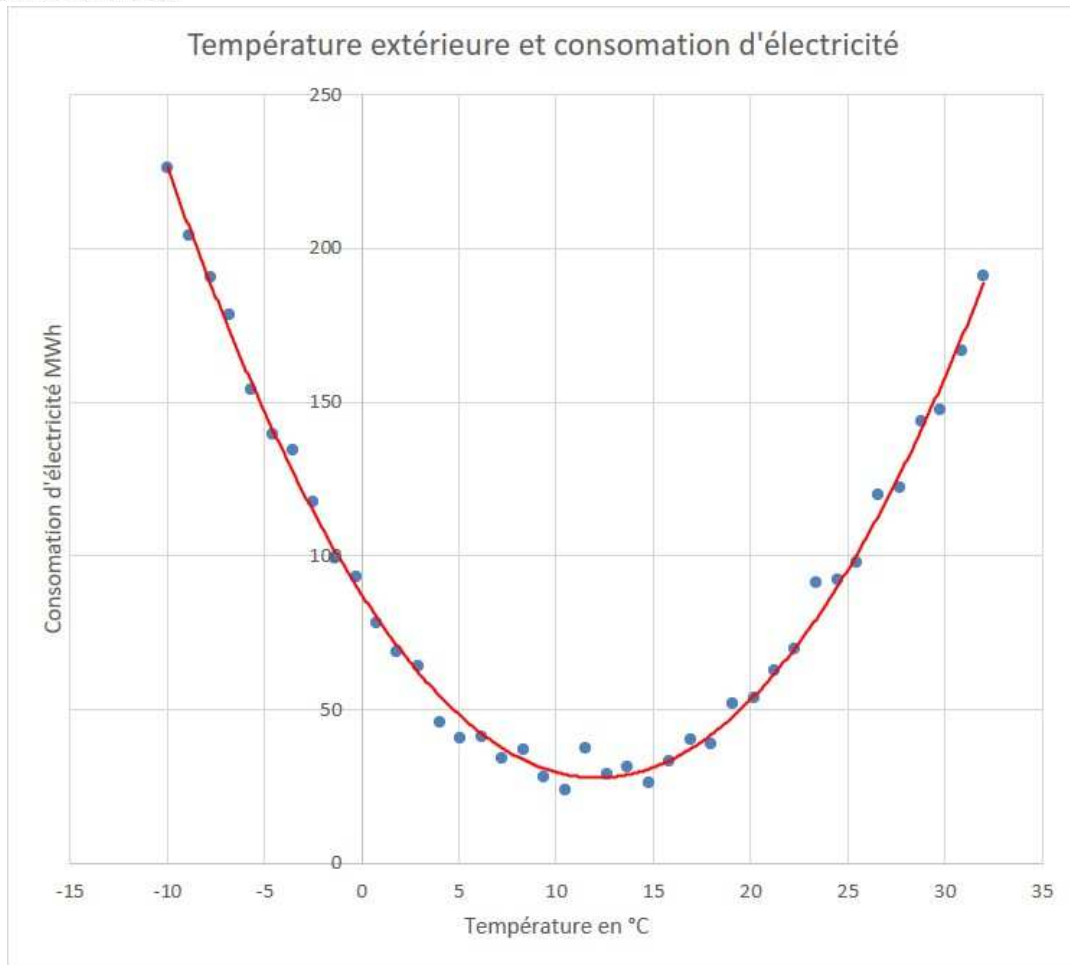
Il peut arriver que deux caractères ne soient pas liés par une relation linéaire (une droite) mais par une relation plus complexe. Dans ce cas, l'utilisation d'un ajustement linéaire peut mener à de mauvaises conclusions.

Exemple :



$$r(\text{Température}, \text{Consommation}) = -0,19$$

On pourrait en conclure que les caractères ne sont pas corrélés. Il convient alors, de s'orienter vers des ajustements plus complexes.

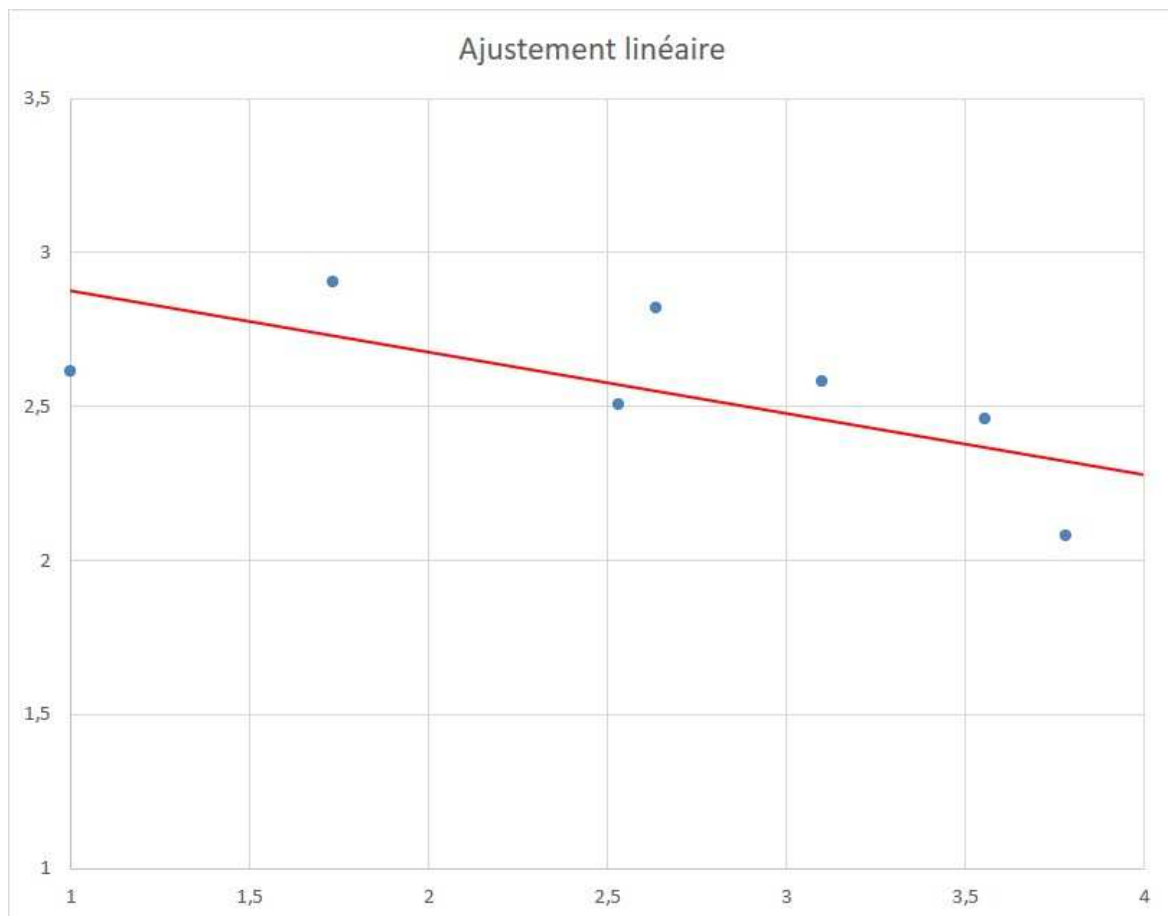
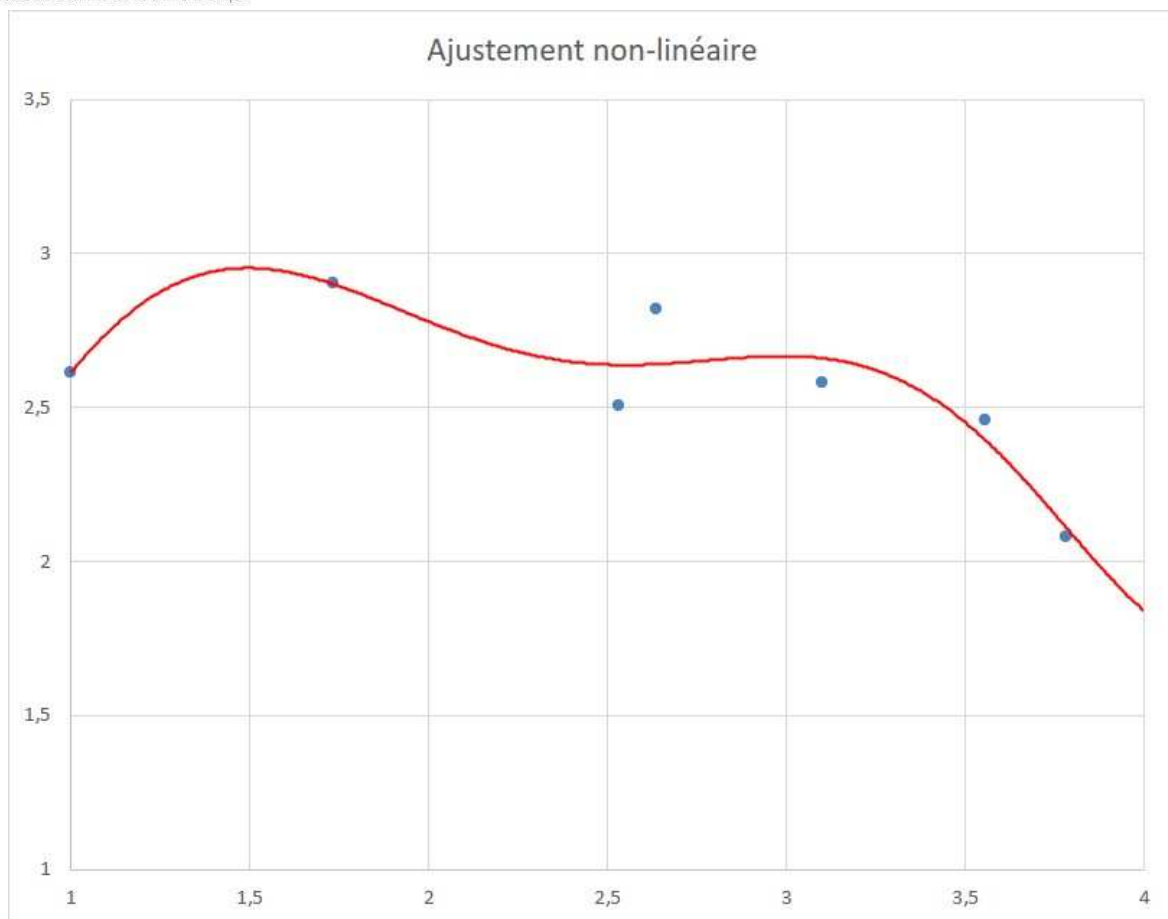


Ici, nous avons un ajustement polynomial du deuxième degré qui permet de très bien caractériser la relation entre les deux caractères.

Sans entrer dans les détails, il existe un coefficient similaire au coefficient de corrélation, le coefficient de détermination.

Attention, il convient d'être extrêmement prudent avec les ajustements non-linéaires dans la mesure où, on peut toujours trouver une courbe qui ajuste des points. Je conseille d'utiliser des ajustements non-linéaires uniquement si on connaît le type de relation qui lie les deux caractères.

L'exemple suivant vous permet de comparer un ajustement polynomial du 6^{ème} degré avec un ajustement linéaire sur des caractères qui sont liés par une relation linéaire.



3. Eléments de probabilité

3.1 Définitions

3.1.1 Expérience aléatoire

Une expérience aléatoire est une expérience qui donne lieu à plusieurs résultats possibles dont on ne peut prédire lequel se réalisera. Chacun de ces résultats porte le nom d'épreuve de l'expérience.

Exemple : On lance un dé et on regarde la face supérieure.

3.1.2 Catégorie d'épreuve

La catégorie d'épreuves, notée Ω , est l'ensemble de tous les résultats possibles de l'expérience aléatoire.

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

3.1.3 Événement

Un événement est un sous-ensemble de la catégorie d'épreuves. Il est noté par une majuscule latine. Les événements se limitant à un seul résultat sont appelés événements élémentaires.

Exemples :

- A = tirer un chiffre pair = $\{2, 4, 6\}$
- B = tirer un chiffre strictement supérieur à 4 = $\{5, 6\}$
- C = tirer un 7 = ϕ (ensemble vide)

Un événement se produit lorsque le résultat obtenu par l'expérience appartient à cet événement.

En particulier,

L'événement Ω se produit **toujours**, c'est l'événement certain.

L'événement ϕ ne se produit **jamais**, c'est l'événement impossible.

Exemples :

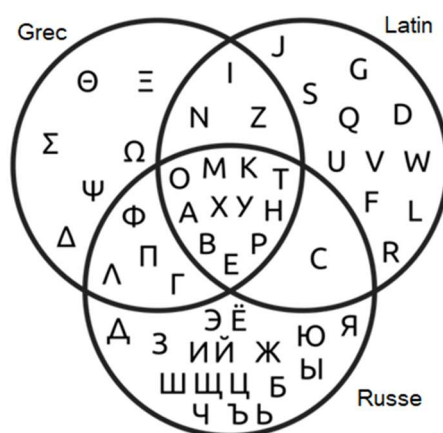
- Événement certain = tirer un chiffre entre 1 et 6.
- Événement impossible = tirer le chiffre 7.

3.2 Diagramme de Venn

3.2.1 Introduction

Les diagrammes de Venn sont utilisés pour enseigner la théorie des ensembles ainsi que pour illustrer des relations simples en probabilité.

A titre d'exemple, voici un diagramme de Venn qui illustre les caractères en majuscules partagés par l'alphabet grec, latin et russe.

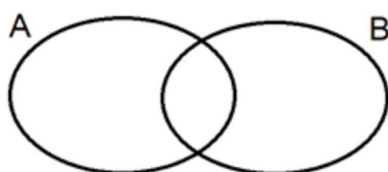


Grace à ce diagramme, on peut constater que le Latin partage douze lettres majuscules avec le Russe (O, M, K, T, A, X, Y, H, B, P, E et C). La lettre C n'est pas utilisée en Grec et les lettres I, N et Z sont utilisées en Grec et Latin mais pas en Russe.

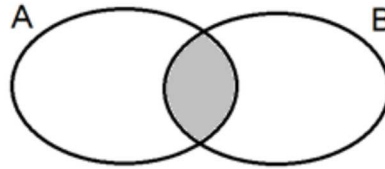
3.2.2 Théorie des ensembles

Afin d'utiliser pleinement les possibilités de ce diagramme, nous allons rappeler quelques notions de la théorie des ensembles.

Soit deux ensembles A et B :

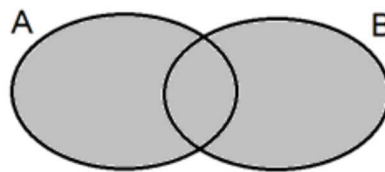


- $A \cap B$: A **inter** B, Intersection de A et de B, A **et** B.



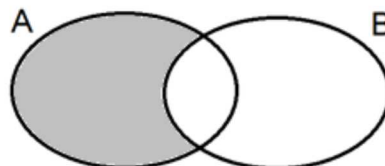
En probabilité, l'événement $A \cap B$ se produit si l'événement A et l'événement B se produisent simultanément.

- $A \cup B$: A **union** B, A **ou** B.



En probabilité, l'événement $A \cup B$ se produit si l'événement A ou l'événement B se produit.

- $A \setminus B$: A **sans** B.



En probabilité, l'événement $A \setminus B$ se produit si l'événement A se produit sans que l'événement B ne se produise.

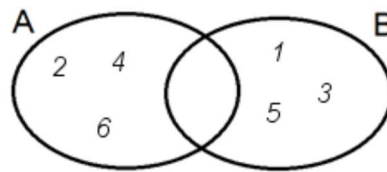
3.2.3 Evénements contraires

Des événements contraires sont des événements dont l'intersection est ϕ et dont l'union est Ω .

Exemple : Tirer un dé à six faces, $\Omega = \{1, 2, 3, 4, 5, 6\}$

$A = \text{tirer un nombre pair} = \{2, 4, 6\}$

$B = \text{tirer un nombre impair} = \{1, 3, 5\}$



$$A \cap B = \{2, 4, 6\} \cap \{1, 3, 5\} = \{\} = \phi$$

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

En conséquence A et B sont des événements contraires. L'événement contraire de A est noté $\bar{A}$.

Dans notre exemple, $\bar{A} = B$

3.2.4 Cardinal d'un événement

Le nombre de résultats d'un événement A est le nombre d'éléments de A . On le note $\#A$, qui se lit cardinal de A .

Dans l'exemple précédent :

$$\#A = 3$$

$$\#B = 3$$

$$\#\Omega = \#(A \cup B) = 6$$

$$\#(A \cap B) = 0$$

3.2.5 Fréquence et probabilité

La fréquence de réalisation d'un événement, $f(A)$ est définie par le rapport suivant.

$$f(A) = \frac{\text{Nombre de fois ou } A \text{ se produit}}{\text{Nombre d'expériences}}$$

Comme le nombre de fois ou A se produit est $\leq$ au nombre d'expériences,

$$0 \leq f(A) \leq 1$$

Par définition, la probabilité que A se produise, $P(A)$ est donnée par :

$$P(A) = \lim_{N \rightarrow \infty} f(A) = \lim_{N \rightarrow \infty} \frac{\text{Nombre de fois ou } A \text{ se produit}}{N}$$

Avec N , le nombre d'expériences.

En pratique, on ne doit pas faire une infinité d'expériences, on doit en faire un grand nombre. Pour fixer les idées, généralement, on considère qu'un nombre supérieur à 30 est un grand nombre.

3.2.6 Axiomes de Kolmogorov

Les axiomes de Kolmogorov sont les trois piliers fondateurs qui permettent de définir mathématiquement une probabilité sur un ensemble.

- Positivité

$$P(A) \geq 0$$

La probabilité d'un événement est positive.

- Unitalité

$$P(\Omega) = 1 = 100\%$$

La probabilité d'un événement certain vaut 1.

Exemple : Tirer un dé à six faces, $\Omega = \{1, 2, 3, 4, 5, 6\}$

La probabilité de tirer un nombre entre 1 et 6 : $P(\Omega) = 1 = 100\%$

- Additivité

$$\text{Si } A \cap B = \emptyset \text{ alors } P(A \cup B) = P(A) + P(B)$$

Si deux événements sont disjoints (pas d'intersection) alors, la probabilité que l'un ou l'autre se produise est la somme de leurs probabilités.

Exemple : Tirer un dé à six faces, $\Omega = \{1, 2, 3, 4, 5, 6\}$

Soit l'événement A consistant à tirer un nombre entre 1 et 2.

$$A = \{1, 2\}$$

Soit l'événement B consistant à tirer un nombre entre 5 et 6.

$$B = \{5, 6\}$$

$$A \cap B = \{\} = \phi$$

$$P(A \cup B) = P(A) + P(B)$$

La probabilité de tirer un nombre entre 1 et 2 ou un nombre entre 5 et 6 est égale à la probabilité de tirer un nombre entre 1 et 2 plus la probabilité de tirer un nombre entre 5 et 6.

Conséquences directes des axiomes de Kolmogorov :

- $0 \leq P(A) \leq 1$
La probabilité de n'importe quel événement est comprise entre 0 et 1.
- $P(\phi) = 0$
La probabilité d'un événement impossible est nulle.

Exemple : Tirer un dé à six faces, $\Omega = \{1, 2, 3, 4, 5, 6\}$

Soit l'événement A consistant à tirer le nombre 7.

$$A = \{\} = \phi$$

$$P(A) = 0$$

- $P(\bar{A}) = 1 - P(A)$
La probabilité d'un événement contraire vaut 1 moins la probabilité de l'événement.

Exemple : Tirer un dé à six faces, $\Omega = \{1, 2, 3, 4, 5, 6\}$

Soit l'événement A consistant à tirer un nombre supérieur à 5.

$$A = \{6\}$$

Alors,

$$\bar{A} = \{1, 2, 3, 4, 5\}$$

Et

$$P(\bar{A}) = 1 - P(A)$$

La probabilité de tirer un nombre inférieur ou égal à 5 vaut 1 moins la probabilité de tirer un 6

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
La probabilité qu'un événement ou l'autre se produise est la somme de leurs probabilités moins la probabilité qu'ils se produisent simultanément.

Exemple : Tirer un dé à six faces, $\Omega = \{1, 2, 3, 4, 5, 6\}$

Soit l'événement A consistant à tirer un nombre supérieur à 3.

$$A = \{4, 5, 6\}$$

Soit l'événement B consistant à tirer un nombre impair.

$$B = \{1, 3, 5\}$$

Alors,

$$A \cap B = \{5\}$$

et,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

La probabilité de tirer un nombre supérieur à 3 ou un nombre impair est égale à la probabilité de tirer un nombre supérieur à 3 plus la probabilité de tirer un nombre impair moins la probabilité de tirer un 5.

3.2.7 Événements équiprobables

Des événements qui ont la même probabilité d'occurrence sont dit équiprobables.

Si tous les événements d'une expérience aléatoire sont équiprobables, alors la probabilité de tout événement A de cette expérience est donnée par :

$$P(A) = \frac{1}{\#\Omega}$$

Avec $\#\Omega$, le nombre d'événements de l'expérience.

Exemple : Tirer un dé à six faces, $\Omega = \{1, 2, 3, 4, 5, 6\}$

La probabilité de tirer le nombre 3 : $P(3) = \frac{1}{\#\Omega} = \frac{1}{6} = 0,17 = 17\%$

Cette dernière formule sera très largement utilisée pour déterminer une probabilité, la détermination du nombre d'événements d'une expérience est une des facettes essentielles du calcul de probabilités.

3.2.8 Événements indépendants

Considérons deux expériences aléatoires successives soit A un événement de la première expérience et B un événement de la deuxième expérience.

A et B sont indépendants si et seulement si :

$$P(A \cap B) = P(A).P(B)$$

Exemple : Tirer deux jetons numérotés dans un sac contenant 10 jetons,

$$\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

Déterminons la probabilité de tirer un 3 puis de tirer un 7.

Les événements sont supposés équiprobables, si on tire un jeton dans le sac.

Soit l'expérience A , tirer le nombre 3 .

$$\text{La probabilité de tirer le nombre 3 : } P(A) = \frac{1}{\#\Omega} = \frac{1}{10} = 0,1 = 10\%$$

Soit l'expérience B , tirer le nombre 7 .

$$\text{La probabilité de tirer le nombre 7 : } P(B) = \frac{1}{\#\Omega} = \frac{1}{10} = 0,1 = 10\%$$

Ici nous devons considérer deux cas de figure.

Le premier, après avoir tiré le premier jeton on le remet dans le sac, ce que l'on appelle un tirage avec remise. Dans ce cas, les événements sont indépendants et la probabilité de tirer un 3 puis un 7 est donnée par :

$$P(A \cap B) = P(A).P(B) = 0,1 . 0,1 = 0,01 = 1\%$$

Par contre si on ne remet pas le jeton dans le sac, les événements ne sont plus indépendants. En effet, $\#\Omega$ est modifié entre la première et la deuxième expérience, il ne reste plus que 9 jetons dans le sac.

La probabilité de tirer le nombre 3 : $P(A) = \frac{1}{\#\Omega} = \frac{1}{10} = 0,1 = 10\%$

Mais la probabilité de tirer le nombre 7 lors de la deuxième expérience, à condition que l'on ne l'a pas tiré dans la première expérience, est donnée par :

$$P(B) = \frac{1}{\#\Omega} = \frac{1}{9} = 0,11 = 11\%$$

$$P(A \cap B) = P(A) \cdot P(B) = 0,1 \cdot 0,11 = 0,011 = 1,1\%$$

3.2.9 Probabilité conditionnelle

Une probabilité conditionnelle est la probabilité d'un événement d'une expérience si un événement d'une autre expérience s'est produit.

Soient A et B deux événements de deux expériences successives, on note $P(B|A)$ la probabilité que B se produise sachant que A s'est produit.

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Dans l'exemple précédent, sans remise, si on tire un 3 à la première expérience. La probabilité de tirer un 7 lors de la deuxième expérience est donnée par :

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0,011}{0,1} = 0,11 = 11\%$$

3.3 Diagramme en arbre

Le diagramme en arbre est un outil qui nous permet de déterminer les probabilités d'événements lors d'expériences successives en représentant les différents cas possibles de chaque expérience, en indiquant pour chaque cas sa probabilité d'occurrence.

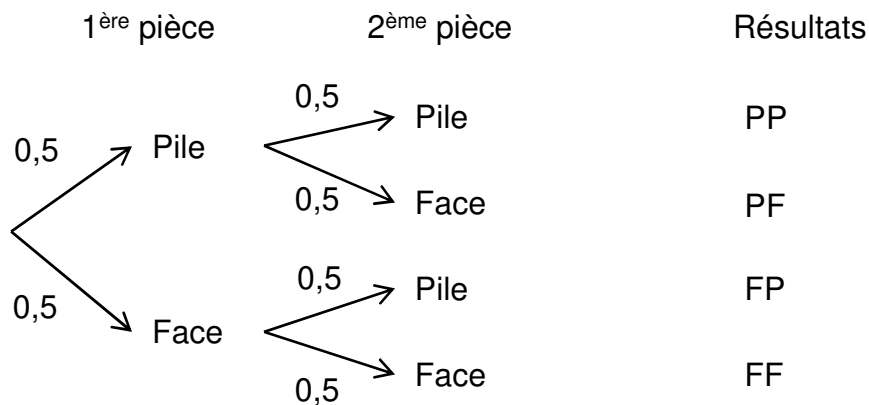
Considérons deux expériences successives qui consistent à tirer à pile ou face.

$$\Omega = \{P, F\}$$

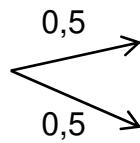
$$\Omega = 2$$

$$P(P) = \frac{1}{\#\Omega} = \frac{1}{2} = 0,5$$

$$P(F) = \frac{1}{\#\Omega} = \frac{1}{2} = 0,5$$



- La somme des probabilités issues d'un point vaut toujours 1.



$$0,5 + 0,5 = 1$$

- Lorsqu'on parcourt un chemin (branches consécutives) d'un arbre pour réaliser un événement, la probabilité composée correspondante s'obtient en multipliant les probabilités rencontrées sur chacune des branches.

Exemple : La probabilité de faire face puis pile (FP) est donc donnée par

$$P(FP) = 0,5 * 0,5 = 0,25 = 25\%$$

- Lorsque, pour rencontrer tous les cas d'un événement, on doit parcourir plusieurs chemins d'un même arbre, la probabilité totale correspondante est égale à la somme des probabilités relatives à chaque chemin.

Exemple : La probabilité de faire une fois face et une fois pile (FP) ou (PF) est donc donnée par

$$P(F \text{ et } P) = P(FP) + P(PF) = 0,25 + 0,25 = 0,5 = 50\%$$

3.4 Analyse combinatoire

Cette branche des mathématiques permet de calculer le nombre d'événements d'une expérience, de calculer $\# \Omega$.

3.4.1 Factorielle d'un naturel

Par définition, si n est un Naturel non nul la factorielle de n , notée $n !$ est donnée par :

$$n! = n. (n - 1). (n - 2) \dots (2). (1)$$

De plus, par convention

$$0! = 1$$

Exemples :

$$5! = 5.4.3.2.1 = 120$$

$$2! = 2.1 = 2$$

Remarque : La factorielle d'un réel est définie par la fonction gamma (Γ)

$$x! = \Gamma (x + 1) = \int_0^{\infty} t^x . e^{-t} dt$$

3.4.2 Permutations

Supposons que nous disposions de trois jetons, un rouge, un vert et un bleu, de combien de manière différente peut-on les classer ?

R, V, B

R, B, V

V, R, B

V, B, R

B, R, V

B, V, R

Soit 6 classements possibles. Ces différents classements sont appelés des permutations, il n'y a pas de répétition d'une même couleur dans un classement.

On peut montrer que le nombre de permutations de n éléments différents, $P(n)$ est donné par :

$$P_n = n!$$

Dans notre cas, on retrouve la valeur déterminée :

$$P_3 = 3! = 6$$

Exemple : On dispose d'un jeu de 52 cartes, si on le mélange, de combien de manière différentes le jeu peut-il être agencé ?

$$P_{52} = 52! = 8,06 \cdot 10^{67}$$

Soit, un peu plus que 80 milliards de milliards de milliards de milliards de milliards de milliards ..., c'est beaucoup, c'est même plus que le nombre d'atomes sur terre. Même si on joue beaucoup aux cartes, on a peu de chances de tomber deux fois sur la même distribution.

3.4.3 Permutations avec répétitions

Considérons le problème suivant, combien de mots différents (qu'ils aient un sens ou non) peut-on écrire avec les lettres du mot « ananas » ?

Le problème est, ici, un peu différent dans la mesure où certaines lettres vont se répéter. En effet, dans le mot « ananas », il y a : 3 a, 2 n et 1 s.

On appelle permutation avec répétitions de n éléments non nécessairement différents tout groupement ordonné formé de ces n éléments. Ce nombre a pour formule :

$$P_n^{i,j,\dots} = \frac{n!}{i!.j!.k! \dots}$$

Dans notre exemple, le nombre de mots formé par les lettres qui composent « ananas » vaut :

$$P_6^{3,2,1} = \frac{6!}{3!.2!.1!} = \frac{720}{6.2.1} = \frac{720}{12} = 60$$

Exemple : On dispose de 13 jetons, 6 rouges, 5 verts et 2 bleus, de combien de manière différente peut-on les classer ?

$$P_{13}^{6,5,2} = \frac{13!}{6!.5!.2!} = \frac{6227020800}{720 \cdot 120 \cdot 2} = \frac{6227020800}{172800} = 36036$$

3.4.4 Arrangements avec répétitions

Considérons la première série des nouvelles plaques d'immatriculations qui commencent par 1 suivi de 3 lettres puis de 3 chiffres (ex : 1-AZE-123). Les lettres et les chiffres pouvant être répétés (AAA, CZZ, 122 sont valides).

On appelle arrangement avec répétitions de n éléments différents pris p à p , tout groupement ordonné, formé de p de ces n éléments, dans lequel chaque élément peut figurer jusqu'à p fois.

$$B_n^p = n^p$$

Remarque importante : p peut être plus grand que n .

En conséquence, le nombre de combinaisons de lettres est donné par :

$$B_{26}^3 = 26^3 = 17\,576$$

Le nombre de combinaisons de chiffres vaut quant à lui :

$$B_{10}^3 = 10^3 = 1000$$

La première série de plaques d'immatriculations compte donc :

$$B_{26}^3 \cdot B_{10}^3 = 17\,576 \cdot 1000 = 17\,576\,000$$

Exemple : De combien de manières peut-on désigner, parmi 5 personnes, un président, un secrétaire et un trésorier sachant que les charges sont cumulables ?

$$B_5^3 = 5^3 = 125$$

3.4.5 Arrangements sans répétition

Reprenons l'exemple des plaques d'immatriculations mais, imaginons que les lettres ne peuvent plus se répéter (AAA, CZZ, ne sont plus valides).

On appelle arrangement sans répétition de n éléments différents pris p à p , tout groupement ordonné, formé de p de ces n éléments, dans lequel chaque élément figure une fois au plus.

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

Remarque importante : p est plus petit ou égal à n .

Le nombre de combinaisons de lettres devient :

$$A_{26}^3 = \frac{26!}{(26-3)!} = 15\,600$$

La première série de plaques d'immatriculations comptera donc :

$$A_{26}^3 \cdot B_{10}^3 = 15\,600 \cdot 1000 = 15\,600\,000$$

Exemple : Si 50 personnes participent à une course, combien de podiums différents sont possibles ?

$$A_{50}^3 = \frac{50!}{(50-3)!} = 117\,600$$

3.4.6 Combinaisons sans répétition

Essayons, à présent, de calculer le nombre de tirages possibles au Lotto. Dans ce cas-ci, On tire, 6 boules parmi 45, l'ordre du tirage n'a pas d'importance.

On appelle combinaison sans répétition de n éléments différents pris p à p , tout groupement non ordonné, formé de p de ces n éléments, dans lequel chaque élément figure une fois au plus.

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

Dans notre exemple, le nombre de tirage possible vaut :

$$C_{45}^6 = \frac{45!}{6! \cdot (45 - 6)!} = 8\,145\,060$$

Exemple, dans une classe de 32 élèves, de combien de manières peut-on former des groupes de 4 ?

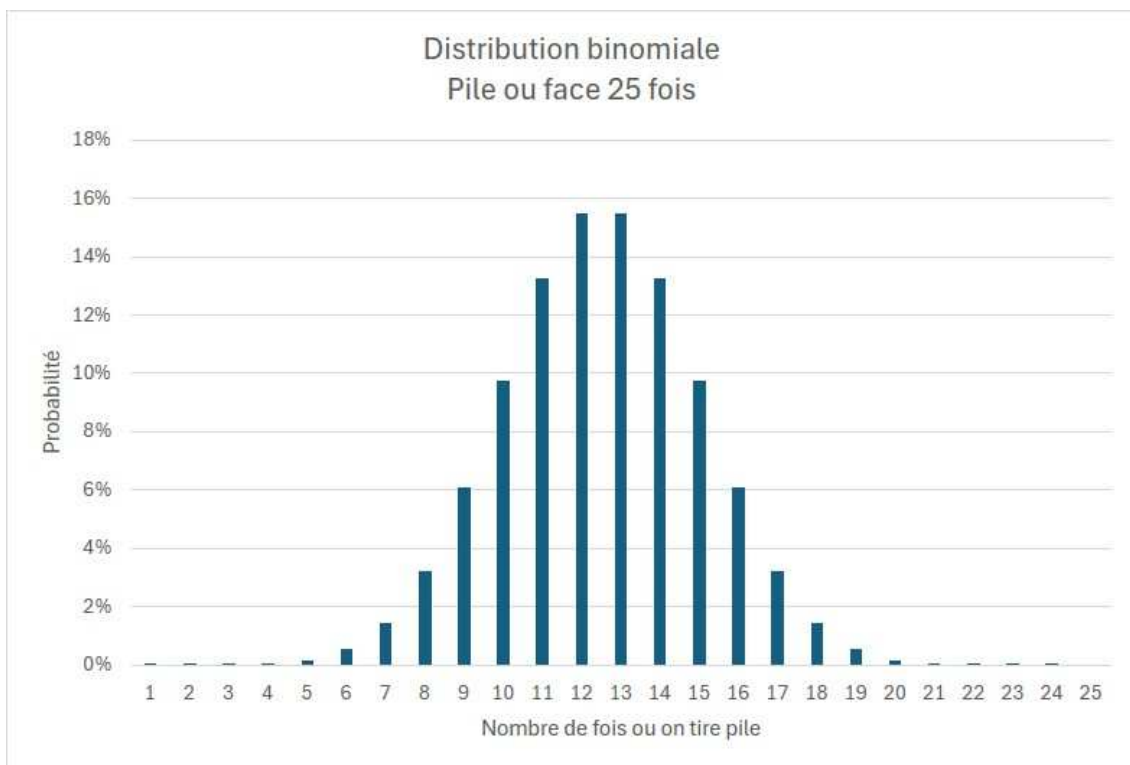
$$C_{30}^4 = \frac{32!}{4! \cdot (32 - 4)!} = 35\,960$$

4. Distributions théoriques

Une distribution théorique est une distribution de probabilités modélisée mathématiquement qui décrit comment on s'attend à ce que des données se répartissent, même avant d'avoir observé toutes les données.

4.1 Distribution binomiale

La loi binomiale modélise le nombre de succès obtenus lors de n répétitions indépendantes d'une expérience qui n'a que deux résultats possibles : succès ou échec, pile ou face,



Pour une distribution binomiale, la probabilité qu'un événement se produise k fois si on réalise N fois l'expérience est donnée par :

$$P(k) = C_N^k p^k \cdot (1 - p)^{N-k}$$

Avec

$C_N^k = \frac{N!}{k! \cdot (N-k)!}$ la combinaison sans répétition de N éléments différents pris k à k

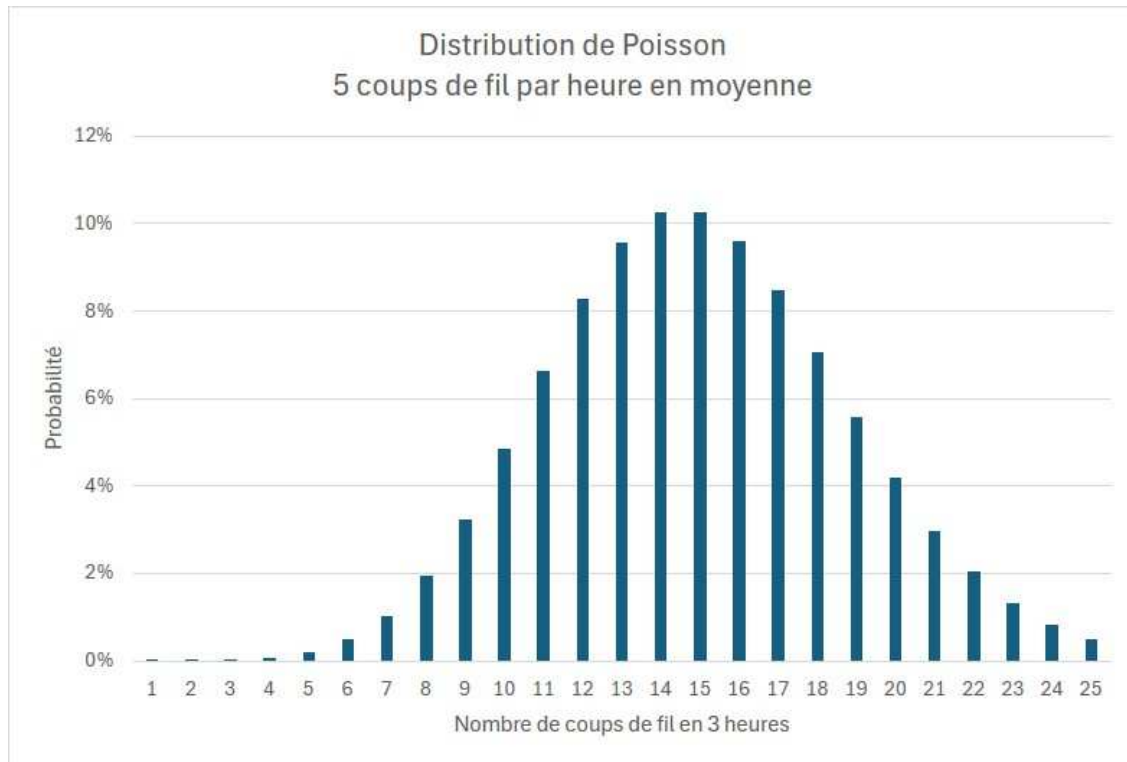
p la probabilité que l'événement se produise

Exemple : la probabilité de tirer 8 fois pile si on tire 25 fois pile ou face :

$$P(8) = \frac{25!}{8! \cdot (25 - 8)!} (0,5)^8 \cdot (1 - 0,5)^{25-8} = 3,2\%$$

4.2 Distribution de Poisson

La loi de Poisson modélise le nombre d'événements qui se produisent pendant un intervalle donné lorsque ces événements se produisent de manière indépendante et avec un taux moyen constant.



Pour une distribution de Poisson, la probabilité qu'un événement se produise k fois durant un intervalle I si le nombre moyen d'occurrence de cet événement par unité d'intervalle vaut m

$$P(k) = e^{-m.I} \cdot \frac{(m.I)^k}{k!}$$

Exemple, le secrétariat reçoit, en moyenne 5 coups de fil par heure, quelle est la probabilité qu'il reçoive 6 coups de fils en trois heures.

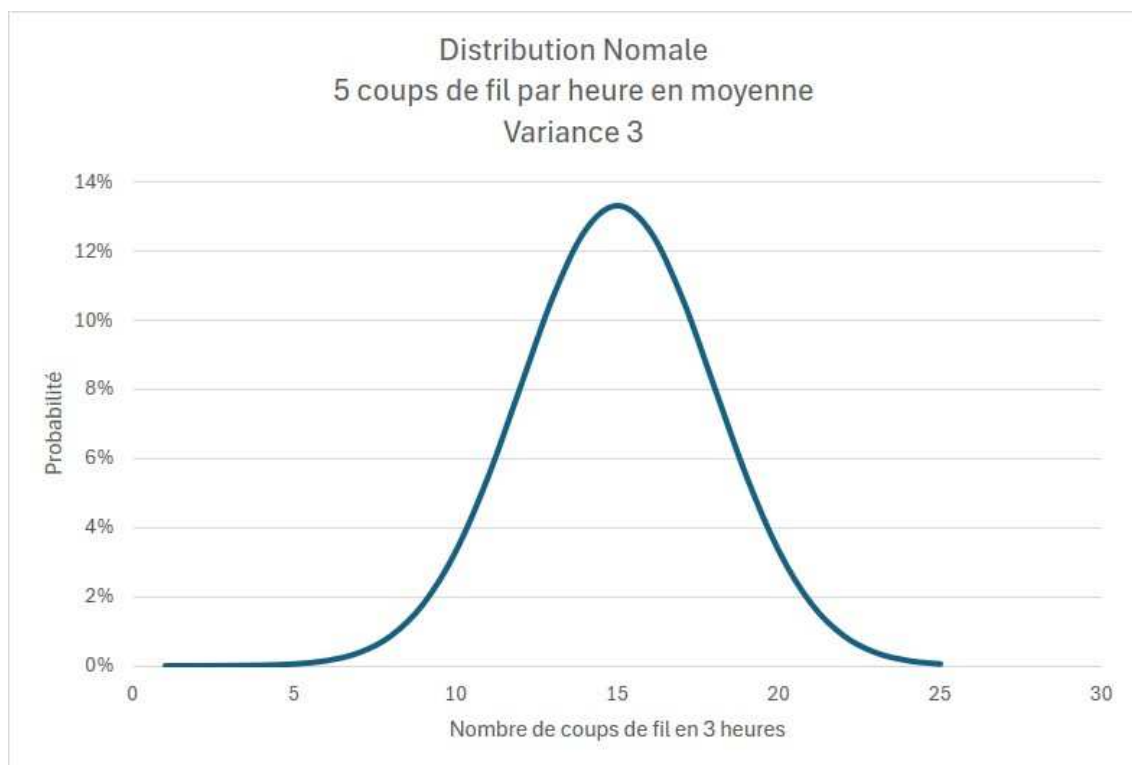
$$P(6) = e^{-5.3} \frac{(5.3)^6}{6!} = 0,48 \%$$

4.3 Distribution normale

La loi normale modélise la répartition des mesures d'une grandeur soumise à de nombreux petits effets aléatoires. La moyenne indique la valeur autour de laquelle les mesures se concentrent et l'écart-type caractérise l'importance de ces erreurs par rapport à la mesure.

La loi normale a de nombreuses applications, car elle modélise relativement bien des mesures expérimentales en sciences, des caractéristiques biologiques, des intervalles de confiance dans les sondages, ...

A titre d'exemple, On peut modéliser les coups de fil du secrétariat à l'aide d'une loi normale :



A la différence des lois binomiale et de Poisson, la loi normale est continue. D'autre part, elle est symétrique et centrée sur sa valeur moyenne.

4.4 Tables de probabilité

Pendant longtemps, les distributions théoriques ont été étudiées à l'aide de tables, depuis le développement de logiciel comme Excel, les tables sont devenues désuètes.